Monte Carlo Markov Chains simple examples in R

Danny Kaye

November 29, 2009

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 the two main MCMC systems, Metropolis Hastings and Gibbs, coded directly in R.

1 introduction

Consider a model for data $\boldsymbol{x}$ assumed to be observed values of some random variable $\boldsymbol{X}$. This model defines a probability distribution for $\boldsymbol{X}$ in terms of parameters $\boldsymbol{\theta}$, from a parameter space $\boldsymbol{\Theta}$, by a density function $p(\boldsymbol{x}|\boldsymbol{\theta})$. The value of this density is often called the likelihood function, as it describes the likelihood of this particular sample $\boldsymbol{x}$ in terms of the parameters $\boldsymbol{\theta}$.

In a Bayesian model initial knowledge about $\boldsymbol{\theta}$ is represented as a prior distribution having density $p(\boldsymbol{\theta})$. This may come from some 'expert' assessment of the parameter value or from some previous measurement or experiment. Statistical inference for $\boldsymbol{\theta}$ is obtained by using Bayes' Theorem to update knowledge about $\boldsymbol{\theta}$ in the light of the sample data $\boldsymbol{x}$.

This inference about $\boldsymbol{\theta}$ is provided by the posterior density given by Bayes' Theorem as:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \tag{1}$$

It is often convenient and sufficient to express Bayes' theorem to proportionality as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2}$$

Note that

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{3}$$

is obtainable as the constant needed to make $p(\boldsymbol{\theta}|\boldsymbol{x})$ a proper density, so that

$$\int p(\boldsymbol{\theta}|\boldsymbol{x}) = 1. \tag{4}$$

2 Example

This is a simple example of obtaining inference from some data and a model for it that allows an analytic solution. A small deviation from that model (replacing the Normal prior with a Teachers prior with $\nu = 4$) makes it impossible to treat the example analytically but makes little difference to numerical methods.

Consider some experiment in which a particular measurement, x, is of interest. This may be modelled as being a value of a random variable X having a Normal distribution with mean μ and variance σ^2, $N(\mu, \sigma^2)$. In this context the mean μ represents the average value of the measurement, while the variance σ^2 represents its variability.

The parameter μ is unknown, but prior knowledge about it may be represented by a $N(\mu_0, \sigma_0^2)$ density, and for this example values of $\mu_0 = 3.5$, $\sigma_0^2 = 1$ are suitable. The choice of σ_0^2 is in fact largely uninformative as a wide range (0.5 to 6.5 say) of values of μ are all quite likely.

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

The sample data to hand

3.51	3.50	3.52	3.50	3.51	3.50	3.52	3.50	3.51	3.51

can be summarised as $n = 10$, and $\sum x_i = 35.08$.

For this example both analytic and numerical solutions are readily available.

2.1 Analytic approach

The interest is in inference about μ given $\sigma = 0.01$.

With prior $\mu \sim N(3.5, 1)$

$$p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\mu - 3.5)^2} \tag{5}$$

$\bar{x}$, the sample mean, is a sufficient statistic for μ, as the likelihood, $l(\mu|\boldsymbol{x})$, depends only on $\boldsymbol{x}$ and σ (fixed), through the sampling distribution of $\bar{x}$ which is $N(\mu, \frac{\sigma^2}{n})$

$$\begin{aligned} l(\mu|\boldsymbol{x}) &\propto p(\bar{x}|\mu, \sigma^2) & (6) \\ &\propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{x})^2\right) & (7) \end{aligned}$$

so the posterior for μ is

$$\begin{aligned} p(\mu|\boldsymbol{x}) &\propto p(\mu) l(\mu|\boldsymbol{x}) & (8) \\ &\propto p(\mu) p(\bar{x}|\mu, \sigma^2). & (9) \end{aligned}$$

So, by inspection

$$p(\mu|\boldsymbol{x}) = \left(\frac{n}{2\pi\sigma^2}\right)^2 \exp\left(\frac{-n}{2\sigma^2}(\mu-\bar{x})^2\right) \qquad -\infty < \mu < \infty \tag{10}$$

$\boldsymbol{x}$ 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

$$\mu_n = \frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2} \tag{11}$$

and

$$\sigma_n^{-2} = n\sigma^{-2} + \sigma_0^{-2}. \tag{12}$$

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.

2.2 Metropolis Hastings

An MCMC solution to a problem consists of an MCMC constructed so that its stationary distribution is the distribution about which it is wished to draw inference, (see, for example, Gilks et al., 1996; Tierney, 1994). This problem is amenable to this approach.

A Metropolis-Hastings MCMC has the advantage of allowing the use of a symmetrical distribution from which to choose the candidate. Once the process has entered the state space of the target distribution it will not leave it.

Here

$$\begin{array}{rcl} p(\theta) & \sim & N(3.5, 1) \\ l(\theta|\boldsymbol{x}) & \propto & \exp\left(\frac{-n}{2\sigma^2}(\theta-\bar{x})^2\right) \\ & \propto & p(\bar{x}|\theta,\sigma) \end{array} \tag{13}$$

so the probability of transition from θ_1 the current state to θ_2 the candidate state is

$$p(\theta_1, \theta_2) \quad = \quad \min\left(\frac{p(\theta_2)p(\bar{x}|\theta_2,\sigma)}{p(\theta_1)p(\bar{x}|\theta_1,\sigma)}, 1\right) \tag{14}$$

The code shown in figure 1 is for running a Metropolis-Hastings, random walk MCMC sampler. With the transition probability above, and allowing 1000 samples as a burn in period, the output is a sample of 1000 values with $\hat{\mu} = 3.507999$ and $sd = 0.00304$. (Note that this result would, of course, be slightly different if run again, due to the small sample used and random differences. However, if a large enough sample is taken, run differences tend to zero with probability 1).

It is possible to draw a predictive sample from this density with the R command `y←rnorm(1000, samp, 0.01)`[1], where `samp` is the output from the sampler, this gives the histogram shown in Figure 2.

[1] `rnorm(n, m, s)` is the R command that generates n random numbers mean m, standard deviation s. So this generates 1000 samples using the MCMC sample from the posterior distribution of the mean as a vector of means (`samp`) and the known parameter σ as standard deviation.

```
function(theta, w = 1000, mu = 3.508, sigma = 0.01, n = 10)
{
#simple mh sampler for the example with mean 3.05
#sigma 0.01
    f <- function(theta, mu = 3.508, sigma = 0.01, n = 10)
    {
        z <- exp( - n/2/sigma^2 * (theta - mu)^2)
        z
    }
    z <- rep(0, n)
    p <- rep(0, w + n)
    theta <- 3.5
    pmu <- 3.5
    psigma <- 1
    for(i in 1:(w + n)) {
        new <- theta + rnorm(1, 0, 0.01)
        prob<-(dnorm(new,pmu,psigma)*f(new))
            /(dnorm(theta,pmu,psigma)*f(theta))
        alpha <- min(prob, 1)
        p[i] <- alpha
        if(runif(1) < alpha) theta <- new
        if(i > w) z[i - w] <- theta
    }z
}
```

Figure 1: *R code for the MCMC example.*

2.3 Asymptotic approach

As above the likelihood is

$$l(\theta|\boldsymbol{x}) \propto \exp\left(\frac{-n}{2\sigma^2}(\mu - \bar{x})^2\right) \tag{15}$$

so the log-likelihood is

$$L(\theta|\boldsymbol{x}) = C - \frac{n}{2\sigma^2}(\mu - \bar{x})^2 \tag{16}$$

where C is a constant. Differentiating and equating to zero gives

$$\frac{n}{\sigma^2}(\theta - \bar{x}) = 0 \tag{17}$$

so that

$$\hat{\theta} = \bar{x} \tag{18}$$

and a second differentiation gives

$$\sigma_n^{-2} = \frac{n}{\sigma^2} \tag{19}$$

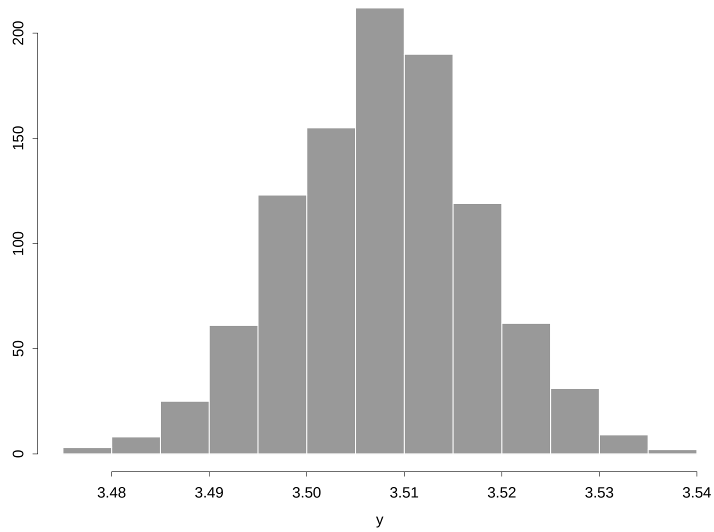


Figure 2: *Histogram of predictive sample of 1000 drawn from the density* μ $N(3.507999, 0.00304^2)$.

giving, in this case, $\mu = 3.508$ and $\sigma_n = 0.00316$.

A histogram of 1000 samples from $N(3.508, 0.00316^2)$ is shown in figure 3. While not a good density estimator, a histogram is useful for the gross comparison of two samples needed here. Note that this histogram appears to come from a distribution with a smaller standard deviation than that shown in Figure 2; the predictive sample is a safer estimate as it allows for the uncertainty in the point density estimation.

2.4 Numerical approach

The interest is in what may be inferred about $\boldsymbol{\theta} = (\mu)$ from a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$. Each component of the random sample is modeled as a value of X, being independently and identically distributed as $N(\mu, \sigma^2)$. Hence the probability density function (pdf) for the whole sample, the likelihood function, can be written as

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right). \tag{20}$$

The prior is as given above, which may, in practice, be a value from a previous experiment (μ) and a measure of variation derived from history (σ^2).

It is possible to numerically apply Bayes' Theorem. Of the many approaches available, the Bayes4 suite of software developed by Dr J C Naylor (see Naylor, 1982; Naylor and Shaw, 1983; Naylor and Smith, 1982) is chosen here. Bayes4 uses Gauss-Hermite quadrature to obtain values for posterior density parameters and applying it gives the posterior densities shown in Figure 4 and also a

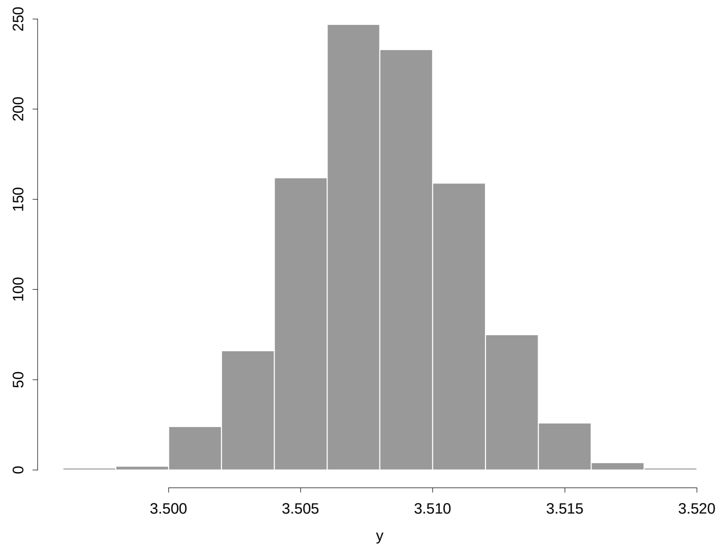


Figure 3: *Histogram of estimative density (Maximum likelihood approximation).*

numerical approximation to the posterior mean for μ of 3.508.

Repeating the exercise using a Student's t-distribution with 4 degrees of freedom ($t_{\nu=4}$) as prior, renders the calculation analytically intractable. However, using Bayes4 gives the posterior densities shown in figure 5, where the posterior estimate for μ is again 3.508.

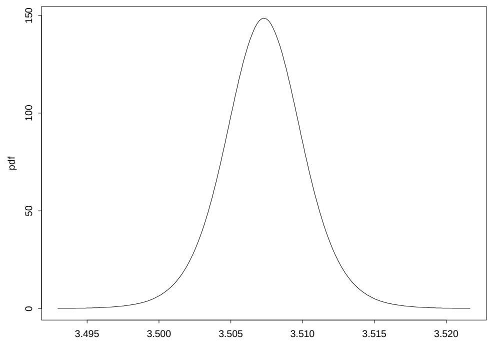
All of the examples above are extremely easy to describe. However, as soon as a $t_{\nu=4}$ prior is chosen their solution becomes, analytically at least, intractable. At the same time the result of the analysis is remarkably similar to that obtained with a Normal prior. Performing the analysis using Bayes4 is no more difficult than when a Normal prior is used.

The graphs in Figures 4 and 5 were produced using standard R routines, which apply spline curves, to interpolate between points.

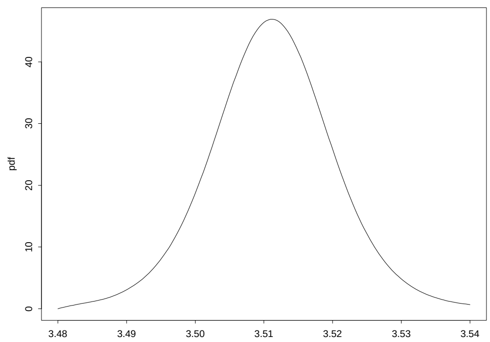
3 Markov Chains

The example above is trivial but the change in prior from Normal to Student's t renders it intractable to analytic methods without making either numerical quadrature or sampling more difficult. The main limitation to the use of both of these has been their computational intensity but, as Moore's Law (Moore, 1965) predicts, this is becoming less of a problem.

In Bayesian analysis the problem dimensionality is the same as the number of parameters of the problem. In problems of up to 10 parameters it is possible to use Monte Carlo integration or numerical quadrature (Smith et al., 1987; Naylor, 1982; Naylor and Smith, 1988), beyond that it rapidly becomes difficult. MCMC algorithms – especially Gibbs (Geman and Geman, 1984) when the density is specified as a full set of conditionals and Metropolis-Hastings (Hastings, 1970;



(a) Posterior Mean.



(b) predictive density.

Figure 4: *Posterior mean and predictive density, Normal prior.*

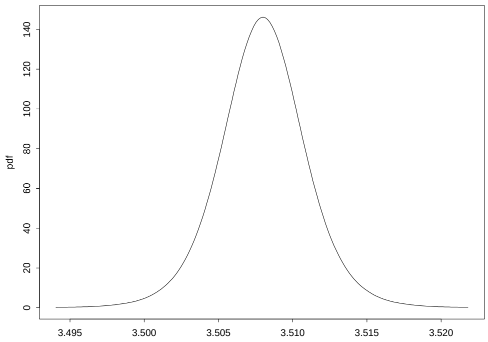
Metropolis et al., 1953) when a complete density is specified – are now commonly used when it is required to produce a sample from a density. These methods can be used when the density is specified either as a complete set of conditional densities or as a multivariate density and are very powerful, able to handle problems of the very high dimensionality needed.

The basic strategy is to construct a Markov Chain sampling scheme for which it can be shown that the equilibrium distribution is the density of interest. The great advantage of MCMC is that it is possible to devise methods such that the density of interest need only be known to proportionality.

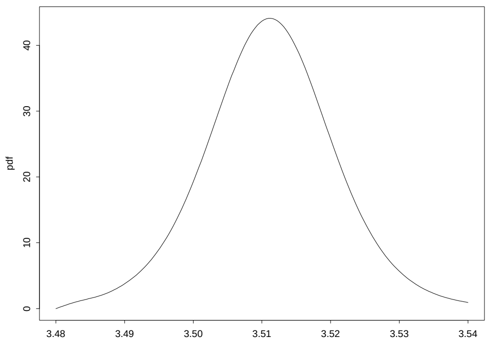
3.1 Some theory

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

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



(a) Posterior Mean.



(b) predictive density.

Figure 5: *Posterior mean and predictive density, $t_{\nu=4}$ prior.*

.

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 $X_n = i$.

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

$$P_i(X_t = j \text{ for some } t \geq 0) > 0. \tag{21}$$

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.

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 (Feller, 1970, p 385).

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 (Gilks et al., 1996).

3.2 Gibbs samplers

Gibbs samplers are the most popular and versatile of the MCMC methods discussed here, introduced to mainstream statistical analysis by Geman and Geman (1984) and developed in, amongst others,Gelfand and Smith (1990); Smith and Roberts (1993); Gilks et al. (1996). Gibbs samplers were originally used to analyse Gibbs distributions on lattices and in statistical physics where they were known as the *heat bath algorithm*. For an introduction and history of MCMC samplers see Gilks et al. (1996, Chapter 1).

3.2.1 Theory

Given a vector of random variables $T = (X, Y_1, \ldots, Y_n)$, some distribution $f(T) = f(x, y_1, \ldots, y_n)$, and the requirement to find characteristics (for example, but not limited to, position and shape parameters) of the marginal density

$$f(x) = \int \ldots \int f(x, y_1, \ldots, y_n) dy_1 \ldots dy_n,$$

then the obvious route is to calculate $f(x)$ and use it to find the required information. However, there are few cases where the integrations can be performed.

The Gibbs sampler provides an alternative method of obtaining a sample $X_1, \ldots, X_m \sim f(x)$ without having access to a direct mathematical description of $f(x)$. By simulating a large enough sample, the mean, variance and any other characteristic of $f(x)$ can be found. The MCMC converges to a sample from the required density and, assuming good quality random numbers, as the sample gets larger the effect on the precision of the estimates of a single extra sample on the values decreases. This means that by taking sufficient samples, values can be estimated to any desired degree of precision. This is contrary to the experience of sampling from a 'real' population where a sample can only increase to the size of the population; in MCMC there is no limit on the size of the sample. For example if some value that has a value in the region of 1 is being estimated and 1000000 samples which vary between 1.001 and 0.999 are taken, then the effect of adding another sample will only appear in the 8th decimal place of the answer.

The end results of any calculation, although based on a simulation, are the population quantities. For example $\overline{f(x)} = 1/m \sum_{i=1}^{m} X_i$ as

$$\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} X_i = \int_{-\infty}^{\infty} x f(x) dx = E[X]$$

By taking m large enough, any population statistic can be obtained to any required degree of precision.

As an example, starting with the pair of random variables (RVs) (X, Y), the Gibbs sampler generates the following sequence from $f(x)$ by sampling from the conditional distributions $f(x|y)$ and $f(y|x)$

$$Y'_0, X'_0, Y'_1, X'_1, Y'_2, X'_2, \ldots, Y'_k, X'_k$$

the initial value $Y'_0 = y'_0$ is specified, the remainder are obtained by alternatively generating

$$X'_j \sim f(x|Y'_j = y'_j)$$

$$Y'_{j+1} \sim f(y|X'_j = x'_j)$$

This is, clearly, a Markov Chain as it fulfils the condition, laid down in section 3.1, that the current state depends only on the previous state and no other. Under quite general conditions, (see Brooks and Roberts, 1995), the distribution of X'_k converges to $f(x)$ (the true marginal of X) as $k \to \infty$. Thus for k large enough, the final point of the sequence is effectively a sample point from $f(x)$. Also, the distribution of (X'_k, Y'_k) converges to $f(x, y)$.

3.2.2 Practice

The Gibbs sampler is a system that allows the generation of a sample from a density defined in terms of all its full conditional densities. Smith (1991) suggested that MCMC benefited from running many chains in parallel and the R system's ability with vectorised arithmetic makes running multiple chains easy.

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 $\boldsymbol{\theta}^{(1)}$.

Once the new realisation of $\boldsymbol{\theta}$ has been obtained the process is repeated. Theory states that the algorithm converges to a sample from the density, however the convergence can be extremely slow and it is common to throw away a large number of samples prior to establishing convergence using one of the well established tests (see, for example, Brooks and Roberts, 1995).

Figure 6 shows the first 5 samples, including the starting position $(20, -20)$, of a simple Gibbs sampler. The four frames show the progression from the start, each frame showing the two samples obtained in each step from the two

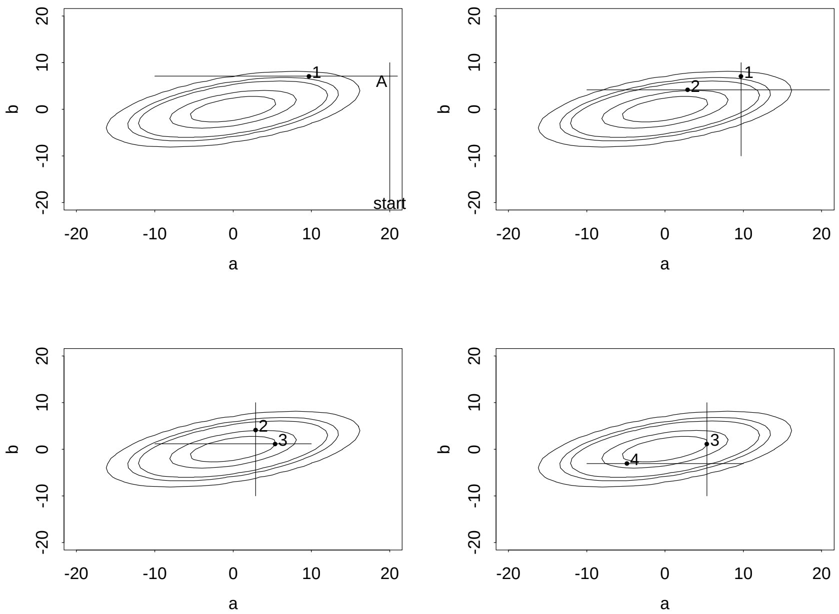


Figure 6: *The first 5 samples of a Gibbs sampler in state space $\Re^2$. The contour lines are percentage contour lines for the target density.*

conditional densities (following the line from start in the first diagram leads from the point $(x^{(0)}, y^{(0)})$ to the point $(x^{(1)}, y^{(0)})$ and the next line leads to the point $(x^{(1)}, y^{(1)})$). Note that the start point can be anywhere in the state space of the distribution. The chain quickly enters the 90% contour line shown for the target distribution; in the case of a more complex distribution it can take many more steps for the chain to reach this point. The chain converges to the required distribution but it is sometimes difficult to predict the rate of convergence (but see Rosenthal (1994) for some interesting discussion of convergence rates). The R code used to generate this sampler is shown in figure 7.

The density shown in figure 6 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 for such a density may be found, for example, in Feller (1970) and are:

the conditional density

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

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

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

```
function(x = c(0, 0, 5), y = c(0, 0, 2.5), rho = 0.5, n = 20, burn
= 0) {
    # set up the constants
    FAC <- sqrt(1 - rho^2)      #
    n <- n + burn
    # plot the contours
    con <- normgrid(a, b, mu = c(x[2], y[2]),
                sig = c(x[3], y[3]), rho = rho)
    contour(a, b, con, labex = 0, levels = c(0.1,
        0.05, 0.01, 0.005, 0.001))      #
    # generate the samples
    for(i in 1:n) {
        # generate a new Y from the current X
        ynew <- rnorm(1, (rho * y[3])/x[3] * x[1], FAC * y[3])
        # generate a new X from the new Y
        xnew <- rnorm(1, (rho * x[3])/y[3] * ynew, FAC * x[3])
        # output if past the burn in,
        # in this case output the number i
        if(i > burn)
            text(x[1], y[1], as.character(i - burn))
        # update the output
        x[1] <- xnew
        y[1] <- ynew
        out <- c(out, c(x[1], y[1]))
    }
    # output the plotting info for reuse
    out <- list(out, con)
    out
}
```

Figure 7: *R Gibbs sampler.*

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

The 10,000 samples in Figure 8 were generated in a similar manner, however this time the R routines make use of the ability of R to handle vectorised arithmetic to reduce the use of the for loop. This gives a much more efficient program. The code used is shown in Figure 9.

3.3 The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a method first proposed in Metropolis et al. (1953) the abstract might demonstrate how far from Bayesian statistics the authors were:

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting

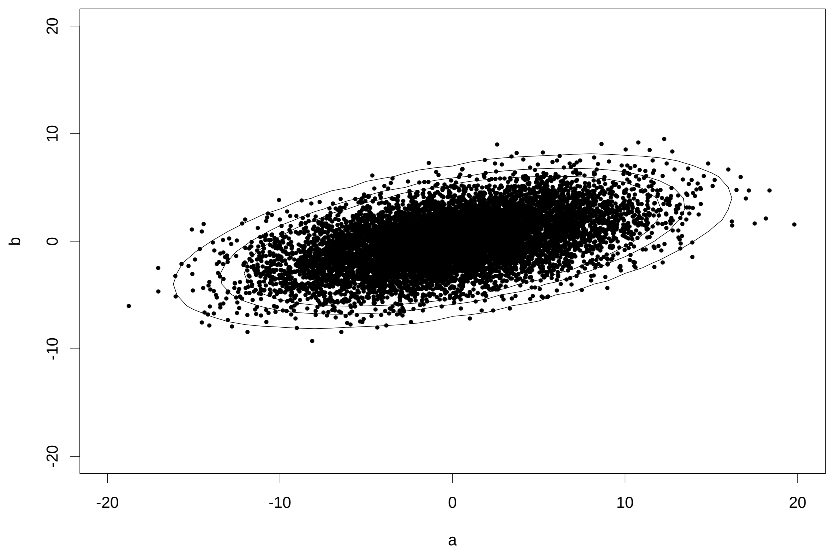


Figure 8: *10,000 samples from a bivariate Normal as described above.*

individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion. Metropolis et al. (1953)

The results were generalised in Hastings (1970).

3.3.1 Theory

If the target density $\pi(x)$ is known to proportionality, and a Markov process with candidate-generating probability $q(x, y)$ is constructed, so the probability of transition from state x to state y is $q(x, y)$, then, with the chain at point $X_n = x_n$, a candidate value Y for X_{n+1} is generated from the density $q(x, y)$. 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. Details of the derivation of the acceptance probability have been extensively discussed in the literature, see, for example Chib and Greenberg (1994).

The acceptance probability only depends on π through the ratio $\pi(y)/\pi(x)$, so π is only needed to proportionality. If the step is accepted the chain moves to $X_{n+1} = y$, otherwise it remains at $X_{n+1} = x$.

```
function(x = c(0, 0, 5), y = c(0, 0, 2.5), rho = 0.5,
        n = 10000, burn = 10,cont=T)
{
    FAC <- sqrt(1 - rho^2)      #
# draw the contour lines
    if(cont)
        contour(a, b, normgrid(a, b, mu = c(x[2], y[2]),
                        sig = c(x[3], y[3]), rho = rho), labex =0,
                        levels = c(0.1, 0.05, 0.01, 0.005, 0.001)) # replicate the x and y
    x2 <- rep(x, n)
    dim(x2) <- c(3, n)
    x2 <- t(x2)
    y2 <- rep(y, n)
    dim(y2) <- c(3, n)
    y2 <- t(y2) #
# burn in
    for(i in 1:burn) {
        ynew <- x2[,2] + rnorm(n, (rho * y2[, 3])/x2[, 3]
                        * (x2[, 1] - x2[,2]), FAC * y2[,3])
        xnew <- y2[,2] + rnorm(n, (rho * x2[, 3])/y2[, 3]
                        * (ynew - y2[,2]), FAC * x2[, 3])
        x2[, 1] <- xnew
        y2[, 1] <- ynew
    }
    points(x2, y2)
    list(x2, y2)
}
```

Figure 9: *Vectorised R Gibbs sampler.*

If the candidate generating density is symmetric, an important special case, $q(x, y) = q(y, x)$ 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 $\pi(y)/\pi(x)$. In other words, 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, see, again, Chib and Greenberg (1994) and Metropolis et al. (1953).

If the Markov process is irreducible and the candidate generating density is symmetric, then $q(x, y) > 0$ and $q(x, y) = q(y, x)$ for all x and y and

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

3.3.2 Random Walk Chains

If the target density is of dimension k and $\boldsymbol{E} = R^k$ where $\boldsymbol{E}$ is the state space of the target density $f(\cdot)$, then the candidate $\boldsymbol{Y}$ is generated by drawing $\boldsymbol{Z}$ independently from $f(\cdot)$ and setting $\boldsymbol{Y} = \boldsymbol{x} + \boldsymbol{z}$, $q(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{y} - \boldsymbol{x}) = f(\boldsymbol{z})$, and the kernel driving the chain is a random walk.

In this case q is symmetric, the process is irreducible and the acceptance probability is

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

3.3.3 Independence Chains

Candidate steps Y can also be chosen from a fixed density f. In this case $q(x, y) = f(y)$ and

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

where $w(x) = \pi(x)/f(x)$. The function w is the importance weight function that would be used in importance sampling if observations were generated from f.

3.3.4 Practice

The easiest algorithm to implement is the random walk. It has the major advantage that the candidate can be generated from a symmetrical distribution, in the sense that $q(x, y) = q(y, x)$, and so the transition probability is

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

In addition, the process will not leave the state space once it has entered it, so $\pi(x) = 0$ is not possible.

3.4 Re-sampling

Re-sampling is a technique that allows a sample from one density to be used to generate a sample from a similar or related density. For example a sample from a prior density may be used to obtain a sample from a posterior density.

As a specific example of this approach, let us consider the **Rejection method**.

Suppose that given a sample from a continuous density $g(\theta)$ a sample is required from a density $h(\theta)$ that is absolutely continuous with respect to $g(\theta)$ [2].

More generally, given a function $f(\theta)$ which is normalisable to a density

$$h(\theta) = \frac{f(\theta)}{\int f(\theta) d\theta}$$

[2] All areas of $h(\theta)$ can be reached from $g(\theta)$, i.e. $g(\theta)$ is an envelope of $h(\theta)$.

and there exists an identifiable constant $M > 0$ such that $\frac{f(\theta)}{g(\theta)} \leq M$, it is possible to obtain a sample from $h(\theta)$ in the following way, (see Smith and Gelfand, 1992).

1. Generate θ from $g(\theta)$.
2. Generate u from uniform $(0, 1)$.
3. If $u \leq \frac{f(\theta)}{Mg(\theta)}$ accept, θ.
4. repeat steps 1-3.

Any accepted θ is then a random variate from $h(\theta)$.

References

S. Brooks and G. Roberts. Diagnosing convergence of mcmc algorithms. Technical report, Cambridge University, 1995.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. Technical report, Washington University, 1994.

W. Feller. *An introduction to probability theory and its applications*. Wiley, New York, 3 edition, 1970.

A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *jasa*, 85(410):398–409, 1990.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *ieeep*, 6:721–741, 1984.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (editors). *Markov Chain Monte Carlo in practice*. Chapman and Hall, London, 1996.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrics*, 57(1):97–109, 1970.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equations of state calculated by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

G. E. Moore. Cramming more components onto integrated circuits. *Electronics Magazine*, 4, 1965.

J. C. Naylor. *Some numerical aspects of Bayesian inference*. PhD thesis, University of Nottingham, October 1982.

J. C. Naylor and J. E. H. Shaw. Bayes 4 user guide. Private, The Nottingham Trent University, 1983.

J. C. Naylor and A. F. M. Smith. An archaelogical inference problem. *Journal of the American Statistical Association*, 83(4):588–595, 1988.

J. C. Naylor and A. F. M. Smith. Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, 31:214–225, 1982.

JS Rosenthal. Theoretical rates of convergence for markov chain monte carlo. In *Interface '94*. Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1, June 1994.

A. F. M. Smith. Bayesian computational methods. *Philosophical Transactions of the Royal Society A*, 337:369–386, 1991.

A. F. M. Smith and A. E. Gelfand. Bayesian statistics without tears: a sampling - resampling perspective. *The American Statistician*, 146(2):84–88, 1992.

A. F. M. Smith and G. O. Roberts. Bayesian computation via the Gibbs sampler and related Markov-chain and Monte-Carlo methods. *Journal of the Royal Statistical Society*, 55(1):3–23, 1993.

A. F. M. Smith, A. M. Skene, J. E. H. Shaw, and J. C. Naylor. Progress with numerical and graphical methods for practical Bayesian statistics. *The Statistician*, 36:75–82, 1987.

L. Tierney. Markov Chains for exploring posterior distributions. Technical report, University of Minnesota, 1994.