Contents

11

Markov chain Monte Carlo

Until now we have simply assumed that we can draw random variables, vectors and processes from any desired distribution. For some problems, we cannot do this either at all, or in a reasonable amount of time. It is often feasible however to draw dependent samples whose distribution is close to and indeed approaches the desired one. In Markov chain Monte Carlo (MCMC) we do this by sampling $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$ from a Markov chain constructed so that the distribution of $\boldsymbol{x}_i$ approaches the target distribution.

The MCMC method originated in physics and it is still a core technique in the physical sciences. The primary method is the Metropolis algorithm, which was named one of the ten most important algorithms of the twentieth century. MCMC, whether via Metropolis or modern variations, is now also very important in statistics and machine learning.

In MCMC problems, the quantity $\boldsymbol{x}$ could be an ordinary vector, or the path of a process, or an image or any more complicated object that can be generated by Monte Carlo. The desired distribution of $\boldsymbol{x}$ is conventionally denoted by π in MCMC problems, so $\pi(\boldsymbol{x})$ is a probability mass function for discrete state spaces, and $\pi(\boldsymbol{x})$ is a probability density function for continuous state spaces. It is very common that we cannot compute $\pi(\boldsymbol{x})$ but have access instead to an unnormalized version $\pi_u(\boldsymbol{x})$. That is $\pi(\boldsymbol{x}) = \pi_u(\boldsymbol{x})/Z$ for an unknown constant $Z = \int_{\mathcal{X}} \pi_u(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ (or $\sum_{x\in\mathcal{X}} \pi_u(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ in the discrete case). This leads us to prefer computations that use π only through ratios such as $\pi(\boldsymbol{y})/\pi(\boldsymbol{x}) = \pi_u(\boldsymbol{y})/\pi_u(\boldsymbol{x})$, avoiding the unknown Z.

The constant Z is called the ***partition function*** in physics. In later examples Z will depend on some other quantities, making the term ‘function’ seem more intuitively reasonable.

As usual, we estimate an expectation $\mu = \int f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i). \tag{11.1}$$

Our new context brings some challenges. First, because $\boldsymbol{x}_i$ have a distribution that approaches π but is usually not equal to π, the estimate (11.1) is biased. Second, the $\boldsymbol{x}_i$ are (in general) statistically dependent. Therefore the variance of $\hat{\mu}$ is not as simple as in the case where $\boldsymbol{x}_i$ are independent, and it is harder to estimate. In extreme cases, the $\boldsymbol{x}_i$ can get stuck in some subset of their domain and then $\hat{\mu}$ will fail to converge to μ. These two issues run through MCMC developments.

Not every use of MCMC output looks at first like estimating an integral as in (11.1). In statistical and machine learning applications we might not have such an f in mind when we look at scatterplots, histograms or other graphical displays of $\boldsymbol{x}_i$. However, the height of a bar in a histogram does translate into such an integral as does the proportion of points in some region of a scatterplot. Approximating expectations by averages is never too far away, even in a graphical exploration: we want our sample of $\boldsymbol{x}_i$ to reflect some aspect of the distribution π that we would have gotten from a much larger number of points $\boldsymbol{x}$ (even infinitely many) than we actually used.

This chapter introduces MCMC and develops the Metropolis-Hastings sampler. Chapter 12 presents another important method, known as the Gibbs sampler and also known as the heat-bath method or Glauber dynamics. Chapter **??** considers some more advanced MCMC methods.

11.1 The need for MCMC

Here we consider some distributions that are very hard if not impossible to get IID samples from. First, there is a problem that motivated Metropolis et al. (1953) to invent MCMC. Figure 11.1 shows $N = 224$ equally large circular disks packed into a square. The square has periodic boundary conditions, commonly called wraparound, so that a point moving continuously off the right side appears on the left and vice versa, while a similar rule connects top and bottom. In the left panel the disks just barely fit while the example on the right has more room.

Their model was that the disks were placed completely at random subject to one condition: no overlap. In principle, one could sample N center points independently and discard the results should any pair of centers be closer than twice the desired disk radius. For the case in the left panel, the disks can only fit if they are close to a hexagonal grid and it would take an enormous amount of computation per accepted configuration to sample that distribution. It would be smarter to sample sequentially, choosing centers one at a time and never placing one too close to a previous point, but that rule could get stuck with a configuration of less than N points to which no new point could be added. The solution they came up with was to place the points in an acceptable configuration

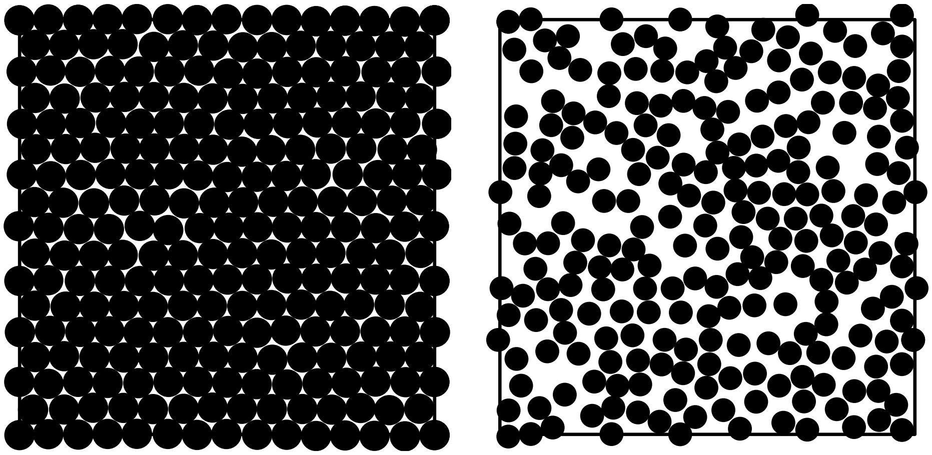

Figure 11.1: The left panel shows $N = 224$ disks of diameter about 0.0692 closely packed into the unit square. The boundary of the square is visible in a few places. The boundary is periodic, and so a disk that intersects an edge is plotted twice, and a disk intersecting the corner is plotted four times. The right panel shows 224 disks of diameter about 0.0536.

and give them random perturbations to simulate random placement. We revisit this problem in §11.7 after considering how to design perturbations consistent with a desired target distribution.

A common problem in statistics and machine learning is to predict a binary label $Y \in \{0, 1\}$ from a vector $\boldsymbol{z} \in \mathbb{R}^d$ of features. One of the simpler models for this problem is the ***probit model*** with $\mathbb{P}(Y = 1 \,|\, \boldsymbol{z}) = \Phi(\boldsymbol{z}^\mathsf{T}\beta)$ for an unknown parameter $\beta \in \mathbb{R}^d$, where Φ is the $\mathcal{N}(0, 1)$ cumulative distribution function. Then $\mathbb{P}(Y = 0 \,|\, \boldsymbol{z}) = 1 - \Phi(\boldsymbol{z}^\mathsf{T}\beta)$ and we cover both cases with $\mathbb{P}(Y = y \,|\, \boldsymbol{z}) = \Phi(\boldsymbol{z}^\mathsf{T}\beta)^y \times (1 - \Phi(\boldsymbol{z}^\mathsf{T}\beta))^{1-y}$.

In a Bayesian analysis of the probit model using data $(\boldsymbol{z}_1, y_1), \dots, (\boldsymbol{z}_m, y_m)$, we begin with a prior distribution $p(\beta)$. Next, suppose that $\boldsymbol{z}_i$ are independent and identically distributed with some distribution $g(\boldsymbol{z})$. With this information, the posterior distribution of β is

$$\begin{aligned} p(\beta \,|\, (\boldsymbol{z}_i, y_i), 1 \leqslant i \leqslant m) &\propto p(\beta) \prod_{i=1}^m g(\boldsymbol{z}_i) \prod_{i=1}^m \Phi(\boldsymbol{z}_i^\mathsf{T}\beta)^{y_i} (1 - \Phi(\boldsymbol{z}_i^\mathsf{T}\beta))^{1-y_i} \\ &\propto p(\beta) \prod_{i=1}^m \Phi(\boldsymbol{z}_i^\mathsf{T}\beta)^{y_i} (1 - \Phi(\boldsymbol{z}_i^\mathsf{T}\beta))^{1-y_i}. \end{aligned}$$

The first $\propto$ above stems from the denominator in Bayes rule not depending on β. The second $\propto$ comes from the fact that the distribution of $\boldsymbol{z}_i$ does not play a role in the posterior distribution of β.

Now suppose that we want some property of the posterior distribution of β. Let it be the posterior mean of $f(\beta)$; for instance a posterior moment or

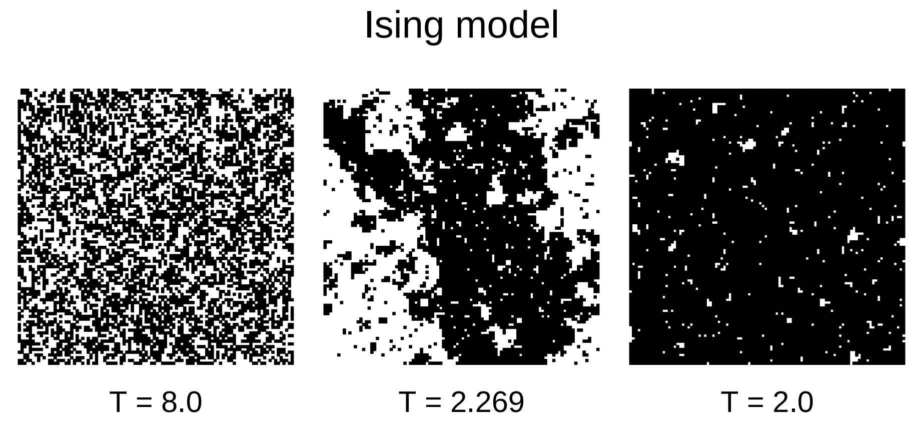

Figure 11.2: The Ising model with $J = 1$ and $B = 0$, sampled at 3 temperatures T on a 100×100 grid, with periodic boundary conditions. The left panel is roughly half black but some renderings make it look like a higher fraction black.

probability. Then we want to compute

$$\frac{\int_{\mathbb{R}^p} f(\beta)p(\beta)\prod_{i=1}^{d} \Phi(\boldsymbol{z}_i^{\mathsf{T}}\beta)^{y_i}(1-\Phi(\boldsymbol{z}_i^{\mathsf{T}}\beta))^{1-y_i}\,\mathrm{d}\beta}{\int_{\mathbb{R}^p} p(\beta)\prod_{i=1}^{d} \Phi(\boldsymbol{z}_i^{\mathsf{T}}\beta)^{y_i}(1-\Phi(\boldsymbol{z}_i^{\mathsf{T}}\beta))^{1-y_i}\,\mathrm{d}\beta}. \tag{11.2}$$

The integrals in (11.2) can be very hard to compute. In many cases of interest, d is large giving the integrals a high dimension. It is also common for the posterior distribution to concentrate within a very small region of space. Indeed we hope for that because it means the data are informative about β. Finally, if the number m of data points is large then the product within these integrands can be very small at any β, even the posterior mode, possibly underflowing in floating point computation.

To fit (11.2) into the methods of this chapter we first let our variable $\boldsymbol{x}$ be the unknown parameter β. Then take π_u to be the unnormalized posterior density, that is $\pi_u(\beta) = p(\beta)\prod_{i=1}^{m} \Phi(\boldsymbol{z}_i^{\mathsf{T}}\beta)^{y_i}(1-\Phi(\boldsymbol{z}_i^{\mathsf{T}}\beta))^{1-y_i}$. We revisit this model in §12.3.

Our third hard problem is sampling the ***Ising model***. It is a model for points on a regular grid that take either the value $+1$ or -1 with a dependence structure described below. Figure 11.2 shows some example outcomes at three different values of a temperature parameter T described below. Our discussion of the Ising model involves terms from physics such as energy, temperature and the Boltzman distribution. These notions are also used in some MCMC problems that have no connection to physics.

We begin with points (r, s) for $1 \leqslant r, s \leqslant L$ of a two dimensional grid. At each of $N = L^2$ grid points there is a variable equal to 1 or -1. Each grid point has 4 immediate neighbors in the grid, using a periodic boundary. For convenience let $\boldsymbol{x} \in \{-1, 1\}^N$ represent all of these values. For instance we might have the (r, s) point at index $j = r + (s-1)L$ of $\boldsymbol{x}$.

We write $j \sim k$ if points j and k are neighbors. To describe $\pi(\boldsymbol{x})$ we introduce the function

$$H(\boldsymbol{x}) = -J\sum_{j\sim k} x_j x_k - B\sum_{j=1}^{N} x_j. \tag{11.3}$$

This function is called the ***Hamiltonian***. Here $\sum_{j\sim k} x_j x_k$ is the number of neighbor pairs with matching signs minus the number that differ.

The Hamiltonian has units of energy and the conversion from energy to probability is via Boltzmann's law

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x}) \propto \exp\Bigl(-\frac{H(\boldsymbol{x})}{k_B T}\Bigr). \tag{11.4}$$

Here, $T > 0$ is the temperature of the system in degrees Kelvin, and $k_B > 0$ is Boltzmann's constant. It is mathematically convenient to rescale the temperature parameter so that $k_B = 1$. Then we write

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x}) \propto \exp\Bigl(-\frac{H(\boldsymbol{x})}{T}\Bigr) = \exp\bigl(-H(\boldsymbol{x})\beta\bigr), \tag{11.5}$$

where β is an inverse temperature. We will call T the temperature, though it is now measured in energy units. Even in problems not originating from physics, it can still be useful to interpret $-\log\pi(\boldsymbol{x})$ as an energy (divided by temperature).

Lower energy levels have higher probability. Anthropomorphizing a little, physical systems 'prefer' to be at lower energy levels. In statistical terms, the energy $H(\boldsymbol{x})$ is a random variable and T or β is the parameter in its distribution.

When $J > 0$, any matching neighbors, $x_j = x_k$ for $j \sim k$, lower the energy $H(\boldsymbol{x})$ thereby raising the probability $\pi(\boldsymbol{x})$ and mismatched neighbors have the opposite effect. This is known as the ***ferromagnetic*** case, from uses where the x_j are positive or negative charges. If $B \neq 0$ then the model is biased towards more charges of the same sign as B. If $J < 0$ then we have the ***antiferromagnetic*** case and the neighbors tend to differ from each other.

If $Y = f(\boldsymbol{x})$ and we want $\mathbb{E}(Y)$ then we need to compute

$$\mathbb{E}(Y) = \frac{1}{Z}\sum_{\boldsymbol{x}\in\{-1,1\}^N} e^{-H(\boldsymbol{x})/T} f(\boldsymbol{x}), \tag{11.6}$$

where,

$$Z = Z(T) = \sum_{\boldsymbol{x}\in\{-1,1\}^N} e^{-H(\boldsymbol{x})/T}. \tag{11.7}$$

The normalizing constant Z depends on the temperature T and $Z(T)$ is called the partition function.

The expectation in (11.6) is an average over 2^N configurations. Even for a small 100×100 grid there are then over 10^{3000} states in the sum. As described by MacKay (2005, Chapter 29.1), the interesting versions of the Ising model might have most of their probability distributed over $2^{N/2}$ cases. There are then an enormous number of these cases while at the same time they constitute an extremely rare subset of all the cases. Plain random sampling won't often find them. Importance sampling is not well suited either, because there is no obvious alternative distribution to sample IID from.

To conclude this section, we consider the role of the temperature T. Let $\boldsymbol{x}$ and $\widetilde{\boldsymbol{x}}$ be two states with $H(\widetilde{\boldsymbol{x}}) > H(\boldsymbol{x})$. Then

$$\frac{\mathbb{P}(\widetilde{\boldsymbol{x}})}{\mathbb{P}(\boldsymbol{x})} = \exp\Bigl(\frac{H(\boldsymbol{x}) - H(\widetilde{\boldsymbol{x}})}{T}\Bigr) = \exp\bigl(H(\boldsymbol{x}) - H(\widetilde{\boldsymbol{x}})\bigr)^{1/T} < 1. \tag{11.8}$$

As T increases, the probability ratio above increases, and approaches 1 as $T \to \infty$. This means that in a very hot system, the Ising model is nearly a uniform distribution on all of its 2^N states and a uniform distribution is easy to sample. In many other problems, sampling $\pi(\boldsymbol{x})^{1/T}$ for some $T > 1$ is easier than sampling $\pi(\boldsymbol{x})$. That approach can be thought of as increasing the temperature. We consider it in §**??**.

For $H(\widetilde{\boldsymbol{x}}) > H(\boldsymbol{x})$, as T decreases the probability ratio in (11.8) decreases, approaching 0 as $T \to 0$. In a very cold system, the Ising model puts almost all of its probability on states that achieve the minimum value of H. It produces a uniform distribution on those states, which are called ***ground states***.

The examples in Figure 11.2 have $B = 0$ and $J = 1$. The sampling for these images (with a caveat) is described in §11.8. At $T = \infty$, the points x_j are independent $\mathbf{U}\{-1, 1\}$ random variables. The Ising model (with $J > 0$ and $B = 0$) then looks like salt and pepper mixed together. The first panel of Figure 11.2 shows a hot system at $T = 8$.

For $T = 0$, we must find the ground states. If possible, they would have $x_j = x_k$ whenever $j \sim k$. We can in fact attain the ground state by either having every $x_j = 1$ or every $x_j = -1$. Furthermore, no other state attains as low an energy. Thus the Ising model at $T = 0$ plots as either a solid black image, or a solid white image, each with probability 1/2. The third panel of Figure 11.2 shows a cold system at $T = 2$.

As we consider in §11.8, the interesting temperatures are intermediate. One realization is shown in the middle panel of Figure 11.2.

11.2 Markov chains

Our solution to hard sampling problems like the Ising model will be to run a Markov chain for a long time so that the values of the chain have a distribution which approaches π. This section presents some results about Markov chains, describing conditions under which they have a unique stationary distribution π and for which long run averages settle down to expectations under π. For the purposes of studying MCMC, we treat most of these results as given facts to be built upon. Readers looking for derivations can find them in Feller (1968), Norris (1998) or Levin et al. (2009).

We will consider random variables $X_i \in \Omega$ for $0 \leqslant i < \infty$. The set Ω is the ***state space***. The random variables $X_i \in \Omega$ satisfy the ***Markov property*** if

$$\mathbb{P}(X_{i+1} \in A \mid X_j = x_j,\ 0 \leqslant j \leqslant i) = \mathbb{P}(X_{i+1} \in A \mid X_i = x_i)$$

for all $A \subset \Omega$. The Markov property is one of memorylessness. The distribution of X_{i+1} given the past depends only on X_i and not on how the chain got to X_i.

The Markov property can be useful for quite general state spaces. In this section we are concerned mostly with finite state spaces

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_M\}$$

for an integer $M > 0$. Countably infinite state spaces such as the positive integers are briefly considered.

A ***Markov chain*** is a sequence of random variables $X_i \in \Omega$ for $0 \leqslant i < \infty$ which satisfy the Markov property. To fully describe the distribution of the Markov chain we have to specify $\mathbb{P}(X_0 = x_0)$ as well as $\mathbb{P}(X_{i+1} = x_{i+1} \mid X_i = x_i)$.

The Markov chain $X_i \in \Omega$ is a ***time-homogenous*** Markov chain if

$$\mathbb{P}(X_{i+1} = y \mid X_i = x) = \mathbb{P}(X_1 = y \mid X_0 = x).$$

We will call them homogenous Markov chains for short. We will emphasize homogenous Markov chains, but nonhomogenous ones are needed in a few places. For homogenous chains on a finite state space, the transition probabilities are completely described by the $M \times M$ matrix P with elements $P_{jk} = \mathbb{P}(X_1 = \omega_k \mid X_0 = \omega_j)$. We will also write the elements of this matrix as $P(\omega_j \to \omega_k)$.

Suppose that $X_0 \sim \boldsymbol{p}_0 = (p_0(\omega_1), p_0(\omega_2), \dots, p_0(\omega_M))$, that is $\mathbb{P}(X_0 = \omega_j) = p_0(\omega_j)$. Then

$$p_1(\omega_k) \equiv \mathbb{P}(X_1 = \omega_k) = \sum_{j=1}^{M} p_0(\omega_j) P(\omega_j \to \omega_k) \tag{11.9}$$

which we write as $X_1 \sim \boldsymbol{p}_1$. From equation (11.9) we see that $\boldsymbol{p}_1 = \boldsymbol{p}_0 P$.

Notice that the probabilities are interpreted here as a row vector, multiplied on the right by the transition probability matrix. The more usual convention in matrix algebra is to treat a vector as a column and then multiply it on the left by a matrix. That convention for vectors conflicts with the perhaps stronger convention specific to Markov chains, in which P_{jk} is the probability of going from j to k. Furthermore, we have another use for multiplication of P by a vector on the right.

Let $\boldsymbol{f}$ be the column vector with values $f(\omega_j)$ for $j = 1, \dots, M$. Let $\boldsymbol{h}$ be the column vector with values $h(\omega_j) = \mathbb{E}(f(X_{i+1}) \mid X_i = \omega_j)$. Then (Exercise **??**) $\boldsymbol{h} = P\boldsymbol{f}$. In words, multiplying a function of Ω by P gives all the conditional expected values of that function one step into the future. That is, $P\boldsymbol{f}$ is a kind of forecast.

The same argument which gave $\boldsymbol{p}_1 = \boldsymbol{p}_0 P$ also gives $\boldsymbol{p}_2 = \boldsymbol{p}_1 P = p_0 P^2$. More generally $\boldsymbol{p}_n = \boldsymbol{p}_0 P^n$ for $n \geqslant 1$. Similarly, $P^n \boldsymbol{f}$ is a conditional expectation

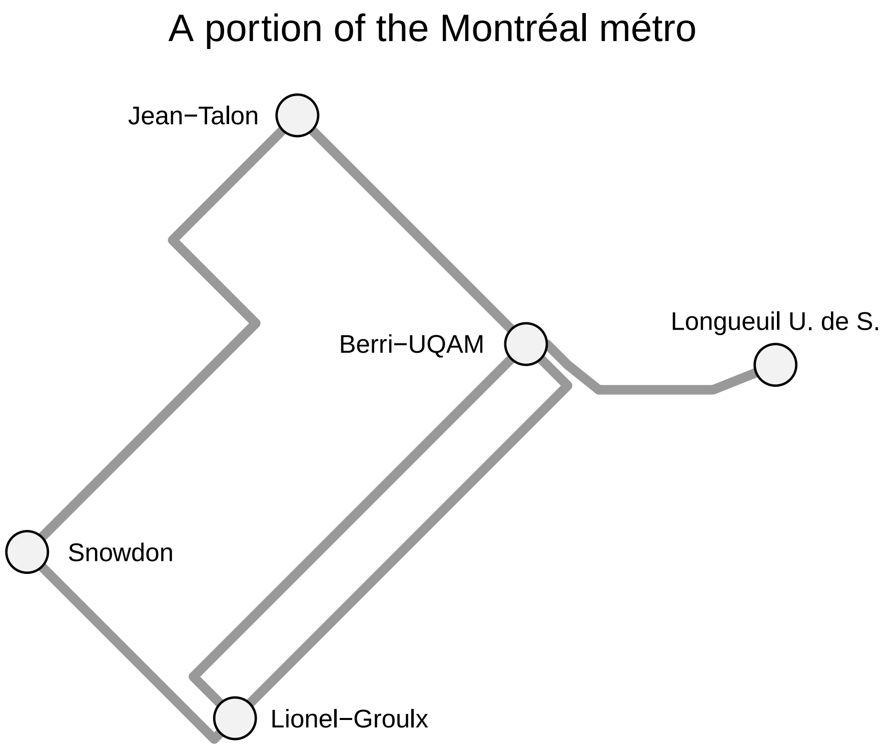

Figure 11.3: Five stations of the Montréal metro system used in a Markov chain example.

forecasting $f(X_{i+n})$ as a function of the the present value of X_i. Combining these expressions yields $\mathbb{E}(f(X_n)) = \boldsymbol{p}_0 P^n \boldsymbol{f}$.

When $\boldsymbol{p}_0 = \boldsymbol{q}$, that is $X_0 \sim \boldsymbol{q}$, then we write $\mathbb{E}_{\boldsymbol{q}}(\cdot)$ and $\mathbb{P}_{\boldsymbol{q}}(\cdot)$ for expectations and probabilities, respectively, for the Markov chain. If $\boldsymbol{q}$ is a point-mass, $\boldsymbol{q}(\omega) = 1$, concentrated on one state ω then we write $\mathbb{E}_\omega(\cdot)$ and $\mathbb{P}_\omega(\cdot)$.

Now we look at an example, based on a random walk, except the walker will take the subway. Figure 11.3 shows a central portion of the Montréal metro. The transition probability matrix P between the stations is as follows:

$$\begin{array}{l} \\ \text{Jean-Talon} \\ \text{Snowdon} \\ \text{Lionel-Groulx} \\ \text{Berri-UQAM} \\ \text{Longueuil} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{JT} & \text{S} & \text{LG} & \text{B} & \text{L} \end{array} \\ \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/3 & 0 & 2/3 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix} \end{array}. \tag{11.10}$$

This walker ordinarily just follows a random subway line out of the present station. The exception is at the Longueuil Université de Sherbrooke station.

Our walker sometimes decides to remain there (for coffee). Each row of P has nonnegative values that sum to one, as they must because the walker has to be somewhere at the next time step. The columns of a transition matrix do not necessarily sum to one, and indeed they don't for this P.

If the walker goes through 100 steps, then the transition matrix relating the final position to the starting position is

$$P^{100} \doteq \begin{array}{c|ccccc} & \text{JT} & \text{S} & \text{LG} & \text{B} & \text{L} \\ \hline \text{JT} & 0.1546 & 0.1530 & 0.2319 & 0.3063 & 0.1541 \\ \text{S} & 0.1530 & 0.1547 & 0.2296 & 0.3091 & 0.1536 \\ \text{LG} & 0.1546 & 0.1530 & 0.2319 & 0.3064 & 0.1541 \\ \text{B} & 0.1532 & 0.1546 & 0.2298 & 0.3089 & 0.1536 \\ \text{L} & 0.1541 & 0.1536 & 0.2311 & 0.3073 & 0.1539 \end{array},$$

where the approximation is in rounding the entries to four places. The probability that X_{100} is Jean-Talon given X_0 ranges from about 0.153 (when X_0 is Snowdon) to 0.1546 (when X_0 is Jean-Talon or Lionel-Groulx). The other four columns are similarly very nearly constant. As a result, the starting point X_0 has almost been forgotten by time 100. For this transition matrix, we will see that $\mathbb{P}_{x_0}(X_n = \omega_j)$ converges to a limit $\pi(\omega_j)$ as $n \to \infty$, which does not depend on the starting position x_0.

Berri-UQAM is clearly the favorite station. Exercise 11.1 asks you to investigate how $\mathbb{P}_{\text{Snowdon}}(X_n = \text{Berri-UQAM})$ develops as n increases. The trajectory of such a probability changes if the walker either ceases lingering at Longueuil or, on the other hand, takes up the habit of lingering at every station.

Whatever value X_0 had is nearly forgotten in X_{100}. Because the chain is homogenous, the value of X_{100} must be nearly forgotten by time 200. If we take a widely separated sequence of equispaced samples we should get a nearly IID sample. That property underlies Markov chain Monte Carlo, though as we will see, it is not necessary and not always wise to skip over the intermediate values.

The distribution π on Ω is a ***stationary distribution*** of the transition matrix P if $\pi(\omega) = \sum_{\widetilde{\omega}\in\Omega} \pi(\widetilde{\omega})P(\widetilde{\omega} \to \omega)$ holds for all $\omega \in \Omega$. If $X_0 \sim \pi$ then $X_i \sim \pi$ for all $i \geqslant 0$.

In matrix terms, stationarity means that $\pi = \pi P$. That is, π is a left eigenvector of Ω, with eigenvalue 1. Not just any left eigenvector can be a stationary distribution. We must have $\pi(\omega) \geqslant 0$ and $\sum_{\omega\in\Omega} \pi(\omega) = 1$.

Software for eigenvectors usually computes right eigenvectors. By writing $\pi^{\mathsf{T}} = P^{\mathsf{T}}\pi^{\mathsf{T}}$, we see that π^{T} is a right eigenvector of P^{T}. The eigenvalues of P^{T}, where P is the metro matrix (11.10) above are, to three decimal places,

$$1.000 \quad -0.946 \quad 0.551 \quad -0.203 \quad \text{and} \quad 0.098.$$

The first eigenvalue takes the desired value 1, and the corresponding eigenvector, arranged as a row, is

$$\boldsymbol{v} = \begin{pmatrix} -0.329 & -0.329 & -0.493 & -0.658 & -0.329 \end{pmatrix}.$$

Obviously $\boldsymbol{v}$ cannot be a stationary distribution, because it is negative. However if $\boldsymbol{v}$ is an eigenvector, then so is $-\boldsymbol{v}$. Either of $\pm\boldsymbol{v}$ is a valid result from an eigenvalue algorithm, so we must be prepared for the possibility of getting a negative vector. Even $-\boldsymbol{v}$ is not the stationary distribution because it does not sum to 1. The remedy to both of these problems is to normalize v, setting $\pi = \boldsymbol{v}/\sum_{j=1}^N v_j$. For the metro walk, this normalization yields

$$\pi \doteq \begin{pmatrix}0.154 & 0.154 & 0.231 & 0.308 & 0.154\end{pmatrix},$$

which sums to 1.001 due to rounding.

Every transition matrix on a finite state space has at least one stationary distribution as the next two results show.

Theorem 11.1 (Perron-Frobenius Theorem)**.** *Let $P \in [0,\infty)^{N\times N}$ with (possibly complex) right eigenvalues $\lambda_1,\dots,\lambda_N$. Let $\rho = \max_{1\leqslant j\leqslant N}|\lambda_j|$. Then P has an eigenvalue equal to ρ with a corresponding eigenvector v with all non-negative entries.*

Proof. See Serre (2002, Chapter 5), who calls this result the weak form of the Perron-Frobenius Theorem. □

Corollary 11.1. *Let P be a transition matrix on a finite state space. Then P has at least one stationary distribution π.*

Proof. Since P is a transition matrix it has nonnegative entries. So therefore does P^{T}. Let $\boldsymbol{v}$ be the nonnegative right eigenvector of P^{T} that Theorem 11.1 provides. Then $\pi = \boldsymbol{v}/\sum_{j=1}^N v_j$ is a stationary distribution for P. The denominator $\sum_{j=1}^N v_j$ is never zero because $v_j \geqslant 0$ and the definition of an eigenvector excludes the vector of all zeroes. □

Transition matrices on infinite state spaces do not necessarily have a stationary distribution. Consider for example, the chain on $\Omega = \{1,2,\dots\}$ in which the transitions are from ω to $\omega+1$ with probability 3/5 and from ω to $\max(\omega-1,0)$ with probability 2/5. This chain will wander off to positive infinity and fail to have a stationary distribution.

Simply having a stationary distribution is not enough. We need two further conditions to ensure that $X_n \overset{\mathrm{d}}{\to} \pi$ as $n\to\infty$. To understand those conditions, consider the transition matrices

$$P_1 = \begin{pmatrix}1/2 & 1/2 & 0 & 0\\ 1/2 & 1/2 & 0 & 0\\ 0 & 0 & 1/2 & 1/2\\ 0 & 0 & 1/2 & 1/2\end{pmatrix} \quad\text{and}\quad P_2 = \begin{pmatrix}0 & 0 & 1/2 & 1/2\\ 0 & 0 & 1/2 & 1/2\\ 1/2 & 1/2 & 0 & 0\\ 1/2 & 1/2 & 0 & 0\end{pmatrix}.$$

A chain with transition matrix P_1 will stay forever in either $\Omega_1 = \{\omega_1,\omega_2\}$ or $\Omega_2 = \{\omega_3,\omega_4\}$. Both $\boldsymbol{v}_1 = \begin{pmatrix}1/2 & 1/2 & 0 & 0\end{pmatrix}$ and $\boldsymbol{v}_2 = \begin{pmatrix}0 & 0 & 1/2 & 1/2\end{pmatrix}$ are stationary distributions of P_1. More generally, if $0\leqslant\theta\leqslant 1$ then $\theta\boldsymbol{v}_1 + (1-\theta)\boldsymbol{v}_2$

is a stationary distribution of P_1. Sets like Ω_1 and Ω_2 from which a Markov chain cannot escape are called ***closed sets***.

A chain with transition matrix P_2 will forever alternate between Ω_1 and Ω_2. It has a unique stationary distribution $\boldsymbol{u} = \begin{pmatrix}1/4 & 1/4 & 1/4 & 1/4\end{pmatrix}$. If $\boldsymbol{p}_0 = \boldsymbol{u}$ then $\boldsymbol{p}_n = \boldsymbol{u}$ for $n \geqslant 0$. But $\mathbb{P}_{\omega_1}(X_n = \omega) = 0$ if n is even and $\omega \in \Omega_2$ or if n is odd and $\omega \in \Omega_1$. As a result $\mathbb{P}(X_n = \omega)$ does not approach $1/4$ in this case.

The problem with P_1 is that the state space splits into two pieces that don't communicate with each other. The state x ***leads to*** the state y if $\mathbb{P}_x(X_n = y) > 0$ for some $n \geqslant 0$. The states x and y ***communicate*** if x leads to y and y leads to x. The transition matrix P is ***irreducible*** if x communicates with y whenever $x, y \in \Omega$. A chain with an irreducible transition matrix can get from any one state to any other. The only closed set for an irreducible transition matrix is Ω itself.

The problem with P_2 is that the transitions have a periodic behavior. For a transition matrix P, the state ω is ***periodic*** with period $t \in \{2, 3, \dots\}$ if

1) $\mathbb{P}_\omega(X_n = \omega) > 0$ only holds for $n = tm$ with integer $m \geqslant 0$, and

2) t is the largest member of $\{2, 3, \dots\}$ for which **1)** holds.

Put another way, the period t is the greatest common divisor (GCD) of $\{n \geqslant 1 \,|\, \mathbb{P}_\omega(X_n = \omega) > 0\}$. If all states ω have period $t \geqslant 2$ then we say that P has period t. The example matrix P_2 has period 2.

A state that is not periodic is ***aperiodic***. For an aperiodic state we ordinarily find that the GCD above is 1 and we might say that the state has period 1. If the state ω has period 1, that does not imply that a self transition $\omega \to \omega$ is possible. If, for example, the smallest two $n \geqslant 1$ with $\mathbb{P}_\omega(X_n = \omega) > 0$ are 19 and 20, then ω is aperiodic. Some transition matrices have states where $\mathbb{P}_\omega(X_n = \omega) = 0$ for all $n \geqslant 1$. These states are also aperiodic. We might say that they also have period 1, with or without referring to the GCD of $\varnothing$. A transition matrix is aperiodic, if every state is aperiodic.

Theorem 11.2. *If the transition matrix P is irreducible and aperiodic and has stationary distribution π, then for all $\omega_0 \in \Omega$,*

$$\lim_{n\to\infty} \mathbb{P}_{\omega_0}(X_n = \omega) = \pi(\omega). \tag{11.11}$$

Proof. This follows from Theorem 1.8.3 of Norris (1998). □

If Ω is finite and P is irreducible and aperiodic then Corollary 11.1 supplies the stationary distribution π needed for Theorem 11.2. Theorem 11.2 also holds for infinite state spaces, though as we have seen, not all infinite transition matrices have stationary distributions.

We cannot have two different values for the limit in (11.11). Therefore Theorem 11.2 shows that the stationary distribution π is unique when P is irreducible and aperiodic. In fact, uniqueness of π holds when P is irreducible whether it is periodic or not (Durrett, 1999, §1.4).

Theorem 11.3. *Let X_i be a time-homogenous Markov chain on a finite set Ω with transition matrix P and stationary distribution π. If P is irreducible, then*

$$\mathbb{P}_{\omega_0}\biggl(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}f(X_i)=\sum_{\omega\in\Omega}\pi(\omega)f(\omega)\biggr)=1$$

holds for any starting state $\omega_0 \in \Omega$ and any real-valued function f on Ω.

Proof. This follows from Levin et al. (2009, Chapter 4.7). □

Theorem 11.3 is a law of large numbers for Markov chains. It shows that long term averages of the chain settle down to the corresponding expectations under the stationary distribution. A Markov chain with this property is often called ***ergodic***. The usage is not universal. Sometimes that term is used for irreducible chains, especially in physics.

In MCMC we have to choose a transition matrix P in order to get the desired asymptotic behavior. In §11.4 we will see how to pick P with a desired stationary distribution π. We also want to avoid reducible P and periodic P.

We would not deliberately choose a matrix P that has a closed set (other than Ω of course). But some transition matrices have sets that are effectively closed. A tiny example is $\bigl(\begin{smallmatrix}1-\epsilon & \epsilon\\ \epsilon & 1-\epsilon\end{smallmatrix}\bigr)$ for $0 < \epsilon < 1$. This matrix is irreducible and aperiodic and Theorem 11.3 applies to it. But Theorem 11.3 describes the limit as $n \to \infty$. That limit does not provide a very good description of the sample behavior when $n\epsilon$, the expected number of state changes, is negligible. We will see more complicated examples where Ω contains states that just rarely communicate.

As for periodicity, there are very natural Markov chains that exhibit it. Suppose for example that we form a graph whose nodes represent movies and actors, and the only edges connect actors to movies in which they have appeared. This is a bipartite graph and the random walk on any bipartite graph has period 2. There is an important fact about states that lets us remedy a periodic transition matrix.

Theorem 11.4. *Let P be an irreducible transition matrix on the state space Ω. If any one state $\omega \in \Omega$ is aperiodic, then P is aperiodic. If any one state $\omega \in \Omega$ has period $t \geqslant 2$ then P has period t.*

Proof. Citation to go here! □

If P is irreducible and $P(\omega \to \omega) > 0$ for some $\omega \in \Omega$ then ω is aperiodic. Therefore by Theorem 11.4, P is aperiodic. As an example, the metro walk above would have been periodic except that the walker sometimes lingered at the Longueuil station. Many of the chains used in MCMC have the property of lingering at one or more states. This has the benefit of removing periodicity, although of course we don't want them to linger too much because then the chain would not explore the state space very quickly.

Periodicity is a less severe flaw than reducibility. A Markov chain does not have to be aperiodic for the law of large numbers (Theorem 11.3) to hold. For example, if we alternately sample movies that a given actor appeared in and actors that a given movie employed, then half of the sample values will be actor nodes and half will be movie nodes (for even n). The stationary distribution for such a walk also puts half of its probability on actor nodes and half on movie nodes. Theorem 11.3 then tells us that the Markov chain will also apply proper weighting within the sets of movie and actor nodes.

11.3 Detailed balance

We are interested in pairs P and π for which the transition matrix P has stationary distribution π. This is a joint property of the pair (P, π). For an irreducible P there is only one π. For a given stationary distribution π, there can be many irreducible P.

The stationarity condition can be written

$$\sum_{x\in\Omega} \pi(x)P(x \to y) = \pi(y) = \sum_{x\in\Omega} \pi(y)P(y \to x) \tag{11.12}$$

for all $y \in \Omega$. The first equality in (11.12) is the definition, $\pi P = \pi$, of stationarity. The second equality multiplies $\pi(y)$ by $\sum_x P(y \to x)$ which equals 1.

Fixing y and subtracting $\pi(x)P(x \to x)$ from both sides of (11.12) yields

$$\sum_{x:x\neq y} \pi(x)P(x \to y) = \sum_{x:x\neq y} \pi(y)P(y \to x). \tag{11.13}$$

Equation (11.13) may be interpreted as a balance condition. The left side shows the probability flowing into y from other states. The right side shows the probability flowing out of y to other states. In equilibrium, those two flows are equal, or balanced.

A sufficient condition for equation (11.13) is that

$$\pi(x)P(x \to y) = \pi(y)P(y \to x), \quad \forall x, y \in \Omega. \tag{11.14}$$

Equation (11.14) is the ***detailed balance*** condition. The probability flow within any pair x and y is the same in both directions.

Suppose that a chain has detailed balance and that $X_0 \sim \pi$. Then

$$\begin{aligned}
\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) &= \pi(x_1)\prod_{j=2}^{n} P(x_{j-1} \to x_j)\\
&= \pi(x_1)\prod_{j=2}^{n} \frac{\pi(x_j)P(x_j \to x_{j-1})}{\pi(x_{j-1})}\\
&= \pi(x_n)\prod_{j=2}^{n} P(x_j \to x_{j-1})\\
&= \mathbb{P}(X_1 = x_n, X_2 = x_{n-1}, \dots, X_n = x_1). \qquad (11.15)
\end{aligned}$$

The probability of observing the sequence $x_1, \dots, x_n$ is the same whether we observe the chain in its original order or in the reverse order. This is why a chain with detailed balance is said to be ***reversible***. It is important to start the chain with $X_0 \sim \pi$. Without that condition, equation (11.15) need not hold.

Detailed balance makes it easy to study the relationship between P and π. In the examples below, the stationary distribution can be shown to take a simple form.

Example 11.1 (Random walk on graph)**.** Consider a simple random walk on an undirected graph G with no loops or multiple edges. The graph has N nodes. The incidence matrix $A \in \{0,1\}^{N\times N}$ has $A_{jk} = A_{kj} = 1$ if an edge connects nodes j and k and $A_{jk} = A_{kj} = 0$ otherwise. Node j has degree $d_j = \sum_{k=1}^N A_{jk}$ and we assume that every $d_j > 0$. The transition matrix P for this random walk has $P(j \to k) = A_{jk}/d_j$. Let $D = \sum_{j=1}^N d_j$. By Corollary 11.1 we know that P has a stationary distribution. One such stationary distribution is $\pi(j) \propto d_j$. To see this, let $\pi(j) = d_j/D$. Now

$$\pi(j)P(j \to k) = \frac{d_j}{D}\frac{A_{jk}}{d_j} = \frac{A_{jk}}{D} = \frac{A_{kj}}{D} = \pi(k)P(k \to j).$$

Because π and P satisfy detailed balance we know that π is a stationary distribution for P. If the graph is connected, then P is irreducible and π is the only stationary distribution.

Example 11.2 (Doubly stochastic matrix)**.** The N by N matrix P is ***doubly stochastic*** if $\sum_i P_{ij} = \sum_j P_{ij} = 1$ and all $P_{ij} \geqslant 0$. A doubly stochastic matrix P satisfies detailed balance with the uniform distribution $\pi(i) = 1/N$. Therefore the uniform distribution is a stationary distribution for any Markov chain with a doubly stochastic transition matrix. As a special case, suppose that $P = P^{\mathsf{T}}$. Then P is doubly stochastic. Any symmetric transition matrix has the uniform distribution as a stationary distribution.

11.4 Metropolis-Hastings

We are going to construct a Markov chain making sure that it has π as a stationary distribution. The method is called the Metropolis-Hastings algorithm. It is named after Metropolis et al. (1953), which was a major breakthrough in Monte Carlo methods and Hastings (1970), which was a significant generalization. The original Metropolis algorithm, in §11.5, remains an important special case of Metropolis-Hastings.

The transitions in Metropolis-Hastings work as follows. We start with some point x_0, whether deterministic or randomly sampled. For $i \geqslant 0$, given that $X_i = x$ we sample a random proposal Y from a distribution $\mathbb{P}(Y = y \mid X = x) = Q(y \mid x)$. To remind us about the directionality we write $Q(x \to y)$. This proposal y is then either accepted or rejected. With probability $A(x \to y)$ we accept it and put $X_{i+1} = Y$. With probability $1 - A(x \to y)$ the proposal is rejected and then $X_{i+1} = X_i$.

As examples, the proposal Y might be a perturbation of x, either large or small, as in random walk Metropolis of §11.5. Alternatively, Y might be an independent draw from a heavy tailed distribution as in the independence sampler of §11.6.

Metropolis-Hastings is obviously quite similar to acceptance-rejection sampling. We saw in §4.7 how to use acceptance-rejection to get samples from one distribution by randomly accepting some that are generated from another distribution. Now we do it for a Markov chain.

The choice of Q and A determines the transition probability matrix P. For $x \neq y$ we find that $P(x \to y) = Q(x \to y)A(x \to y)$. Given a choice for Q we seek values for A that provide detailed balance. For $x \neq y$ we need

$$\pi(x)Q(x \to y)A(x \to y) = \pi(y)Q(y \to x)A(y \to x). \tag{11.16}$$

From (11.16) we see that if a given acceptance function A provides detailed balance then so will $A/2$. We would rather double A than halve it because accepting more transitions should make the chain move faster. What stops us from using arbitrarily large multiples of an A that satisfies (11.16) is that A is a probability. We need $A \leqslant 1$.

We want to choose $A(x \to y)$ and $A(y \to x)$ as large as possible subject to (11.16) and $\max(A(x \to y), A(y \to x)) \leqslant 1$. We need to consider the possibility that one or more of the six factors in (11.16) might be zero. For instance, it may be much simpler to allow $Q(x \to y) > 0$ when $\pi(y) = 0$ than to make up a complicated proposal that never does that. We will assume that the Markov chain starts at point x with $\pi(x) > 0$.

The current state x must have $\pi(x) > 0$. The starting point has $\pi(x) > 0$ and so does any other point that could have have accepted. The proposal y must have $Q(x \to y) > 0$ or we would not have made it. Combining these gives $\pi(x)Q(x \to y) > 0$.

Because $\pi(x)Q(x \to y) > 0$, we may write

$$A(x \to y) = \frac{\pi(y)}{\pi(x)}\frac{Q(y \to x)}{Q(x \to y)}A(y \to x). \tag{11.17}$$

Next write $A(y \to x) = \lambda\pi(x)Q(x \to y)$ for some $\lambda = \lambda(x, y) \geqslant 0$. Substituting this expression for $A(y \to x)$ into (11.17) we find that

$$A(x \to y) = \lambda\pi(y)Q(y \to x).$$

That is, the same $\lambda = \lambda(x, y) = \lambda(y, x)$ appears in both acceptance probabilities.

To maximize λ, we choose

$$\lambda \max\bigl(\pi(y)Q(y \to x), \pi(x)Q(x \to y)\bigr) = 1.$$

That gives

$$\lambda = \frac{1}{\max\bigl(\pi(y)Q(y \to x), \pi(x)Q(x \to y)\bigr)} < \infty. \tag{11.18}$$

Substituting λ from (11.18) into $\lambda\pi(y)Q(y \to x)$ yields

$$A(x \to y) = \min\left(1, \frac{\pi(y)Q(y \to x)}{\pi(x)Q(x \to y)}\right). \tag{11.19}$$

Equation (11.19) is the ***Metropolis-Hastings acceptance probability***. Given a desired stationary distribution π and a proposal mechanism Q, it shows how to construct the acceptance probability A in order to obtain detailed balance and thus arrange for π to be a stationary distribution.

The ratio inside the minimum in (11.19) has a factor $\pi(y)/\pi(x)$. Other things being equal, we favor moves to higher probability states y. The second factor is $Q(y \to x)/Q(x \to y)$. Other things being equal, we hesitate to move to y if it would be hard to get back to x.

If $\pi(x) = \pi_u(x)/Z$ for an unnormalized distribution π_u, then (11.19) can be replaced by

$$A(x \to y) = \min\left(1, \frac{\pi_u(y)Q(y \to x)}{\pi_u(x)Q(x \to y)}\right)$$

because the factor Z cancels from numerator and denominator. Thus Metropolis-Hastings sampling does not require that we can compute the partition function. As written, $A(x \to y)$ requires the normalized version Q of the proposal probability. We will see some special cases where Q can be used in unnormalized form too. In general, the proposals for y given x and x given y may be from different distributions that have different normalizing constants.

The Metropolis-Hastings update is sometimes derived assuming that $Q(x \to y)$ and $Q(y \to x)$ are either both 0 or both positive, for any given pair x, y. That is very good advice, though as we see above, it is not required for detailed balance. If we did have $Q(x \to y) > Q(y \to x) = 0$, then from x we would sometimes propose a value y with acceptance probability $A(x \to y) = 0$ by (11.19). This is inefficient, because the probability placed on those proposals could have instead been placed on proposals that had a chance to move the chain.

Another possible inefficiency is to have $Q(x \to y) > 0$ when $\pi(y) = 0$. All of those proposals are certain to be rejected. But making them does not violate detailed balance, and sometimes it is computationally easier to make such a proposal. If for example we are following a random walk inside a complicated domain, it may be difficult to construct a proposal constrained to that domain, but easy to check whether just one proposed point is in the domain. The Metropolis algorithm for the hard shell model, in §11.7, makes some proposals to impossible states.

The Metropolis-Hastings acceptance probability is not the only one that attains detailed balance. A proposal due to Barker (1965) is

$$\widetilde{A}(x \to y) = \frac{\pi(y)Q(y \to x)}{\pi(x)Q(x \to y) + \pi(y)Q(y \to x)}. \tag{11.20}$$

Exercise 11.8 asks you to show that $\widetilde{A}(x \to y)$ satisfies detailed balance and that $\widetilde{A}(x \to y) \leqslant A(x \to y)$ holds for any $x \neq y$.

For $x \neq y$, maximizing $A(x \to y)$ for the given $Q(x \to y)$ has the result of maximizing $P(x \to y)$. Maximizing $P(x \to y)$ can be motivated by Peskun's Theorem, which appeared shortly after Hastings (1970).

Theorem 11.5 (Peskun's Theorem)**.** *Let P and $\widetilde{P}$ be irreducible $M \times M$ transition matrices, that both satisfy detailed balance for the same stationary distribution π. Suppose that $\widetilde{P}(x \to y) \leqslant P(x \to y)$ holds for all $x \neq y$. For $i \geqslant 1$, let X_i be sampled from the transition matrix P starting at x_0. Similarly, for $i \geqslant 1$, let $\widetilde{X}_i$ be sampled from the transition matrix $\widetilde{P}$ starting at $\widetilde{x}_0$. Then*

$$\lim_{n\to\infty} n\mathrm{Var}\Bigl(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\Bigr) \leqslant \lim_{n\to\infty} n\mathrm{Var}\Bigl(\frac{1}{n}\sum_{i=1}^{n} f(\widetilde{X}_i)\Bigr).$$

Proof. Peskun (1973). □

The Metropolis-Hastings acceptance probability (11.19) is a default choice for $A(x \to y)$ that provides detailed balance. With this default in hand we can then search for good proposal distributions $Q(x \to y)$ to suit any given problem.

We should bear in mind that detailed balance does not mean that the chain will mix quickly or even that it is irreducible. To take an extreme example, Metropolis-Hastings based on the proposal with $Q(x \to x) = 1$ for all x has detailed balance, but clearly goes nowhere. At the other extreme, Exercise 11.5 considers proposals $Q(x \to y) = \pi(y)$, which are much better than the proposals we can usually make.

It is extremely important that when the proposal y_{i+1} is rejected, giving $x_{i+1} = x_i$, that the repeated value be counted again in the average in equation (11.1). Every once in a while, the Metropolis-Hastings algorithm is wrongly implemented where only newly accepted values get counted into the sum. That is a beginner's error, not a valid alternative. Those repetitions may seem inefficient but they apply a necessary reweighting to the generated points. The original Metropolis article mentions this point at least twice, and other authors warn of it from time to time.

A simple example to remember this point is based on a Markov chain with just two states and transition matrix

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix},$$

for $\alpha, \beta \in (0,1)$. It is easy to verify that this matrix has stationary distribution $\pi = \bigl(\beta/(\alpha+\beta) \quad \alpha/(\alpha+\beta)\bigr)$. If we only count the times when the state changes, then for even n half of the sampled X_i will be in each state, and this would be an error if $\alpha \neq \beta$.

The straightforward way to implement the acceptance or rejection is by sampling $U \sim \mathbf{U}(0,1)$ and accepting $Y = y$ if and only if $U \leqslant A(x \to y)$. The acceptance probability is $A = \min(R, 1)$, where $R(x \to y) = \pi_u(y)Q(y \to x)/(\pi_u(x)Q(x \to y))$. If $R > 1$, then of course $U \leqslant R$ and so we can simplify

Algorithm 11.1 Metropolis-Hastings update.

```
MHup( $x$, $\pi_u$, $Q$ )

// $x$ is the current point, having $\pi_u(x) > 0$
// $\pi_u$ is a possibly unnormalized stationary distribution
// $Q$ is a proposal distribution

$y \sim Q(x \to \cdot)$
$R = \frac{\pi_u(y)Q(y\to x)}{\pi_u(x)Q(x\to y)}$
$U \sim \mathbf{U}(0,1)$
if $U \leqslant R$ then
  return $y$
else
  return $x$
```

the decision to one that accepts when

$$U \leqslant R(x \to y) = \frac{\pi_u(y)Q(y \to x)}{\pi_u(x)Q(x \to y)}. \tag{11.21}$$

The quantity $R(x \to y)$ is called the ***Hastings ratio***.

A single Metropolis-Hastings update is shown in Algorithm 11.1. In Metropolis-Hastings sampling, this update is repeated with each delivered point being used as the first argument to the next call of MHup. Taken literally, Algorithm 11.1 recomputes $\pi_u(x)$ at each proposal. If that step is expensive we can simply save the old value for reuse.

In statistical applications, π_u commonly contains a likelihood that is a product of many factors. We might then find that one or both of $\pi_u(y)$ and $\pi_u(x)$ underflow to zero, or overflow to ∞, obscuring the Hastings ratio R. It is then more numerically robust to compute $\log(R)$ and accept y if $\log(U) \leqslant \log(R)$.

The full algorithm needs a starting point x_0 with $\pi(x_0) > 0$ as well as a stopping rule. The simplest stopping rule is just to run for some number n of updates as in Algorithm 11.2.

It might be desirable to run the chain for some fixed amount of elapsed time. That choice has the potential to bias the results if proposals and probability evaluations take different amounts of time in different parts of the state

Algorithm 11.2 Metropolis-Hastings sampler.

```
Given $\pi_u$, $Q$, $x_0$ with $\pi_u(x_0) > 0$ and integer $n \geqslant 1$

for $i = 1$ to $n$ do
  $x_i \leftarrow \mathrm{MHup}(x_{i-1}, \pi_u, Q)$
return $x_1, \dots, x_n$
```

space. Similarly we might run the chain until some convergence diagnostic is satisfactory, but that choice has the potential to introduce subtle biases.

Next we consider and illustrate some of the proposal mechanisms used in Metropolis-Hastings. One of the simplest is random walk Metropolis that we describe in §11.5. There the proposed y is simply the present x plus some IID random offset. Perhaps still simpler is the independence sampler of §11.6, where the proposals themselves are IID. In most problems requiring MCMC the state x is a vector that may well have a very high dimension. In such cases one might want to change only a subset of those components in each proposal. There are many samplers of that kind, they raise special issues, and most of their discussion is saved for Chapter 12.

11.5 Random walk Metropolis

Recall that a random walk is a process where the increments $\boldsymbol{z}_i = \boldsymbol{x}_i - \boldsymbol{x}_{i-1}$ are independent and identically distributed. In ***random walk Metropolis*** (RWM) the proposals take the form $\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{z}_i$ where $\boldsymbol{z}_i$ are independent and identically distributed random vectors.

Suppose that $\boldsymbol{z} \sim Q$. Then $Q(\boldsymbol{x} \to \boldsymbol{y})$ can be written $Q(\boldsymbol{y} - \boldsymbol{x})$ where we now use Q to also represent the probability mass function of $\boldsymbol{z}$. The acceptance probability in RWM is

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\left(1, \frac{\pi_u(\boldsymbol{y})Q(\boldsymbol{x}-\boldsymbol{y})}{\pi_u(\boldsymbol{x})Q(\boldsymbol{y}-\boldsymbol{x})}\right).$$

We will focus on random walks where the distribution Q is symmetric: $Q(\boldsymbol{z}) = Q(-\boldsymbol{z})$. For symmetric random walks we get

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\left(1, \frac{\pi_u(\boldsymbol{y})}{\pi_u(\boldsymbol{x})}\right) \tag{11.22}$$

because the Q ratio cancels.

To illustrate RWM we consider a continuous problem. The stationary distribution is a probability density function $\pi(\boldsymbol{x})$ on $\boldsymbol{x} \in \mathbb{R}^d$ for an integer $d \geqslant 1$. The proposal generator Q is now taken to be a symmetric probability density function on $\mathbb{R}^d$ such as $\mathcal{N}(0, \sigma^2 I)$ or $\mathbf{U}[-\sigma, \sigma]^d$ for some scale parameter $\sigma > 0$. In moving from a discrete state space to a continuous one, we replace probability mass functions by probability density functions. The move leads to some technicalities that we postpone discussing to §**??**. For now, we sample continuous random variables but use the discrete case to form intuition.

Suppose that σ in our random walk proposal is very small compared to the distance scale over which π_u varies. Then $R(\boldsymbol{x} \to \boldsymbol{y}) = \pi_u(\boldsymbol{y})/\pi_u(\boldsymbol{x})$ will be very close to 1 and most proposals will be accepted. However small proposals mean that the Markov chain will move only very slowly and not explore the space. Conversely, suppose that σ is so large that it is much greater than the diameter of some set $S \subset \mathbb{R}^d$ that contains nearly all of the distribution π.

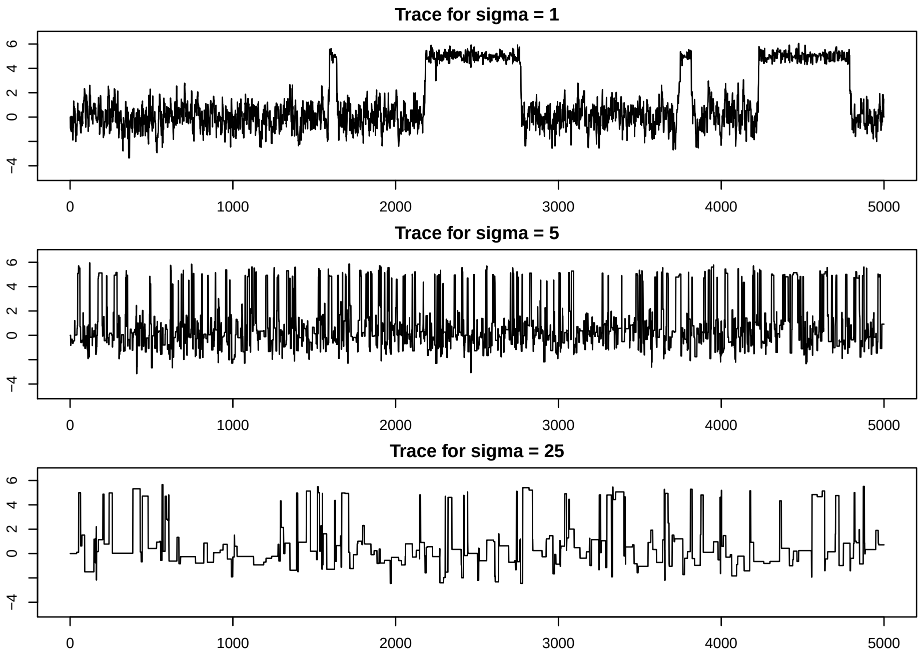

Figure 11.4: This figure shows traces x_i for $i = 1, \dots, 5000$ of random walk Metropolis for $\pi(x)$ given by (11.23), starting at $x_0 = 0$, using $\sigma \in \{1, 5, 25\}$.

Then if $\boldsymbol{x}$ is well inside S and has a large $\pi(\boldsymbol{x})$ we could find that proposals land outside S where π is negligible and get rejected. Once again we would have a Markov chain that does not explore the space, because it gets stuck for many consecutive iterations. The best scale for RWM is a compromise, not too large and not too small.

The next example is simple enough that we would not need MCMC but it serves to illustrate the scaling compromise. For $x \in \mathbb{R}$ we take

$$\pi(x) = \theta\varphi(x) + (1-\theta)\frac{1}{\tau}\varphi\Bigl(\frac{x-\mu}{\tau}\Bigr), \tag{11.23}$$

a mixture of $\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,\tau^2)$ distributions, using $\theta = 5/6$, $\mu = 5$ and $\tau^2 = 1/9$. We start with $x_0 = 0$ and run RWM for $n = 5000$ iterations, using $\sigma = 1$, $\sigma = 5$ and $\sigma = 25$. Figure 11.4 plots x_i versus i for these simulations. These plots are called ***trace plots*** and more general trace plots may show $f(x_i)$ versus i for some function f of particular interest.

A walk with $\sigma = 1$ is not very effective. That walk did not even find the second mode of π (for x near $\mu = 5$) within its first 1000 sample values. When it did find that mode it stayed for a long time before returning to the larger mode for x near 0. A walk with $\sigma = 25$ is also problematic. About 95% of its proposals are in the range $[-50, 50]$. We know from (11.23) that such a large range of proposals will usually generate a value of y well into the region where

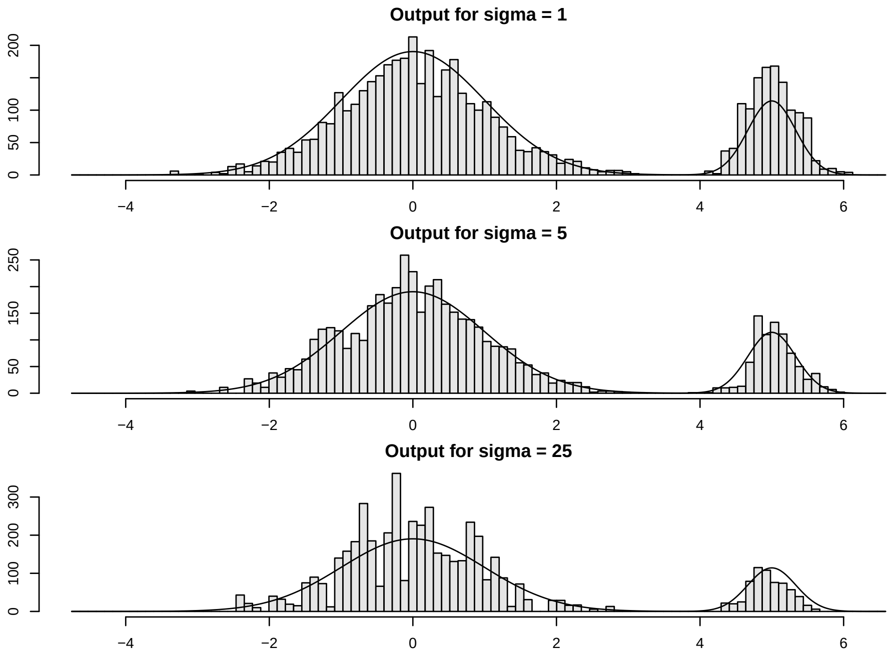

Figure 11.5: This figure shows histograms of x_i sampled by random walk Metropolis for $\pi(x)$ given by (11.23), starting at $x_0 = 0$, and using $\sigma \in \{1, 5, 25\}$. The target density π, multiplied by n, is superimposed.

π is negligible, triggering a rejection. When π is given just by a black box, we might well find ourselves using too large a scale. The third trace in Figure 11.4 has some prominent flat spots where the Markov chain got stuck for a while. The middle trace, using $\sigma = 5$ looks better than the other two.

Figure 11.5 shows histograms of the points x_i from these three RWM runs. The choice $\sigma = 1$ gives fairly nicely Gaussian bumps for the two modes but it has them with the wrong relative size. Because it takes long sojourns with few rejections within each mode, the modes have been well sampled. Unfortunately it seldom switches from mode to mode. Then the fraction of time it spends in each mode becomes unreliable. The choice $\sigma = 25$, has done a better job on the relative probabilities of the two modes but the histogram in the left mode is very spiky due to all of the rejected proposals made from there.

If the consecutive values of an MCMC tend to be very close to each other, then it is a sign that the simulation is not moving quickly through its space. We can quantify this phenomenon by computing the sample correlations between the output values x_i and lagged values $x_{i+\ell}$. The ***sample autocovariance*** of

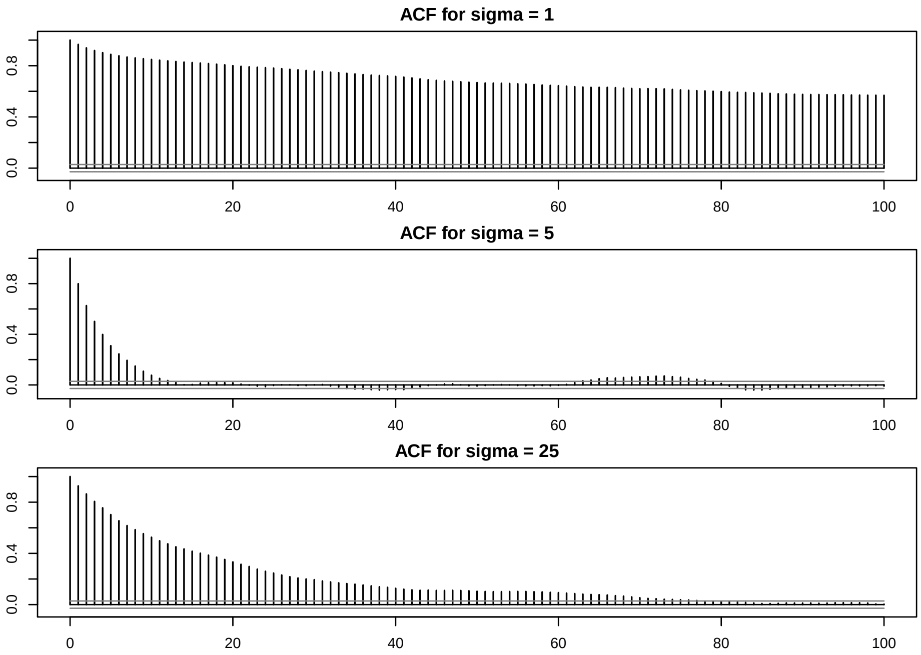

Figure 11.6: This figure shows autocorrelation functions of x_i sampled by random walk Metropolis for $\pi(x)$ given by (11.23), starting at $x_0 = 0$, and using $\sigma \in \{1, 5, 25\}$. Each panel has horizontal bars at $\pm 2/\sqrt{n}$. Uncorrelated data would have a sample correlation with a standard deviation of roughly $1/\sqrt{n}$.

x_i at lag ℓ for $0 \leqslant \ell < n$ is

$$\hat\gamma_\ell = \frac{1}{n}\sum_{i=1}^{n-\ell+1}(x_i - \bar x)(x_{i+\ell} - \bar x), \quad \text{for } \bar x = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The denominator in $\hat\gamma_\ell$ is n instead of the $n - \ell$ that one might expect. This convention gives a more stable $\hat\gamma_\ell$ when ℓ/n is not negligible. For $-n < \ell < 0$, let $\hat\gamma_\ell = \hat\gamma_{|\ell|}$. The ***sample autocorrelation*** (ACF) of x_i at lag ℓ is

$$\hat\rho_\ell = \frac{\hat\gamma_\ell}{\hat\gamma_0}.$$

Figure 11.6 shows the sample ACF of the random walk Metropolis samplers in our example. For $\sigma = 5$, the autocorrelations decrease quickest and at a lag just over $\ell = 10$ we see negligible sample correlation between x_i and $x_{i+\ell}$. The ACF at $\sigma = 25$ decays more slowly and the one for $\sigma = 1$ decays slowest of all. We can see from Figure 11.4 that the $\sigma = 1$ Markov chain only rarely moved from one mode to the other and the slowly decaying ACF reflects this.

Next we consider the population counterpart to $\hat\rho_\ell$. If x_0 has been sampled from π then our sample values would be stationary, meaning that

$$(x_i, x_{i+1}, \dots, x_{i+\ell}) \overset{\mathrm{d}}{=} (x_{i+k}, x_{i+k+1}, \dots, x_{i+k+\ell})$$

for any $k \geqslant 0$ and $\ell \geqslant 0$. Here $\overset{\mathrm{d}}{=}$ means having the same distribution. Then all of the x_i have the distribution π and if $\mathbb{E}(x_i^2) < \infty$ there is a well defined correlation $\mathrm{Corr}(x_i, x_{i+\ell})$ that we label ρ_ℓ. When we are running an MCMC it is usually because we could not start with $x_0 \sim \pi$. However, in a well mixing Markov chain the samples quickly approach the stationary distribution where ρ_ℓ is a meaningful quantity.

In this very simple problem we were able to explore various choices of σ and see which worked better. We knew that there were only two modes and so we knew that the first 1000 x_i from $\sigma = 1$ were not very good and we also knew that there was no undiscovered third mode to worry about after the first 5000 x_i. If we only had a black box function to compute π, then we would not know whether σ was too small or too large or just right.

In many Bayesian applications there is underlying theory, based on the central limit theorem, to suggest that π should be dominated by one mode. Then the Markov chain needs to explore within that mode but does not have to take large enough steps to find any other modes. The problem is not quite a black box. If the acceptance rate is very small, we can infer that σ is too large and then try a smaller value. Conversely, a very high acceptance rate suggests that we should raise σ. Under a famous result from Gelman et al. (1996), it is optimal to tune σ so that about 23.4% of proposals are accepted.

Gelman et al. (1996) use strong assumptions to get their answer, however the range of near optimal acceptance rates is broad and we might expect a similar broad range in other problems. They consider random walk Metropolis when $\pi(\boldsymbol{x})$ is the $\mathcal{N}(0, I_d)$ distribution and the proposal is $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \sigma_d^2 I_d)$, so they can study how the proposal variance should depend on d. They approach the problem numerically for $d = 1$ and theoretically in the limit as $d \to \infty$. Their criterion is based on the variance of averages $(1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)$, and in their formulation, the answer hardly depends on f, within reason. They recommend $\sigma_d = 2.4$ for $d = 1$ and that closely matches their asymptotic result $\sigma_d = 2.38/\sqrt{d}$ for large d. Numerically obtained optimal σ_d are also quite close to $2.38/\sqrt{d}$ for $2 \leqslant d \leqslant 10$.

The acceptance rate when using the optimal σ_d is about 44% for $d = 1$ and decreases rapidly to a limiting value of about 23.4% as $d \to \infty$. As long as the rejection rate is between 15% and 40% the efficiency is close to that of the optimal σ. We can estimate the acceptance rate empirically from one or more pilot runs and adjust our choice of σ to get an acceptance rate in this range.

The efficiency of Metropolis versus IID sampling decreases sharply as d increases. For the Gaussian case, Gelman et al. (1996) show that the efficiency of MCMC with the optimal σ_d relative to IID sampling, is about $0.331/d$. Thus if n IID samples would have been enough then we need about $n/(0.331/d) \approx 3dn$ MCMC samples to compensate.

If $\pi \approx \mathcal{N}(\nu, \Sigma)$ for some positive definite matrix Σ, then a reasonable choice is to take $\boldsymbol{x}_0 = \nu$ and proposals $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \sigma_d^2 \Sigma)$. We can often estimate Σ by finding $\hat{\nu} = \arg\max_{\boldsymbol{x}} \pi(\boldsymbol{x})$ and then using

$$\widehat{\Sigma} = \Bigl(-\frac{\partial^2}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathsf{T}}} \log(\pi(\hat{\nu}))\Bigr)^{-1}.$$

For even modestly large d, seeking 23.4% acceptance leads to $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \sigma_d^2 \widehat{\Sigma}) = \mathcal{N}(\boldsymbol{x}, 2.38^2 \widehat{\Sigma}/d)$.

Another approach is to estimate Σ using the output $\boldsymbol{x}_i$. See §**??** on adaptive MCMC. The target π might not have tails as light as the Gaussian distribution. Just as in importance sampling, it is generally wiser to use heavier than Gaussian tails in the proposals. See §**??** for more.

If $\pi(\cdot)$ has numerous modes of unknown shapes, sizes and distances from each other, then finding the optimal σ is much harder. Problem 11.10 is about a synthetic example with $\boldsymbol{x} \in \mathbb{R}^2$ and

$$\pi_u(\boldsymbol{x}) = \max_{1 \leqslant j \leqslant 8} \exp\Bigl(-\frac{1}{2}\|\boldsymbol{x} - \theta_j\|^2\Bigr) \tag{11.24}$$

where θ_j is the j'th column of

$$\Theta = \begin{pmatrix} 9.11 & 7.89 & -0.24 & 0.50 & 1.41 & -7.97 & -6.50 & -4.21 \\ 4.82 & -2.69 & 1.22 & 0.08 & -0.97 & 2.97 & -5.33 & -0.11 \end{pmatrix}. \tag{11.25}$$

This $\pi_u(\boldsymbol{x})$ is proportional to the maximum of 8 Gaussian probability density functions. Were it a weighted average, we could easily apply mixture sampling and then MCMC would not be needed. The contours of $\pi_u(\boldsymbol{x})$ are shown in Figure 11.7. The centers roughly match bodies in the constellation Orion (not including Meissa). The upper right center corresponds to Betelgeuse and the lower left is Rigel. A good σ to sample π_u must trade off exploration within modes versus communication between modes.

11.6 Independence sampler

Perhaps the simplest proposal mechanism is to take independent and identically distributed proposals from some distribution Q that does not even depend on the present location $\boldsymbol{x}$. Then $Q(\boldsymbol{x} \to \boldsymbol{y})$ can simply be written $Q(\boldsymbol{y})$ for a probability mass or density function Q. The Metropolis-Hastings proposal for this ***independence sampler***, simplifies to

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\Bigl(1, \frac{\pi_u(\boldsymbol{y})}{\pi_u(\boldsymbol{x})} \frac{Q(\boldsymbol{x})}{Q(\boldsymbol{y})}\Bigr).$$

The independence sampler is also called the ***Metropolized independence sampler***.

If we only have an unnormalized proposal Q_u, then we can accept or reject via

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\left(1, \frac{\pi_u(\boldsymbol{y}) Q_u(\boldsymbol{x})}{\pi_u(\boldsymbol{x}) Q_u(\boldsymbol{y})}\right).$$

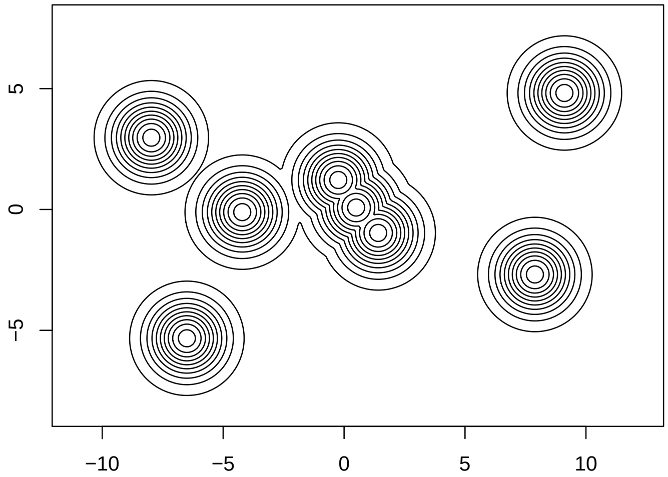

Figure 11.7: This figure shows 10 equispaced contours for the synthetic probability function $\pi(\boldsymbol{x})$ of equation (11.24). The displayed contour levels range from 5% to 95% of the maximum value of π.

We can use the independence sampler if we can sample from Q and compute both π_u and Q_u at any $\boldsymbol{x}$.

The key quantity in the independence sampler is the ***importance ratio*** $w(\boldsymbol{x}) = \pi(\boldsymbol{x})/Q(\boldsymbol{x})$. If $w(\boldsymbol{x})$ is large then it means that $\boldsymbol{x}$ is underrepresented when sampling from Q. The acceptance probability is then

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\Bigl(1, \frac{w(\boldsymbol{y})}{w(\boldsymbol{x})}\Bigr),$$

so that the chain prefers to move towards points with higher importance. We can use unnormalized π and/or Q in w without changing $w(\boldsymbol{y})/w(\boldsymbol{x})$.

A valid proposal distribution Q must sample any region that π does. To do otherwise would introduce bias. On the other hand if Q samples a region that π does not sample, then the consequence is milder. Proposals to move to such a region are automatically rejected and hence some computation is wasted.

The independence sampler is closely related to importance sampling. When we can sample from Q, we could also take observations $\boldsymbol{x}_i \sim Q$ and weight them proportionally to $w(\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)/Q(\boldsymbol{x}_i)$. Then the expected value of $f(\boldsymbol{x})$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^n \frac{\pi(\boldsymbol{x}_i)}{Q(\boldsymbol{x}_i)}\, f(\boldsymbol{x}_i) \tag{11.26}$$

if $w(\boldsymbol{x})$ is computable, or by

$$\sum_{i=1}^n \frac{\pi_u(\boldsymbol{x}_i)}{Q_u(\boldsymbol{x}_i)}\, f(\boldsymbol{x}_i) \Bigm/ \sum_{i=1}^n \frac{\pi_u(\boldsymbol{x}_i)}{Q_u(\boldsymbol{x}_i)} \tag{11.27}$$

if we have only an unnormalized π_u or only an unnormalized Q_u. A normalized version of w ordinarily requires normalized versions of both π and Q. If however, unnormalized π_u and Q_u share the same normalization constant so that $\pi_u(\boldsymbol{x})/Q_u(\boldsymbol{x}) = \pi(\boldsymbol{x})/Q(\boldsymbol{x})$, then we can use those unnormalized versions in (11.26). For an additional edge case see Exercise 11.6.

Plain importance sampling (11.26) weights $f(\boldsymbol{x}_i)$ by $w(\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)/Q(\boldsymbol{x}_i)$ to adjust for sampling from Q instead of π. The independence sampler draws proposals $\boldsymbol{y}_i \sim Q$ and applies random discrete weights equal to n_i/n where $n_i \geqslant 0$ is the number of times that proposal $\boldsymbol{y}_i$ got used. This could be 0 because $\boldsymbol{y}_i$ could have been rejected, or it could be quite large if $\boldsymbol{y}_i$ was accepted and $w(\boldsymbol{y}_i)$ was very large value. We define $\boldsymbol{y}_0 \equiv \boldsymbol{x}_0$ in order to include cases where $\boldsymbol{y}_1$ was rejected.

As in importance sampling, It is generally safer to have slightly heavy tails in an independence sampler proposal Q instead of light tails. Similarly to importance sampling there are advantages in having $w(\boldsymbol{x})$ bounded. A large $w(\boldsymbol{x})$ corresponds to a point that is very underrepresented in Q compared to π, and so the chain is likely to stay stuck there for a long time.

The independence sampler is often used in strategies that mix proposal types. See §11.9. For instance, an independence proposal with very large variance helps to ensure that the chain can reach more of the sample space.

The ***autoregressive sampler*** is an interesting hybrid between the independence sampler and the random walk sampler. In a d-dimensional problem, the proposal is

$$\boldsymbol{y}_i = \boldsymbol{c} + \Gamma(\boldsymbol{x}_i - \boldsymbol{c}) + \boldsymbol{z}_i$$

for a central point $\boldsymbol{c} \in \mathbb{R}^d$, a matrix $\Gamma \in \mathbb{R}^{d\times d}$ and some IID vectors $\boldsymbol{z}_i \in \mathbb{R}^d$. The center point $\boldsymbol{c}$ could be the mode of π or, more simply 0. The matrix Γ could be as simple as γI_d for a scalar γ. Then $\boldsymbol{y}_i = \boldsymbol{c} + \gamma(\boldsymbol{x}_i - \boldsymbol{c}) + \boldsymbol{z}_i$. The independence sampler has $\gamma = 0$ and random walk Metropolis has $\gamma = 1$. Taking $0 > \gamma \geqslant -1$ gives the proposals an antithetic property.

11.7 Random disks revisited

Now we return to the disk placement problem in Figure 11.1, with $N = 224$ disks inside the square $[0,1]^2$ with a periodic boundary. Metropolis et al. (1953) consider different sizes of disks. At one extreme the disks are so large that the only possible configurations are those that are near to a hexagonal grid. For such a small number of disks, a large proportion of them will be near the boundary of the square. The boundary effects would thus be far larger in the simulation than they would be in a real system. By using a periodic boundary, the disks near an edge jostle against other points that are about 15 disks away. This is a better approximation to what happens within an enormous system than a hard square barrier would be. The set $[0,1]^2$ with a periodic boundary is known as the ***flat torus***. It is like the familiar donut-shaped torus, but with different interpoint distances.

The maximum possible diameter for these points is about $d = 1/14$. Metropolis et al. (1953) consider diameters

$$d_0 = d(1 - 2^{\nu-8}), \quad \text{for } 0 \leqslant \nu \leqslant 7. \tag{11.28}$$

The very tightly packed example in the left panel of Figure 11.1 used $\nu = 3$, while the looser configuration on the right used $\nu = 6$.

A point $\boldsymbol{x}$ in this simulation is a 224×2 matrix with the i'th disk centered at (x_{i1}, x_{i2}). The desired distribution is

$$\pi(\boldsymbol{x}) \propto \pi_u(\boldsymbol{x}) = \begin{cases} 1, & \min\limits_{1\leqslant i<j\leqslant 224} d_T\bigl((x_{i1}, x_{i2}), (x_{j1}, x_{j2})\bigr) \geqslant d_0 \\ 0, & \text{else,} \end{cases}$$

where d_T is the distance between two points in the flat torus. In detail

$$d_T\bigl((x_1, x_2), (x_1', x_2')\bigr) = \sqrt{d_W(x_1, x_1')^2 + d_W(x_2, x_2')^2}$$

where for $x, x' \in [0, 1]$ their wraparound distance is

$$d_W(x, x') = \frac{1}{2} - \Bigl| |x - x'| - \frac{1}{2} \Bigr|. \tag{11.29}$$

Exercise 11.12 asks you to derive equation (11.29).

Metropolis et al. (1953) initialize their simulation at $\boldsymbol{x}_0$ which has the disks centered on a 'trigonal grid'. Their trigonal grid is almost hexagonal. They form 16 equispaced horizontal rows at height $i/16$ for $i = 1, \dots, 16$. The top row ($i = 16$) has points at $j/14$ for $j = 0, \dots, 13$. The second row from the top has points at $(j + 1/2)/14$ for $j = 0, \dots, 13$. The even numbered rows are all the same as the top row, and the odd numbered ones are the same as the second row.

Their basic proposal is to move one of the disk centers $\boldsymbol{x}_{i,1:2}$ for $i = 1, \dots, 224$ to $\boldsymbol{z} \equiv \boldsymbol{x}_{i,1:2} + \mathbf{U}[-\alpha, \alpha]^2$ where $\alpha = d - d_0$. The proposal $\boldsymbol{y}$ is the old $\boldsymbol{x}$ with row i replaced by $\boldsymbol{z}$. The Metropolis-Hastings acceptance probability is

$$\min\Bigl(1, \frac{\pi_u(\boldsymbol{y})}{\pi_u(\boldsymbol{x})}\Bigr) = \min(1, \pi_u(\boldsymbol{y})) = \pi_u(\boldsymbol{y}).$$

The acceptance probability is 1 if the newly moved point does not overlap any of the others and is 0 otherwise. They cycle through the 224 points in succession making a proposal for each one in turn.

Their algorithm is not the same as what we now call the Metropolis method. Now it would be more standard for a Metropolis method to choose a new $\boldsymbol{x}$ potentially moving all of the disks. In present terminology, the method they used is called Metropolis within Gibbs. See §12.10.

Their objective was to consider the density of neighboring disks at the boundary of any given disk and they are able to relate that density to numerous physical parameters of interest. Here we look at the distance from each disk to its nearest neighbor.

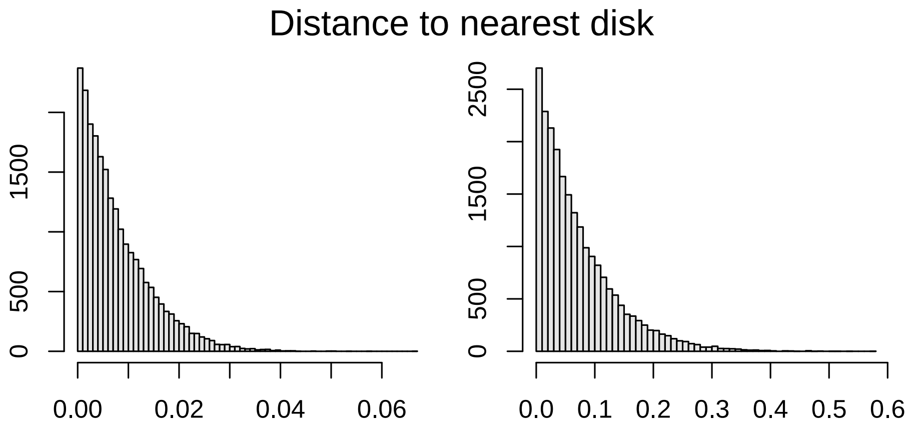

Figure 11.8: This figure shows histograms of the distance from each disk to its nearest neighbor, as a multiple of the disk diameter. On the left the diameters are given by (11.28) with $\nu = 3$ and on the right $\nu = 6$.

Figure 11.8 shows histograms of the distance from each disk to its nearest neighbor. The distances are reported in proportion to the disks' own diameter. For the tightly packed disks, most of them have a neighbor within 3% of their diameter. For the more loosely packed disks, the neighbors are much farther away on average. To make these histograms, the Markov chain was run for 1600 iterations computing nearest neighbor distances only at every 16'th iteration because the interpoint distance computation is relatively expensive. Each histogram then includes 22,400 distances. The distances look to have a roughly exponential distribution. The right tail is in fact somewhat thinner than the exponential distribution and of course the distribution is bounded.

Figure 11.9 shows some traces for this simulation. At each pass over the disks there are 224 accept or reject decisions. The fraction of acceptances is plotted against the iteration index $i = 1, \dots, 1600$. The probability of acceptance starts out high because the disks are initially very well separated. It quickly drops to around 40% and fluctuates in that range. The mean distance to a neighbor, over 224 disks, is plotted at every 16'th iteration.

The trace for mean distance at the tight spacing $\nu = 3$ appears to have a slight downward trend. Perhaps a longer simulation is in order. See Exercise 11.13. For $\nu = 3$ after 1600 steps, the original orientation of the points from the trigonal grid is still evident. In that tight spacing there is no way for the particles to push past each other. It seems reasonable that these potential flaws are not serious problems for the interpoint distances of interest there, though that decision might require input from domain experts.

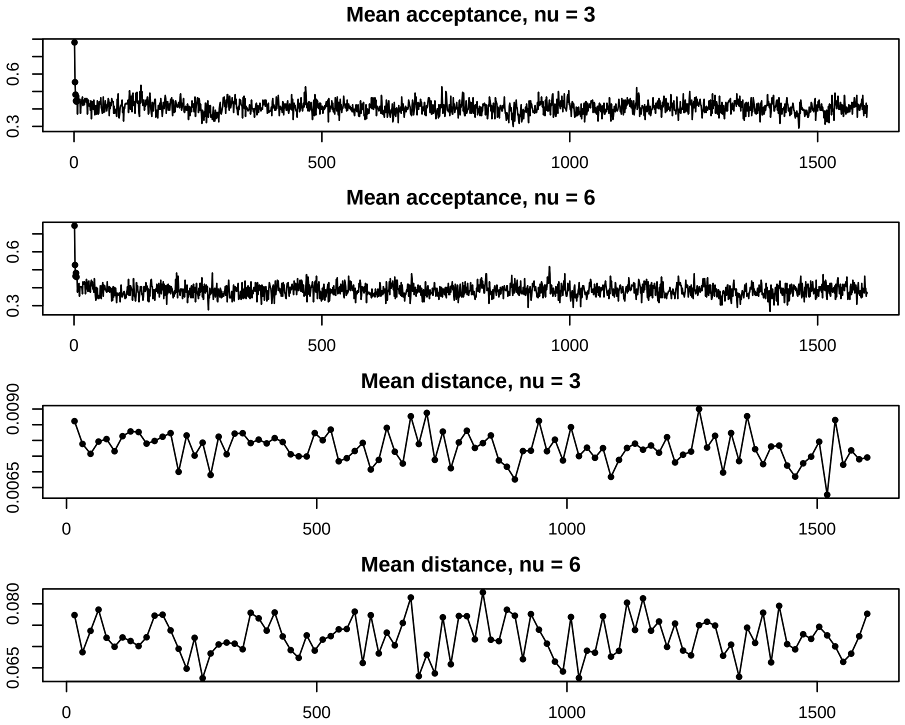

Figure 11.9: Here are four traces versus iteration for $1, 2, \dots, 1600$. All four curves depict averages over 224 disks. The top two panels have the mean acceptance rate for the $\nu = 3$ (dense) and $\nu = 6$ (sparse) simulations, with the first 5 points marked with a bullet. The next two panels show traces of the average (over 224 disks) of the distance to a nearest neighbor, thinned to every 16'th observation.

11.8 Ising revisited

Here we use a very simple proposal mechanism to get a rudimentary sampler for the Ising model. The proposal works by picking a site $\ell \sim \mathbf{U}\{1, \dots, N\}$ and changing x_ℓ to $-x_\ell$. That is $\boldsymbol{y}$ has $y_j = x_j$ for $j \neq \ell$ and $y_\ell = -x_\ell$, so that

$$Q(\boldsymbol{x} \to \boldsymbol{y}) = \begin{cases} 1/N, & \sum_{j=1}^N \mathbb{1}_{x_j \neq y_j} = 1 \\ 0, & \text{else.} \end{cases} \tag{11.30}$$

For this proposal $Q(\boldsymbol{x} \to \boldsymbol{y}) = Q(\boldsymbol{y} \to \boldsymbol{x})$.

The Metropolis-Hastings acceptance probability for this proposal is

$$A(\boldsymbol{x} \to \boldsymbol{y}) = \min\biggl(1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\biggr) = \min\biggl(1, \frac{\exp(-H(\boldsymbol{y})/T)/Z}{\exp(-H(\boldsymbol{x})/T)/Z}\biggr)$$

$$= \min\Bigl(1, \exp\Bigl(\frac{H(\boldsymbol{x}) - H(\boldsymbol{y})}{T}\Bigr)\Bigr).$$

The ratio $Q(\boldsymbol{x} \to \boldsymbol{y}) = Q(\boldsymbol{y} \to \boldsymbol{x})$ canceled out and so did the Z's. Now

$$\begin{aligned} H(\boldsymbol{x}) - H(\boldsymbol{y}) &= J \sum_{j:j\sim\ell} (y_j y_\ell - x_j x_\ell) + B(y_\ell - x_\ell) \\ &= -2Jx_\ell \sum_{j:j\sim\ell} x_j - 2Bx_\ell. \end{aligned}$$

Let $z_\ell = \sum_{j:j\sim\ell} x_j$ be the total spin of the neighbors of site ℓ. We accept $\boldsymbol{y}$ with probability

$$\min\bigl(1, \exp(-2x_\ell(Jz_\ell + B)/T)\bigr).$$

It is customary to measure the number of steps of the Metropolis-Hastings algorithm for the Ising model in terms of sweeps. One sweep corresponds to $N = L^2$ Metropolis-Hastings updates. Each site in the grid is visited, on average, once per sweep.

The images in Figure 11.2 were run with $J = 1$, $B = 0$ and $T \in \{8, 2.269, 2.0\}$. The image for $T = 8$ was sampled using 100,000 sweeps. The image for $T = 2.269$ was sampled using 10,000,000 sweeps. That run was done at the critical temperature where the algorithm above is known to be slow. The image for $T = 2.0$ was done with 1,000,000 sweeps.

The mean spin per site for state $\boldsymbol{x}$ is $\sum_{j=1}^N x_j/N$. Similarly, the mean energy per site is $H(\boldsymbol{x})/N$. Figure 11.10 shows trajectories of these quantities, sampled after each of the first 500 sweeps. There were 4 realizations of the Ising model at $T = 8$, $J = 1$, $B = 0$ with $L = 100$ and a periodic boundary. One run starts with all $x_j = 1$ so the mean spin is 1.0 and the energy is -2 (per site). It rapidly moves into a central region near mean spin 0 and energy just above -1. A second run starting with all $x_j = -1$ performs similarly. Two other runs start with spins that are independent $\mathbf{U}\{-1, 1\}$ random variables. Those starting values are draws from the Ising model with $T = \infty$. The runs from this hot starting distribution approach the same energy level as the other two runs. The average energy over 200,000 sweeps from the random start runs is -0.817. Although the two ground states have the lowest energy, and hence the highest probability, there are vastly more states at higher energy levels and so some higher energy levels are more probable than the minimal one.

When $T = 8$, the Ising simulations quickly reach a central region in the spin-energy plane, with average spin near 0 and average energy near -0.8. Once there they oscillate. We know by symmetry that the expected value (over $\boldsymbol{x} \sim \pi$) of the average (over $j = 1, \dots, N$) of the spin is 0. A more interesting quantity is $\mathbb{E}(f(\boldsymbol{x}))$ where $f(\boldsymbol{x}) = |(1/N)\sum_{j=1}^N x_j|$. This quantity is near zero at high temperatures, near 1 at low temperatures, and for very large N, has a sharp transition near T_c.

Figure 11.11 plots four ACFs from the Ising simulations of this section. The top row is at temperature $T = 8$, and it shows the ACF when $f(\boldsymbol{x})$ is the absolute mean spin (on the left) and the ACF when $f(\cdot)$ is the mean energy (on

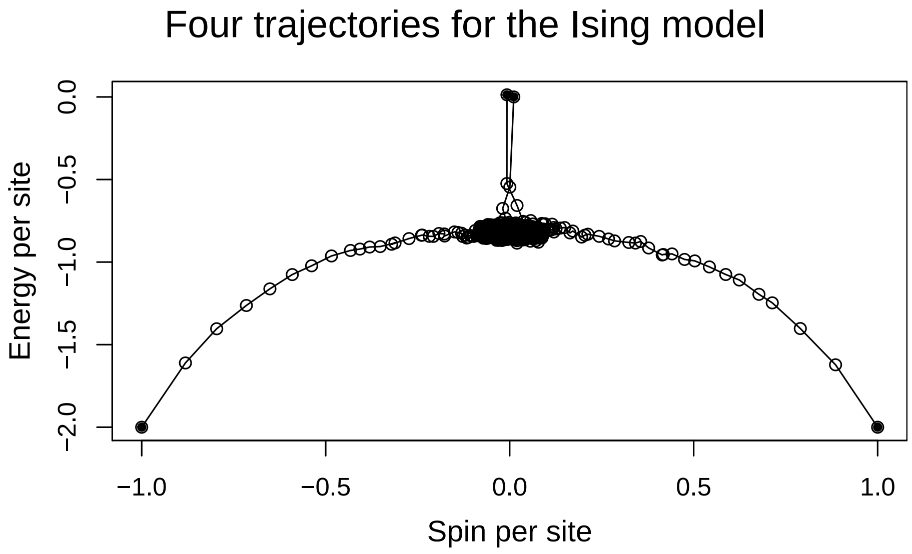

Figure 11.10: Mean energy versus mean spin for the Ising model with $J = 1$ and $B = 0$, temperature $T = 8.0$ on a 100×100 grid. Four trajectories of 500 sweeps are shown as described in the text. The starting points are solid.

the right). The autocorrelation drops off very quickly for energy and somewhat more slowly for the spin. At the critical temperature, the autocorrelations drop off much more slowly.

Even when we are interested in $|(1/N)\sum_{j=1}^N x_j|$, it is informative to plot the trace of $f(\boldsymbol{x}) = (1/N)\sum_{j=1}^N x_j$. Figure 11.12 plots this trace at ever 200'th sweep. We can see that over the course of 10^7 sweeps, the mean spin has moved back and forth between positive and negative values several times. Had the mean spin maintained the same sign over the whole simulation, then we would know that the space had not been well sampled.

When an MCMC is run with a large value of n and the autocorrelations are high then it can be reasonable to only compute $f(\boldsymbol{x}_i)$ on a subset of the values of i. For samples at temperature $T = 8$, the values were computed after each sweep, with one sweep corresponding to 10^4 Metropolis-Hastings steps. For $T = T_c$ the mean energy and spin values were recorded only at every 10'th sweep. The autocorrelations were thus available only at lags k that are multiples of 10 and only lags equal to multiples of 20 were actually plotted. The trace in Figure 11.12 was plotted with a point for every 400'th sweep. The original data $\boldsymbol{x}_{10i}$ for $i = 1, \dots, 10^6$ contained 38 times at which the simulation changed over from mostly positive to mostly negative. If we define a 'switch' to be an upcrossing above the level 0.5 followed by a downcrossing below -0.5, then there were 38 switches. It is possible that this number is an undercount due

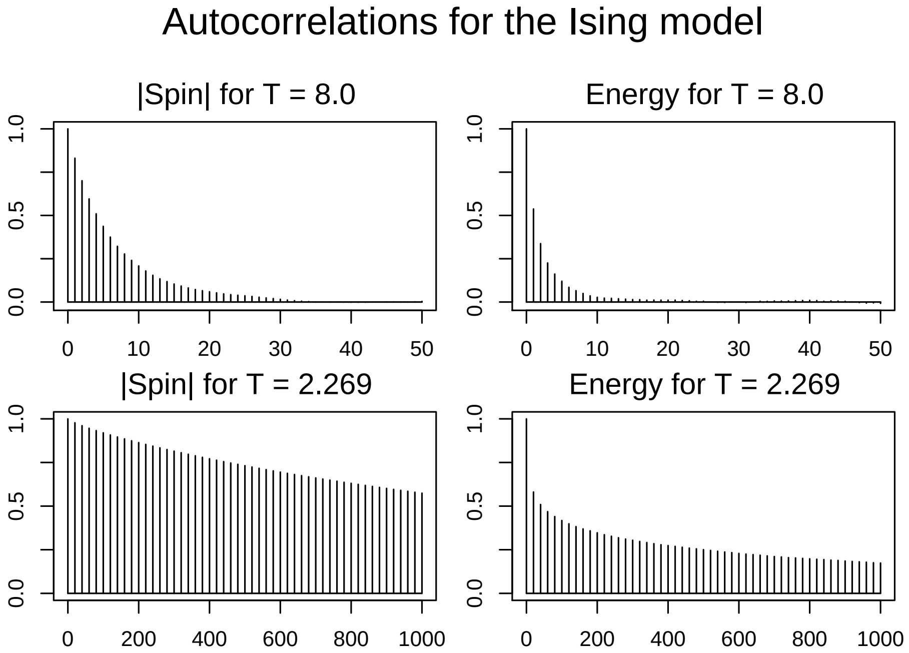

Figure 11.11: Autocorrelation functions for the Ising model at temperatures 8 and $T_c = 2.269$. The ACFs for mean absolute spin are on the left and the ACFs for mean energy are on the right. The lags for $T = 8$ go up to 50 while those for $T = T_c$ go up to 1000 in steps of 20.

to undetected upcrossings or downcrossings, but it is unlikely that many were missed because the chain is very sticky.

Looking at only every k'th output of a Markov chain is known variously as thinning or subsampling. See §11.13 for a discussion.

At the low temperature, $T = 2$, four simulations were done. One started at an image of all 1s, a second at all -1s, and two more started with IID $\mathbf{U}\{-1, 1\}$ images. All four ran for one million sweeps. None of them showed even modestly strong alternations. As a result, we must attach a caveat to Figure 11.2. The chain failed to sample anywhere near the all white image, so it clearly did not explore the stationary distribution well. At best the chain gave a reasonable sampling for nearly all black images at $T = 2$. But we cannot be sure.

While simple spin flip proposals illustrate some MCMC computations they are not the most powerful enough way to sample the Ising model at low temperatures. For some Monte Carlo problems, there is as yet no good solution. In the specific case of the Ising model on a square grid at low temperatures, however, there is a good method. The Swendsen-Wang algorithm §**??**, is efficient for this setting.

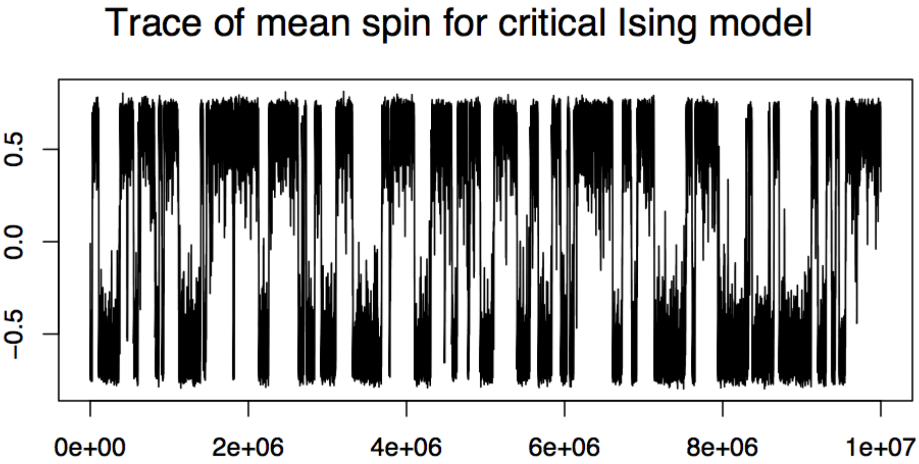

Figure 11.12: This figure shows the mean spin per site in $\boldsymbol{x}_i$, versus the simulation index i, after every 200'th sweep, for the Ising simulation at the critical temperature $T_c = 2.269$.

11.9 New proposals from old

Suppose that we have a list of M different proposal functions $Q_1, \dots, Q_M$ with corresponding acceptance rules $A_1, \dots, A_M$ and transition matrices $P_1, \dots, P_M$. Perhaps only one of these is really good, the others mix slowly, and we don't know which one is the good one. Then strategies that try them all will at least never miss the good one. It could also be that one transition matrix is good at jumping between modes of π while another one is good for exploring within each mode. In that case a strategy combining both could be better than using any single one of the transition matrices.

There are two main ways to proceed. We can sample proposals randomly and we can do them in sequence. We will use the following very simple results on random combinations later when we study proposal combinations.

Proposition 11.1. *Let $P_1, \dots, P_M$ be transition matrices for which π is a stationary distribution. Let $\alpha_m \geqslant 0$ for $m = 1, \dots, M$ satisfy $\sum_{m=1}^M \alpha_m = 1$. Then π is also a stationary distribution of the transition matrix $P_\alpha = \sum_{m=1}^M \alpha_m P_m$.*

Proof.

$$\pi P_\alpha = \sum_{m=1}^M \alpha_m \pi P_m = \sum_{m=1}^M \alpha_m \pi = \pi. \quad \square$$

The matrix P_α in Proposition 11.1 is the transition matrix of a chain that picks $m(i)$ with probabilty α_m and then transitions from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$ according to P_m. The P_m do not need to have detailed balance. If they do, then so does P_α.

Proposition 11.2. *If the transition matrices P_m in Proposition 11.1 satisfy detailed balance with respect to π, then so does P_α.*

Proof. Expand πP_α and apply $\pi(\boldsymbol{x})P_m(\boldsymbol{x}\to\boldsymbol{y}) = \pi(\boldsymbol{y})P_m(\boldsymbol{y}\to\boldsymbol{x})$. □

A second way to make a hybrid out of proposals P_1, P_2, through P_M is to make them in that order one after the other. If we do that, then the M-step path from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+M}$ has transition matrix $P_{1:M} \equiv P_M P_{M-1}\cdots P_3 P_2 P_1$. The rightmost transition matrix P_1 is applied first and P_M is applied last. We can consider a new Markov chain $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_{Mi}$ for $i \geqslant 0$ with transition matrix $P_{1:M}$. This matrix also preserves the stationary distribution but it does not generally satisfy detailed balance, even if all of the P_m do. There is an example with $M=2$ in §12.4 on the Gibbs sampler.

Instead of mixing the transition matrices we can mix the M proposals. We can sample $\boldsymbol{y}$ from $Q_\alpha(\boldsymbol{x}\to\cdot) = \sum_{m=1}^{m}\alpha_m Q_m(\boldsymbol{x}\to\cdot)$. Then, for $\boldsymbol{x}\neq\boldsymbol{y}$, the Metropolis-Hastings acceptance probability is

$$A_\alpha(\boldsymbol{x}\to\boldsymbol{y}) = \min\biggl(1, \frac{\pi(\boldsymbol{y})Q_\alpha(\boldsymbol{y}\to\boldsymbol{x})}{\pi(\boldsymbol{x})Q_\alpha(\boldsymbol{x}\to\boldsymbol{y})}\biggr)$$

and the transition matrix $\widetilde{P}_\alpha = Q_\alpha A_\alpha$ is

$$\begin{aligned}
\widetilde{P}_\alpha(\boldsymbol{x}\to\boldsymbol{y}) &= \min\biggl(Q_\alpha(\boldsymbol{x}\to\boldsymbol{y}), \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}Q_\alpha(\boldsymbol{y}\to\boldsymbol{x})\biggr)\\
&= \min\biggl(\sum_{m=1}^{M}\alpha_m Q_m(\boldsymbol{x}\to\boldsymbol{y}), \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\sum_{m=1}^{M}\alpha_m Q_m(\boldsymbol{y}\to\boldsymbol{x})\biggr)\\
&\geqslant \sum_{m=1}^{M}\alpha_m \min\biggl(Q_m(\boldsymbol{x}\to\boldsymbol{y}), \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}Q_m(\boldsymbol{y}\to\boldsymbol{x})\biggr)\\
&\geqslant \sum_{m=1}^{M}\alpha_m Q_m(\boldsymbol{x}\to\boldsymbol{y}) \min\biggl(1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\frac{Q_m(\boldsymbol{y}\to\boldsymbol{x})}{Q_m(\boldsymbol{x}\to\boldsymbol{y})}\biggr)\\
&= P_\alpha(\boldsymbol{x}\to\boldsymbol{y}).
\end{aligned}$$

Because $\widetilde{P}_\alpha(\boldsymbol{x}\to\boldsymbol{y}) \geqslant P_\alpha(\boldsymbol{x}\to\boldsymbol{y})$ for $\boldsymbol{x}\neq\boldsymbol{y}$, it follows from Theorem 11.5 of Peskun that $\widetilde{P}_\alpha$ will give an asymptotic variance at least as small as P_α does. In other words, there is an advantage in using a weighted average of the proposals Q_m instead of a weighted average of the transitions P_m.

Now $Q_\alpha(\boldsymbol{x}\to\boldsymbol{y}) = \sum_{m=1}^{M}\alpha_m Q_m(\boldsymbol{x}\to\boldsymbol{y})$. Therefore, to use $\widetilde{P}_\alpha$ we have to evaluate $Q_m(\boldsymbol{x}\to\boldsymbol{y})$ and $Q_m(\boldsymbol{y}\to\boldsymbol{x})$ for all $m = 1,\dots,M$ not just for the one sampled choice $m(i)$. The extra cost of $\widetilde{P}_\alpha$ could outweigh its variance advantage over P_α.

We can make a transition matrix out of $P_1,\dots,P_M$ that preserves detailed balance while sampling in a fixed order. If P_j preserve detailed balance for $j=1,\dots,M$ then so do

$$P_{1:M,M:1} = P_1 P_2 \dots P_M P_M \dots P_2 P_1, \quad \text{and} \tag{11.31}$$

$$P_{1:M:1} = P_1 P_2 \dots P_{M-1} P_M P_{M-1} \dots P_2 P_1. \tag{11.32}$$

We prove (11.32) in Corollary 12.1 in §12.4 on the Gibbs sampler. There we also consider transitioning according to $P_1, \dots, P_M$ arranged in a random order.

11.10 Burn-in

The most basic usage of MCMC output is to estimate an expectation such as $\mu = \int f(\boldsymbol{x})\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. The customary estimate is $\hat\mu = (1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)$ where $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are sampled by an MCMC algorithm with stationary distribution π.

A common practice is ignore the first $b < n$ generated points and estimate μ by

$$\frac{1}{n-b}\sum_{i=b+1}^n f(\boldsymbol{x}_i).$$

The practice is called ***burn-in***. The distribution of $\boldsymbol{x}_i$ usually only approaches π as i increases and so the first few observations might be very unrepresentative of π. Including them could bias the answer. For instance, one sample for the Ising model in Figure 11.10 starts with all pixels white and another starts with all pixels black. Two others start with completely random pixels. None of those four configurations is very representative of the Ising model. The first few $\boldsymbol{x}_i$ in each sample have quite low probability as the Markov chain quickly climbs towards a higher probability region. We could think of the burn-in period as one way of finding a good starting point for the simulation.

The term 'burn-in' comes from the electronics industry. Some products like computer chips have very high failure rates in their first hours of use. Those that survive an initial trial period are expected to last a long time. Burn-in consists of subjecting all the chips to a trial period, perhaps at higher than ordinary temperature, and then keeping only the survivors. The analogy to discarding the first b samples is not perfect. Some authors call the initial period ***warmup*** instead.

The amount of burn-in to use is subject to debate. Geyer (2011) advocates against burn-in and uses other ways of finding good starting points. In the same volume, Gelman and Shirley (2011) advocate discarding the first $n/2$ observations.

Very often using a small amount of burn-in makes no difference. For instance, if $|f(\boldsymbol{x})| \leqslant C$ and bC/n is negligible compare to μ then the estimates with or without burn-in are close. On the other hand if outlying values are possible for f and $f(\boldsymbol{x}_1)$ is an outlier, then burn-in can make a difference. The same happens if $f(\boldsymbol{x}) \in \{0, 1\}$ with the value 1 describing a very rare event. Then if $f(\boldsymbol{x}_1) = 1$, burn-in makes an important difference especially if the Markov chain moves slowly. For instance in Figure 11.10, one of the starting points generated numerous results with mean spin per site below -0.5. The true probability of that event might be far below $1/n$ for even a very large simulation. If so, failing to discard those points would seriously bias the estimate of that rare event probability.

11.11 Convergence diagnostics

Two of the hardest problems in MCMC are to decide whether the distribution of $\boldsymbol{x}_i$ has nearly converged to π and deciding whether the $\boldsymbol{x}_i$ are mixing well. It is plain to see in Figure 11.4 that for $\sigma = 1$ the points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{1000}$ have not sampled π well because they have not yet discovered the smaller mode. From $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{5000}$, we see that the chain is not mixing well, because it has only made four transitions into the smaller mode.

While traces are quite useful, they are only one way diagnostics. When they show us that the chain is not sampling π well, we can believe it. Unfortunately a trace can look perfectly good even when the sample is poor. For instance the trace of the first 1000 samples in Figure 11.4 for $\sigma = 1$ looks good. The ACF has the same problem. If it shows slowly decaying autocorrelations, we know there is a problem, but if they decay rapidly we might still have missed part of the space.

One approach to generating diagnostics is to run multiple independently generated Markov chains, starting them in different places, ideally even more widely separated than samples from π would be. If they end up intermingling that is a sign that they may have converged to π though it is not proof. For instance, the four Ising simulations in Figure 11.10 appear to have merged.

Even this simple concept requires a detailed derivation, which we sketch. The best known version of this diagnostic is due to Gelman and Rubin (1992) with a correction in Brooks and Gelman (1998). It is based on an assumption that the statistic $f(\boldsymbol{x})$ of interest is approximately normally distributed. They use m independent chains of length $2n$. The first n points are discarded from each chain. Using burn-in is logical when the chains are purposely started far apart and we are waiting to see if they merge. Then for a function f of interest, let $y_{ij} = f(\boldsymbol{x}_{ij})$. We measure the extent of variation of these y_{ij} within and between chains by

$$
\begin{aligned}
W &= \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=n+1}^{2n} (y_{ij} - \bar{y}_{\bullet \mathrm{j}})^2, \quad \text{and}\\
B &= \frac{n}{m-1} \sum_{j=1}^{m} (\bar{y}_{j\bullet} - \bar{y}_{\bullet\bullet})^2
\end{aligned}
$$

where $\bar{y}_{j\bullet} = (1/n)\sum_{i=n+1}^{2n} y_{ij}$ and $\bar{y}_{\bullet\bullet} = (1/m)\sum_{j=1}^{m} \bar{y}_{j\bullet}$. The denominators are constructed so that $\mathbb{E}(B) = \mathbb{E}(W) = \mathrm{Var}(f(\boldsymbol{x}))$ if $\boldsymbol{x}_{n+1,j} \sim \pi$. Now let

$$
\hat{\sigma}^2_{+} = \frac{n-1}{n} W + \frac{B}{n},
$$

which would also be unbiased for σ^2 if the $\boldsymbol{x}_{n+1,j} \sim \pi$ but should be biased up by the B/n term otherwise. To account for randomness in $\bar{y}_{\bullet\bullet}$, they replace $\hat{\sigma}^2_{+}$ by $\widehat{V} = \hat{\sigma}^2_{+} + B/(mn)$.

If $\widehat{V}/\sigma^2$ is than 1, that is a sign that the chains have not mixed. Because σ^2 is not known, they estimate it by W to produce the estimate

$$\widehat{R} = \frac{\widehat{V}}{W}.$$

There are other versions of $\widehat{R}$ that include a complicated degrees of freedom adjustment. That adjustment has vanishing importance as $n \to \infty$. If one or more functions f of interest have $\widehat{R} > 1.1$, that is a sign that the chains have not sufficiently merged. If $\widehat{R}$ is close to one for all the functions considered then no lack of convergence has been detected though that does not prove convergence.

For a discussion of the tradeoffs between using one chain and multiple chains, see page 42 of the chapter end notes.

Without any knowledge of π beyond the output $\boldsymbol{x}_i$ it seems unlikely that there can be decisive evidence in favor of convergence. Sometimes however we know a bit about π that can be useful. In Bayesian applications where the $\boldsymbol{x}$ in our simulations is a vector of parameters in a model for data $\boldsymbol{D}$ we can make up a parameter value $\boldsymbol{x}_*$, and sample data $\boldsymbol{D}_*$ as if our $\boldsymbol{x}_*$ were the true parameter. Corresponding to that generated data $\boldsymbol{D}_*$ there is a new posterior π_* where we actually know the true parameter value. It is the $\boldsymbol{x}_*$ we chose. We can then run an MCMC to simulate from π_* and see if the sample values cluster around $\boldsymbol{x}_*$ in the way that we think the simulation from π should cluster around the unknown true $\boldsymbol{x}$. For instance, if we think that π should have its mean close to the true $\boldsymbol{x}$ then we can check whether the simulation from π_* has a mean near $\boldsymbol{x}_*$. We can do this repeatedly, getting numerous such data sets $\boldsymbol{D}_*$ for a variety of different $\boldsymbol{x}_*$.

It is possible that our simulations from π_* behave in a completely different way than simulations from π do giving us false hope of convergence. However experience and judgment with the problem domain might cast that possibility as an unreasonable doubt.

11.12 Error estimation

Having computed $\hat{\mu}$ we will often want a confidence interval for μ. In IID sampling we would assume that $\int f(\boldsymbol{x})^2\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$, look to the central limit theorem to justify a confidence interval, and then estimate the appropriate variance. We use this same strategy in MCMC, devising an approximate confidence interval under the assumption that $\int f(\boldsymbol{x})^2\pi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} < \infty$.

There is a central limit theorem for Markov chains in §**??**. There

$$\sqrt{n}(\hat{\mu} - \mu) \to \mathcal{N}(0, \sigma_f^2), \quad \text{where} \tag{11.33}$$

$$\sigma_f^2 = \sigma^2\Bigl(1 + 2\sum_{\ell=1}^{\infty}\rho_\ell\Bigr), \tag{11.34}$$

with ρ_ℓ the autocorrelation of $f(\boldsymbol{x}_i)$ under Markov chain sampling. Usually a sample-based estimate of σ_f^2 must be used. Truncating the sum in (11.34) and then plugging in estimates of the covariances $\rho_\ell\sigma^2$ does not work well. Some subtle time series analysis methods are possible, but a very simple alternative called ***batching*** is widely used.

In batching, we take $n = bm$ observations in b batches of m consecutive observations. The b for batching is not related to the b for burn-in in §11.10. For $y_i = f(\boldsymbol{x}_i)$, let

$$\bar{y}_j = \frac{1}{m}\sum_{i=(j-1)m+1}^{jm} y_i, \quad j = 1,\dots,b.$$

The batch analysis treats the $\bar{y}_j$ as if they were nearly IID normal random variables with mean μ. Then an approximate 99% confidence interval for μ is

$$\bar{y} \pm t^{0.995}_{(b-1)}s, \quad \text{for} \quad s^2 = \frac{1}{b(b-1)}\sum_{j=1}^{b}(\bar{y}_j - \bar{y})^2,$$

where $t^{0.995}_{(b-1)}$ is the 99.5'th percentile of the $t_{(b-1)}$ distribution. Notice that $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ is also the average of $\bar{y}_j$. The recommended number of batches is usually small, perhaps 20, and then the batches of length $m = n/20$ are typically quite long.

The normality assumption on $\bar{y}_j$ comes from a central limit theorem and it is meant for large m. The approximate independence comes from the wide separation between the y_i values in $\bar{y}_j$ and $\bar{y}_{j'}$ for $j' \ne j$ when m is large. To sketch how it works, we will consider a very simple situation with $\delta^{|\ell|} \leqslant \rho_\ell \leqslant \gamma^{|\ell|}$ for $0 \leqslant \delta \leqslant \gamma < 1$. That is ρ_ℓ is bounded between two geometrically decaying sequences. That geometric decay is an idealized correlation pattern similar though not necessarily identical to what one sees when correlations decay slowly. Under this model, for $j > j'$ and very large m,

$$\begin{aligned}\mathrm{Cov}(\bar{y}_j, \bar{y}_{j'}) &= \frac{\sigma^2}{m^2}\sum_{i=1}^{m}\sum_{i'=1}^{m}\rho_{(j-j')m+i-i'}\\ &\approx \sigma^2\int_0^1\int_0^1 \gamma^{m(j-j'+u-v)}\,\mathrm{d}u\,\mathrm{d}v \leqslant \sigma^2\gamma^{m(j-j')}.\end{aligned}$$

Next $\mathrm{Var}(\bar{y}_j) \geqslant \sigma^2/m$. Putting together these estimates gives us a rough bound of $m\gamma^m$ for $\mathrm{Corr}(\bar{y}_j, \bar{y}_{j'})$ when $j \ne j'$. If $|\gamma|$ is reasonably well below 1 and m is very large, then this correlation will be negligible. The end notes have further references on batching.

11.13 Thinning

Another alternative to the customary estimate $\hat{\mu} = (1/n)\sum_{i=1}^{n} f(\boldsymbol{x}_i)$ is to use ***thinning*** on the Markov chain. If we thin to every k'th observation, then we use

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} f(\boldsymbol{x}_{ki})$$

for some integer $k \geqslant 2$. Thinning is also called ***subsampling***. It is possible to use thinning after burn-in, but for simplicity we study them separately.

The appeal of thinning is that observations spaced apart by k steps are ordinarily more nearly independent than consecutive observations are. Despite lowered autocorrelations, thinning increases variance compared to using $\hat\mu = (1/n)\sum_{i=1}^n f(\boldsymbol{x}_i)$ for $n = kn_k$. It is ignoring $k-1$ observations for every one it uses.

In some circumstances thinning can indirectly lower variance. If it costs on average one time unit to advance the chain from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$ and then θ time units to compute $f(\boldsymbol{x}_i)$, then $\hat\mu$ has cost $n(1+\theta)$. Computing $\hat\mu_k$ only costs $n(1+\theta/k)$. In the amount of time it takes to compute $\hat\mu$ we could instead run the thinned chain for about $n(1+\theta)/(1+\theta/k)$ steps. If θ is large and the autocorrelations decay slowly, then thinning reduces variance for a given amount of computer time.

In many applications, such as Bayesian computing, $f(\boldsymbol{x})$ might just be $\boldsymbol{x}$ and then θ is effectively zero. For the Metropolis hard disk model the function f of interest used the interpoint distances, yielding a meaningfully large cost θ. In the Ising model, computing statistics after every sweep is a form of thinning compared to computing them after every pixel has been visited. Thinning can be a practical alternative to sophisticated update algorithms for computing $f(\boldsymbol{x}_{i+1})$ from $f(\boldsymbol{x}_i)$. A reasonable guideline suggested by Hans Andersen is to thin if necessary to ensure that at most half of the computer's time is spent evaluating f.

Thinning also reduces the cost to store data by a factor of k. If $\boldsymbol{x}_i$ are large, such as three dimensional images, these savings can be important. The space savings does not always apply. For instance, if $f(\boldsymbol{x}_i)$ are being batched as described in §11.12 then we might only have to store some of the batch means. For other purposes, such as looking at graphical displays of sample points, or exploring various functions f interactively, thinned samples will be more convenient than batch means. Similar to storage costs, data transmission costs can be reduced by thinning.

Chapter end notes

The Metropolis-Hastings update is the generalization by Hastings (1970) of the algorithm by Metropolis et al. (1953). Hastings (1970) also proposed the Metropolized independence sampler (bottom of page 103) and many of the basic proposal combining strategies in §11.9. The derivation of the Hastings rule $A(\boldsymbol{x} \to \boldsymbol{y})$ in §11.4 is based on the account in Newman and Barkema (1999). The two state explanation of the beginner's error in MCMC came from Seth Tribble. Tierney (1994) discusses the autoregressive sampler as does Hastings (1970).

Ising model

For more background on the Ising model, see Snell and Kindermann (1980). Onsager (1944) found an analytic exression for the free energy of the two dimensional Ising model.

Numerical issues in Metropolis-Hastings

Equation (11.19) for $A(x \to y)$ could be affected by numerical underflow and overflow in which values get rounded to 0 or $+\infty$, respectively. If the Hastings ratio turns out to be $0/0$ or ∞/∞ then it may be given the value NaN in floating point (not a number). This complicates the acceptance rejection decision. Both NaN $> x$ and NaN $< x$ are usually considered false statements in numerical computing. An MCMC could get into a bad part of the space and accept every proposal. Working with logarithms helps to avoid this but $\log(R(\boldsymbol{x} \to \boldsymbol{y}))$ can still be NaN. It is worth being vigilant about these numerics and make sure that any junk output is flagged.

Convergence

The multiple chains convergence diagnostic is from Gelman and Rubin (1992). They credit Fosdick (1959) for an earlier use of multiple chains to diagnose convergence. Brooks and Gelman (1998) also consider some multivariate versions of the diagnostic. Cowles and Carlin (1996) discuss numerous other convergence diagnostics. Andrew Gelman has advocated fake data tests for Bayesian MCMC problems since at least Kass et al. (1998). There are some promising methods for computationally confirming that a sample is close to its target distribution, using a theory of Stein discrepancy. The challenges are making them work for high dimensional problems and controlling the computational costs. See Gorham and Mackey (2015).

One chain versus several

The diagnostic measure $\widehat{R}$ in §11.11 required $m \geqslant 2$ independently sampled Markov chains. There has long been controversy over whether it is better to simulate one long chain of length n or m chains of length n/m each.

Neither choice will always be best and in a given situation we might not know which is best for our π. At one extreme, taking n/m absurdly small, too small for the chain to move very far, would require one to have a very good way to pick starting points, almost as good as sampling from π. At the other extreme, consider a Markov chain where the state space is almost reducible, splitting into two parts between which movement is very hard. The low temperature Ising model is of this type. If we have a starting point $\boldsymbol{x}_0$ somehow equidistant from the two parts, then one single chain may well sample just one of the parts while multiple starts could easily discover them both. A toy version of this problem is to have a Markov chain on $\{-M, -M+1, \dots, -1, 0, 1, \dots, M-1, M\}$ with a strong tendency to move away from 0. Multiple starts at 0 will work better than a single start at 0.

Flegal and Jones (2011, §7.4.1.2) have a good survey of the literature on this problem from Bayesian statistics, operations research and physics. There are settings where the single long chain has less bias. An additional issue is that with burn-in, multiple chains require discarding bm observations instead of just b of them.

Parallel computing changes the picture. If there are k processors available, each capable of simulating the Markov chain, then we might as well run k long chains instead of having one long chain and $k-1$ idle processors. The debate doesn't quite end because one might prefer to run km Markov chains, m per processor for some $m \geqslant 2$.

Batching

Geyer (1992) recommends the batch analysis for MCMC and provides references with rigorous justifications. Schmeiser (1982) recommends using 10 to 30 batches and surveys some early literature on batch methods. Glynn and Whitt (1991) show that consistent estimation of σ_f^2 requires a number of batches that increases to infinity. Flegal and Jones (2010) show that the optimal number of batches should grow proportionally to $n^{1/3}$, though it can be hard to know the appropriate constant of proportionality. Their result shows that slow growth is appropriate. They include a well known technique from time series in which the batches are allowed to overlap.

There are other time series techniques that may help. Kitamura (1997) develops empirical likelihood confidence intervals for time series. Politis and Romano (1994) propose a time series bootstrap.

Most of the work has been on confidence intervals for univariate parameters. Consistent estimation of the variance-covariance matrix for a vector $f(\boldsymbol{x}) \in \mathbb{R}^r$ has had much less study. A notable exception is Vats et al. (2015).

Thinning

Geyer (1991) shows that thinning a reversible Markov chain increases variance. MacEachern and Berliner (1994) extend the conclusion to more general chains. Owen (2017) shows that sometimes very large thinning factors k are optimal when $f(\boldsymbol{x})$ is expensive to compute. Much smaller factors can then be nearly optimal. The recommendation to spend at most half of the time computing f is from Hans Andersen.

Exercises

11.1. Consider the Montréal metro walk of §11.2, for a walker starting at $X_0 =$ Snowdon.

- **a)** Plot $\mathbb{P}_{\text{Snowdon}}(X_n = \text{Berri-UQAM})$ as a function of n for $n = 1, \dots, 1000$. Make a second plot just for $n = 1, \dots, 10$.
- **b)** Suppose that the walker changes behavior and does not linger at Longueuil. Instead, every visit to Longueuil is immediately followed by a visit to Berri-UQAM. Write the transition matrix $\widetilde{P}$ for this version of the walk. With this change in the Markov chain, repeat part **a**.
- **c)** Now suppose that the walker always has probability 1/2 of lingering. Specifically, let $\widetilde{P}$ be the transition matrix from part **b**, let I_5 be the 5 by 5 identity matrix, and define $\bar{P} = (\widetilde{P} + I)/2$. Repeat part **a** using the transition matrix $\bar{P}$.
- **d)** Find $\pi(\text{Berri-UQAM})$ for the original walk. One way to do this is by solving the eigenvalue problem $\pi P = \pi$, taking care to get the ***left*** eigenvector corresponding to eigenvalue 1. Compare this probability to that for the other two walks considered here. Comment also on how quickly the limiting probability is approached in the three walks.

11.2. Find numerically, by computing an eigendecomposition, the stationary distributions of the three metro walks in Exercise 11.1.

11.3. Let $W \in [0, \infty)^{N \times N}$ be the incidence matrix of a weighted undirected graph on N nodes. The entry $W_{jk} = W_{kj}$ is the weight on the edge between nodes j and k. The degree of node j is $d_j = \sum_{k=1}^{N} W_{jk}$. We allow $W_{jj} > 0$ and assume only that each $d_j > 0$. Let P be the transition matrix with $P(j \to k) = W_{jk}/d_j$.

- **a)** Show that π is a stationary distribution for P where $\pi(j) \propto d_j$.
- **b)** Use the previous part to get an exact expression for the stationary distribution π of the original metro walk. That is give

$$\pi = \frac{1}{b}\begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \end{pmatrix}$$

where a_i and b are positive integers.

11.4. In this exercise, you prove that the Ising model has a spatial Markov property. Choose a particular point ℓ. Now for $\boldsymbol{x} \in \{-1, 1\}^N$ define $\boldsymbol{x}^+$ to be the point with $x_j^+ = x_j$ for $j \neq \ell$ and $x_\ell^+ = 1$. Similarly let $\boldsymbol{x}^-$ be $\boldsymbol{x}$ after setting $x_\ell = -1$. It is also convenient to use $\widetilde{J} = J/T$ and $\widetilde{B} = B/T$. Let $S = \sum_{j \sim \ell} x_j$. Prove that

$$\mathbb{P}(X_\ell = 1 \mid X_j = x_j, j \neq \ell) = \frac{\exp\bigl(2\widetilde{J}S + 2\widetilde{B}\bigr)}{1 + \exp\bigl(2\widetilde{J}S + 2\widetilde{B}\bigr)}.$$

In statistical terms, the value X_ℓ follows a logistic regression in the sum S of its neighbors with slope $2J/T$ and intercept $2B/T$.

11.5. If we were somehow able to make proposals $Q(x \to y)$ from the stationary distribution $\pi(y)$ then of course we could use ordinary MC sampling. Show that Metropolis-Hastings sampling reduces to ordinary MC sampling with this proposal.

11.6. Suppose that $\pi(\boldsymbol{x})$ is not available but we have an unnormalized $\pi_u(\boldsymbol{x})$. Similarly $Q(\boldsymbol{x})$ is unavailable but an unnormalized $Q_u(\boldsymbol{x})$ is available. Now suppose that although both normalizing constants are unknown we do know that the one for π is c times the one for Q for a known value $c > 0$. How then can we compute the importance sampling estimate (11.26)?

11.7. For the Ising model, what is that stationary distribution if the proposals (11.30) are always accepted?

11.8. Show that Barker's acceptance probability (11.20) satisfies detailed balance. Show *directly* that Barker's acceptance probability is no larger than the Metropolis-Hastings acceptance probability. (It is not enough to just remark that we derived Metropolis-Hastings by maximizing the acceptance probability.)

Construct a simple example for which Barker's probability is strictly smaller than Metropolis-Hastings.

11.9. For the Gaussian mixture problem in §11.5,

a) Find numerically the value of σ that minimizes the first order autocorrelation ρ_1.

b) Find numerically the value $x_* = \arg\min_{0\leqslant x\leqslant 5} \pi(x)$ that separates the two modes.

c) For $f(x) = \mathbf{1}\{x > x_*\}$, find numerically the value of σ that minimizes $\rho(f(x_i), f(x_{i+1}))$.

11.10. For π_u given by equation (11.24) consider RWM using proposals $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2 I)$. Take $\sigma = 2^r$ and search for a choice of r among integers $-3 \leqslant r \leqslant 4$.

a) We say that the walk has 'discovered' star j if $\|\boldsymbol{x}_i - \theta_j\| \leqslant 2$ for any $i = 1, \dots, n$, where θ_j is the j'th column of Θ from (11.25). For each of the eight values of σ under consideration, run RWM 10 times independently, for $n = 10{,}000$ steps starting at $\boldsymbol{x}_0 = (0, 0)$ each time, and report the average number of stars discovered in the 10 runs at each σ.

b) Report the fraction of accepted proposals for each σ.

c) For each σ produce the ACF up to lag 100 for the first component of $\boldsymbol{x}$. You need to pool 10 ACFs for each σ. Which σ has the smallest $\hat\rho_1$ and what acceptance probability does it have?

d) We say that $\boldsymbol{x}$ is 'at star j' if $\|\boldsymbol{x} - \theta_j\| = \min_{1\leqslant k\leqslant 8} \|\boldsymbol{x} - \theta_k\|$. It is possible for a point $\boldsymbol{x}$ to be at more than one star, but that event has probability zero, so we ignore it. For $j = 1, \dots, 8$, let $f_j(\boldsymbol{x})$ be 1 if $\boldsymbol{x}$ is at star j and 0 otherwise and let $\mu_j = \int f_j(\boldsymbol{x})\pi_u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} / \int \pi_u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. For each value of σ find the estimated value of $f_1(\boldsymbol{x})$, the probability that $\boldsymbol{x}$ is at star Betelgeuse. We might expect π to put more probability there than on any other stars because that star is remote, so the answer should be at least 1/8 but a poorly mixing chain could miss that.

e) Estimate the ACF for each σ out to lag 100 for the function $f_1(\boldsymbol{x})$. Which σ has the smallest $\hat{\rho}_1$ and what acceptance probability does it have?

11.11. For the Markov chain of Exercise 11.10, select a value of σ where every star was reliably discovered. Run the chain until x is at either Betelgeuse ($j = 1$) or Rigel ($j = 7$). Then record the number N of additional steps until the chain is at the other of those two stars. Keep counting the transitions until at least $n = 100{,}000$ (which might not be enough to get a good estimate). What is the estimated waiting time for the walk to go from Betelgeus to Rigel? What is the estimated waiting time in the other direction?

11.12. For $x, y \in [0, 1]$, the wraparound distance between x and y is $d_W(x, y) = \min(|x - y|, |x + 1 - y|, |x - 1 - y|)$. Prove that

$$d_W(x, y) = \frac{1}{2} - \Bigl||x - y| - \frac{1}{2}\Bigr|.$$

This formula simplifies some programming. It doesn't require if-then-else statements, or rather, those are hidden within the absolute value function.

The next exercises are based on the Metropolis hard disk simulation, generalizing the original trigonal starting grid to one with $8k$ rows of $7k$ disks each for an integer $k \geqslant 1$. The original simulation has $k = 2$. The $56k^2$ disks have diameters $d(1 - 2^{\nu-8})$ for $d = 1/(7k)$. These exercises are of a project type and are well suited to team work. They require experimentation with different simulations, judgement calls, and an explanation of those judgements using traces and ACFs and perhaps other computations.

It is interesting to animate these simulations. When comparing different values of k it is important to measure spaces between disks as a fraction of their diameters.

11.13. For the Metropolis simulation with $k = 2$ and $\nu = 3$, how long do you think it takes before the average minimum interpoint distance has become stable?

11.14. For the $\nu = 3$ case it looks like each disk must always have the same 6 nearest neighbor disks. Only 6 distances need to be computed when the Markov chain is updated and even then they only need to be computed if the proposal was accepted. This fact allows larger grids and longer simulations to be done in the same amount of time. Compare estimates of the mean distance to a nearest neighbor for $k = 1, 2, 3, 4$.

11.15. The left panel of Figure 11.1 shows a few unusually large gaps. An interesting quantity is then the maximum over $56k^2$ disks of the maximum distance to one of its 6 nearest neighbors. Use MCMC to find the distribution of this maximal gap. Compare the results for different k.

11.16. In the previous exercise we could also be interested in the average over $56k^2$ disks of the largest gap to one of its six neighbors. Use MCMC to find the distribution of this maximal gap. Compare the results for different k.

12

Gibbs sampler

When MCMC methods are used, we typically cannot sample $x \sim \pi$. Sometimes however we can come close. We may be able to sample each component x_j from its conditional distribution given all the other components $x_1, \dots, x_{j-1}, x_{j-1}, \dots, x_d$ of $\boldsymbol{x}$. This distribution is called the ***full conditional distribution*** of x_j.

In the Gibbs sampler we repeatedly sample one component after another from the appropriate full conditional distribution. Many models used in statistics and machine learning have simple full conditional distributions for which the Gibbs sampler is easy to use. This strategy is also known as the ***heat-bath method*** or ***Glauber dynamics***.

12.1 Stationary distribution for Gibbs

The point $\boldsymbol{x}$ is comprised of x_j, where $1 \leqslant j \leqslant d$, and all of its other $d-1$ components, which we lump together into $\boldsymbol{x}_{-j}$. We write the full conditional distribution of x_j given $\boldsymbol{x}_{-j}$ as $\pi_{j|-j}(x_j \,|\, \boldsymbol{x}_{-j})$ and in abbreviated form it is $\pi(x_j \,|\, \boldsymbol{x}_{-j})$. Let the marginal distribution of $\boldsymbol{x}_{-j}$ be $\pi_{-j}(\boldsymbol{x}_{-j})$, or $\pi(\boldsymbol{x}_{-j})$ in abbreviated form.

The Gibbs sampler is shown in Algorithm 12.1. There are two versions. In the ***random scan*** Gibbs sampler, the component to update is chosen at random from $1, \dots, d$. In the ***systematic scan*** Gibbs sampler, the components are updated sequentially.

We can show directly that sampling component j from its full conditional distribution preserves the stationary distribution π. Suppose that $\boldsymbol{x} \sim \pi$ and that we replace x_j by a value $z \sim \pi(x_j \,|\, \boldsymbol{x}_{-j})$, obtaining the point $\boldsymbol{y}$ with $y_j = z$

Algorithm 12.1 The Gibbs sampler

```
Gibbs( $\boldsymbol{x}_0$, $\pi$, $n$ )

// $\boldsymbol{x}_0 \in \mathbb{R}^d$ is the starting point
// $\pi$ is a target distribution with full conditional distributions $\pi_{j|-j}$
// $n$ is the number of points to generate

for $i = 1$ to $n$ do
    $j \sim \mathbf{U}\{1,\dots,d\}$
    $z \sim \pi_{j|-j}(\cdot \mid \boldsymbol{x}_{i-1,-j})$
    $\boldsymbol{x}_i = \boldsymbol{x}_{i-1}$
    $x_{ij} = z$
```

This is the random scan Gibbs sampler. For the systematic scan Gibbs sampler take $j = \ell + 1$ where $\ell = i - 1 \mod d$.

and $y_k = x_k$ for $k \neq j$. Then

$$p(\boldsymbol{y}) = \pi(\boldsymbol{y}_{-j})\pi(y_j \mid \boldsymbol{y}_{-j}) = \pi(\boldsymbol{y}).$$

We may also understand the Gibbs sampler by relating it to Metropolis-Hastings. Suppose that we are at point $\boldsymbol{x}$ and have decided to modify component j of $\boldsymbol{x}$ to take the value z. Let $\boldsymbol{y}$ be the point with $y_j = z$ and $y_k = x_k$ for $k \neq j$. If we use $\boldsymbol{y}$ as the proposal in Metropolis-Hastings, then the Hastings ratio is

$$R(\boldsymbol{x} \to \boldsymbol{y}) = \frac{\pi(\boldsymbol{y})Q(\boldsymbol{y} \to \boldsymbol{x})}{\pi(\boldsymbol{x})Q(\boldsymbol{x} \to \boldsymbol{y})} = \frac{\pi(\boldsymbol{x}_{-j})\pi(z \mid \boldsymbol{x}_{-j})\pi(x_j \mid \boldsymbol{x}_{-j})}{\pi(\boldsymbol{x}_{-j})\pi(x_j \mid \boldsymbol{x}_{-j})\pi(z \mid \boldsymbol{x}_{-j})} = 1. \qquad (12.1)$$

Equation (12.1) shows that if we update component j of $\boldsymbol{x}$ by sampling from its full conditional distribution, then we can view this as a Metropolis-Hastings proposal that is never rejected. Next we show that the Gibbs sampler preserves the stationary distribution. We present the argument for discrete full conditional distributions so that the analysis uses transition matrices.

For $j = 1, \dots, d$, let P_j be the transition matrix corresponding to an update of x_j from its full conditional distribution. In random scan Gibbs, each of the n steps in Algorithm 12.1 is a draw from $\bar{P} = (1/d)\sum_{j=1}^d P_j$. Therefore by Proposition 11.1, each of those n steps preserves the stationary distribution π.

The fixed scan Gibbs sampler can be looked at as a non-homogeneous Markov chain that repeatedly cycles through transitions $P_1, P_2, \dots, P_d$ in that order. It can also be viewed as an homogeneous Markov chain that uses the transition matrix $P_{1:d} = P_d P_{d-1} \dots P_2 P_1$ to advance from $\boldsymbol{x}_{id}$ to $\boldsymbol{x}_{i(d+1)}$ for $0 \leqslant i \leqslant \lfloor n/d \rfloor - 1$. Either way, it preserves the stationary distribution π because each P_j does and hence so does $P_{1:d}$.

We defer a discussion of ergodicity and detailed balance to §12.4. For now we note that random scan has detailed balance, while fixed scan does not but can be modified to do so.

12.2 Example: truncated normal

As an illustrative example, we use the Gibbs sampler on a multivariate normal density truncated to a rectangular region. Let $\boldsymbol{x} \in \mathbb{R}^d$ have density function

$$\pi(\boldsymbol{x}) \propto \begin{cases} e^{-\frac{1}{2}(\boldsymbol{x}-\mu)^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\mu)}, & a_j \leqslant x_j \leqslant b_j,\ j=1,\dots,d \\ 0, & \text{else.} \end{cases}$$

That is, $\boldsymbol{x}$ is a normal random vector truncated to the interval $[\boldsymbol{a}, \boldsymbol{b}]$. It is allowed to have some $a_j = -\infty$ and some $b_j = \infty$. The normalizing constant in the density is $(2\pi)^{d/2}\det(\Sigma)^{1/2}\mathbb{P}(\mathcal{N}(\mu,\Sigma) \in [\boldsymbol{a},\boldsymbol{b}])$, written in terms of the ordinary $\pi \doteq 3.14159$ (not the stationary distribution π). The full conditional distribution of x_j given $\boldsymbol{x}_{-j}$ is a one dimensional truncated normal distribution with density proportional to

$$\exp\Bigl(-\frac{1}{2\,\widetilde{\sigma}_j^2}(x_j - \widetilde{\mu}_j)^2\Bigr) \times \mathbb{1}_{a_j \leqslant x_j \leqslant b_j}$$

where

$$\begin{aligned} \widetilde{\mu}_j &\equiv \mathbb{E}(x_j \mid \boldsymbol{x}_{-j}) = \mu_j + \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}(\boldsymbol{x}_{-j} - \mu_{-j}), \quad \text{and,} \\ \widetilde{\sigma}_j^2 &\equiv \mathrm{Var}(x_j \mid \boldsymbol{x}_{-j}) = \Sigma_{j,j} - \Sigma_{j,-j}\Sigma_{-j,-j}^{-1}\Sigma_{-j,j}. \end{aligned}$$

As a result, we can run the Gibbs sampler on π by using samples from truncated normal distributions. The following sequence updates component j:

$$\begin{aligned} &\widetilde{a} \leftarrow \Phi\Bigl(\frac{a_j - \widetilde{\mu}_j}{\widetilde{\sigma}_j}\Bigr), \quad \widetilde{b} \leftarrow \Phi\Bigl(\frac{b_j - \widetilde{\mu}_j}{\widetilde{\sigma}_j}\Bigr), \quad U \sim \mathbf{U}(0,1), \\ &x_j \leftarrow \widetilde{\mu}_j + \widetilde{\sigma}_j\Phi^{-1}\bigl(\widetilde{a} + (\widetilde{b} - \widetilde{a})U\bigr). \end{aligned}$$

The values $\widetilde{\sigma}_1, \dots, \widetilde{\sigma}_d$ remain constant throughout the sampling. Each $\widetilde{\mu}_j$ depends on $\boldsymbol{x}_{-j}$ and so they change as the sampling proceeds. The vectors $\Sigma_{j,-j}\Sigma_{-j,-j}^{-1} \in \mathbb{R}^{d-1}$ only need to be computed once. As usual, $\Phi^{-1}(y)$ for large positive y may be numerically inferior to using $-\Phi^{-1}(-y)$.

We look at two concrete examples. In both cases the truncation region is $[2, 2.5] \times [2, 2.5]$. In the first case $\boldsymbol{x} \sim \mathcal{N}\bigl(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)\bigr)$, for $\rho = 0.7$, while the second has $\boldsymbol{x} \sim \mathcal{N}\bigl(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1 & -\rho\\ -\rho & 1\end{smallmatrix}\right)\bigr)$, also for $\rho = 0.7$.

Figure 12.1 shows the first 40 moves for each chain, using a systematic scan. The one at a time updates translate into moves parallel to the coordinate axes. Both chains were run for 10,000 moves in total. Figure 12.2 shows the resulting histograms for the component x_1. By symmetry, the distribution of x_2 is the same as that of x_1 in each case. The distribution of $\boldsymbol{x}$ given that it lies inside the box $[2, 2.5]^2$ is quite different in these two cases. When $\rho = -0.7$ each component has a strong tendency to be close to the minimum value 2, while for $\rho = 0.7$ the values are more evenly distributed through the box.

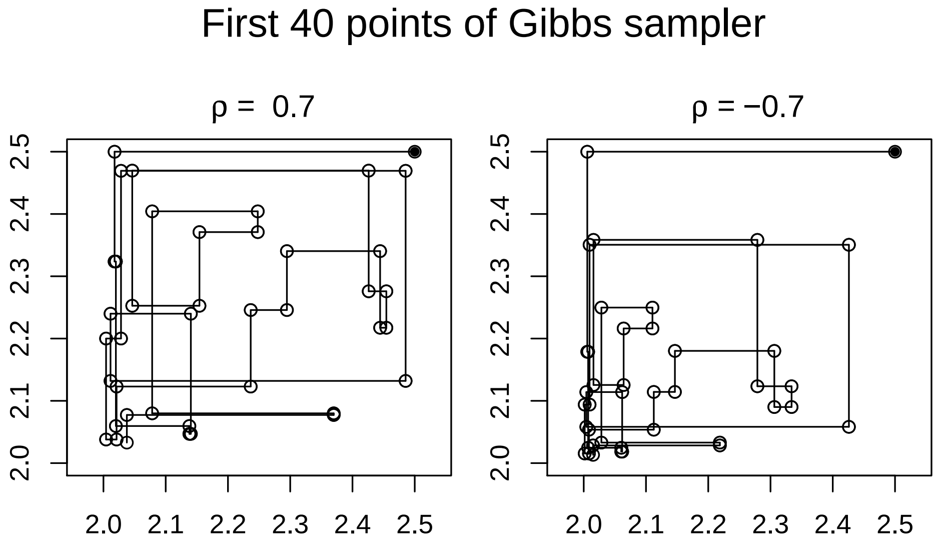

Figure 12.1: This figure shows the first 40 moves of the Gibbs sampler for bivariate normal distributions truncated to the square $[2, 2.5]^2$. Both samplers started at the solid point $(2.5, 2.5)$. In the left panel, $\rho = 0.7$ while in the right panel $\rho = -0.7$.

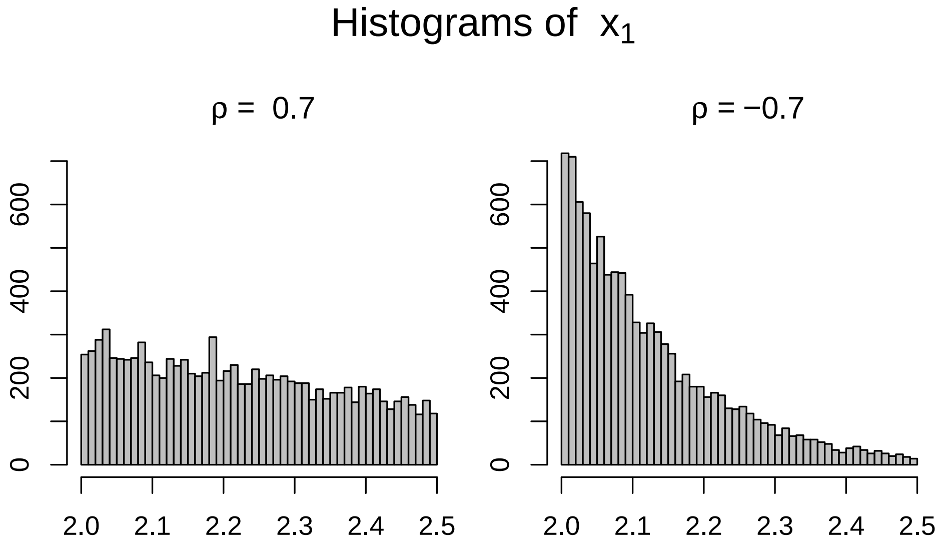

Figure 12.2: This figure shows the x_1 component of the first 10,000 points for the two Gibbs samplers illustrated in Figure 12.1.

12.3 Example: probit model

Here we revisit the probit model from §11.1. The posterior distribution of β given the data y_i is

$$\pi(\beta) \equiv p(\beta \,|\, \boldsymbol{y}) \propto p(\beta) \prod_{i=1}^{m} \Phi(\boldsymbol{z}_i^\mathsf{T}\beta)^{y_i}(1-\Phi(\boldsymbol{z}_i^\mathsf{T}\beta))^{1-y_i},$$

where $p(\beta)$ is the prior density for $\beta \in \mathbb{R}^p$. A common choice is a non-informative (improper) prior, with $p(\beta)$ constant. Here we are treating the $\boldsymbol{z}_i$ values as non-random. The probit model can be represented using latent variables $w_i = \boldsymbol{z}_i^\mathsf{T}\beta + \varepsilon_i$ for $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ as follows:

$$y_i = 1\{w_i > 0\}, \quad \text{for} \quad w_i = \boldsymbol{z}_i^\mathsf{T}\beta + \varepsilon_i.$$

Including the vector $\boldsymbol{w}$ of latent variables in the model yields

$$\pi(\beta, \boldsymbol{w}) \equiv p(\beta, \boldsymbol{w} \,|\, \boldsymbol{y}) \propto p(\beta) \prod_{i=1}^{n} \bigl(\mathbb{1}_{w_i>0}\mathbb{1}_{y_i=1} + \mathbb{1}_{w_i\leqslant 0}\mathbb{1}_{y_i=0}\bigr) e^{-(w_i - z_i^\mathsf{T}\beta)^2/2}.$$

We are increasing the dimension of the parameter space, from p to $n+p$, making this a ***data augmentation*** method. See the end notes for some references.

If we condition on $\boldsymbol{w}$ we get $\pi(\beta \,|\, \boldsymbol{w}) \propto p(\beta)\exp(-\|\boldsymbol{w} - \boldsymbol{Z}^\mathsf{T}\beta\|^2/2)$, where $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ has i'th row $\boldsymbol{z}_i$. For a non-informative prior $p(\beta)$ and a full rank matrix $\boldsymbol{Z}$, we get

$$\beta \,|\, \boldsymbol{w} \sim \mathcal{N}((\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{w}, (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1}). \tag{12.2}$$

More generally, the conjugate prior $p(\beta) = \mathcal{N}(\beta_0, \Sigma)$ yields a Gaussian full conditional for β given $\boldsymbol{w}$. (See Exercise 12.5.)

If we condition on β, we get

$$\pi(\boldsymbol{w} \,|\, \beta) \propto \prod_{i=1}^{n} \bigl(\mathbb{1}_{w_i>0}\mathbb{1}_{y_i=1} + \mathbb{1}_{w_i\leqslant 0}\mathbb{1}_{y_i=0}\bigr) e^{-(w_i - z_i^\mathsf{T}\beta)^2/2}.$$

Because this factors, the components w_i are independent given β. If $y_i = 1$ then the distribution of w_i is proportional to $\varphi(w_i - \boldsymbol{z}_i^\mathsf{T}\beta)$ on $(0,\infty)$. That is, it has the $\mathcal{N}(\boldsymbol{z}_i^\mathsf{T}\beta, 1)$ distribution conditionally on $w_i > 0$. Similarly, if $y_i = 0$, then $w_i \sim \mathcal{N}(\boldsymbol{z}_i^\mathsf{T}\beta, 1)$ conditionally on $w_i \leqslant 0$. Letting $\mu_i = \mu_i(\beta) = \boldsymbol{z}_i^\mathsf{T}\beta$, we can thus sample $\boldsymbol{w}$ given β by taking $u_i \overset{\text{iid}}{\sim} \mathbf{U}(0,1)$ and then

$$w_i = \begin{cases} \mu_i + \Phi^{-1}(\Phi(-\mu_i) + u_i\Phi(\mu_i)), & y_i = 1 \\ \mu_i + \Phi^{-1}(u_i\Phi(-\mu_i)), & y_i = 0. \end{cases} \tag{12.3}$$

The Gibbs sampler for the probit model alternates between equations (12.2) and (12.3). Although there are $n+p$ variables in $(\beta, \boldsymbol{w})$ we update them in two blocks.

12.4 Aperiodicity, irreducibility, detailed balance

While the Gibbs sampler has the right stationary distribution, that is not all that we need for it to be ergodic. We need irreducibility and we would also like aperiodicity. Finally, some versions of the Gibbs sampler are reversible (i.e., they have detailed balance) but others are not.

On a discrete state space, the Gibbs sampler always has some chance of repeating any value $\boldsymbol{x}$ that has positive probability. As a result we do not get periodic Gibbs samplers.

It is however possible for the Gibbs sampler to be reducible. Figure 12.3 depicts such a case. For the Gibbs sampler to work properly, it must be possible to reach any point from any other, using only moves parallel to the coordinate axes. The distribution in Figure 12.3 has two disjoint components and there is no way to go between them by sampling one variable at a time.

For continuous distributions π we have probability zero of reaching any specific point. There what we require is just that we can get arbitrarily close to any of the points, again by making moves parallel to the axes. See Chapter xxx for a more detailed discussion.

A different parameterization of the space would have made this distribution suitable for the Gibbs sampler. If instead of making our moves in the North-South and East-West directions we made our moves along Northwest-Southeast and Northeast-Southwest axes, the sampler would indeed be irreducible. See Exercise 12.3.

Next we turn to detailed balance. An update that replaces x_j by a sample from its full conditional distribution does have detailed balance. This follows from equation (12.1) which expresses that update as a Metropolis-Hastings proposal that is never rejected. We can also make a direct demonstration. For two points $\boldsymbol{x}$ and $\boldsymbol{y}$ we have $P_j(\boldsymbol{x} \to \boldsymbol{y}) = 0$ unless $\boldsymbol{x}_{-j} = \boldsymbol{y}_{-j}$. Let $M_j(\boldsymbol{x}, \boldsymbol{y}) = 1$ if $\boldsymbol{x}$ and $\boldsymbol{y}$ match apart from component j, that is if $\boldsymbol{x}_{-j} = \boldsymbol{y}_{-j}$, and take $M_j(\boldsymbol{x}, \boldsymbol{y}) = 0$ otherwise. Then

$$\pi(\boldsymbol{x})P_j(\boldsymbol{x} \to \boldsymbol{y}) = \pi_{-j}(\boldsymbol{x}_{-j})\pi_{j|-j}(x_j \mid \boldsymbol{x}_{-j})M_j(\boldsymbol{x}, \boldsymbol{y})\pi_{j|-j}(y_j \mid \boldsymbol{x}_{-j})$$

and

$$\pi(\boldsymbol{y})P_j(\boldsymbol{y} \to \boldsymbol{x}) = \pi_{-j}(\boldsymbol{y}_{-j})\pi_{j|-j}(y_j \mid \boldsymbol{y}_{-j})M_j(\boldsymbol{y}, \boldsymbol{x})\pi_{j|-j}(x_j \mid \boldsymbol{y}_{-j}).$$

If $M_j(\boldsymbol{x}, \boldsymbol{y}) = 0$ then $\pi(\boldsymbol{x})P_j(\boldsymbol{x} \to \boldsymbol{y}) = \pi(\boldsymbol{y})P_j(\boldsymbol{y} \to \boldsymbol{x}) = 0$. If $M_j(\boldsymbol{x}, \boldsymbol{y}) = 1$ then $\boldsymbol{y}_{-j} = \boldsymbol{x}_{-j}$ and once again $\pi(\boldsymbol{x})P_j(\boldsymbol{x} \to \boldsymbol{y}) = \pi(\boldsymbol{y})P_j(\boldsymbol{y} \to \boldsymbol{x})$.

The random scan Gibbs sampler has transition matrix $\bar{P} = (1/d)\sum_{j=1}^{d} P_j$. It therefore satisfies detailed balance by Proposition 11.2. Taking every d'th value from the systematic scan Gibbs sampler yields a homogenous Markov chain with transition matrix $P_{1:d} = P_d P_{d-1} \cdots P_2 P_1$. This matrix does not necessarily satisfy detailed balance with respect to π.

To see the failure, consider sampling with $\pi = (1/3, 1/3, 1/3)$, the uniform distribution on the set $\{A, B, C\} \in [0, 1]^2$ with $A = (0, 1)$, $B = (0, 0)$ and

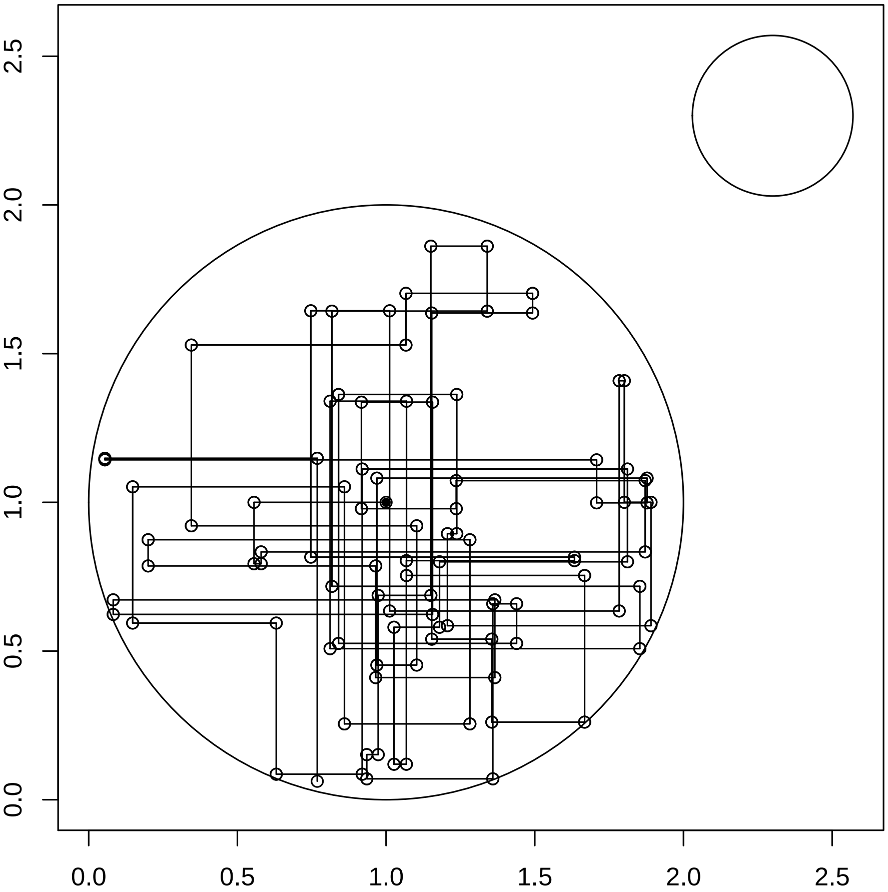

Figure 12.3: This figure shows a setting where the Gibbs sampler is a reducible Markov chain. The target distribution π is uniform within the union of the two circles shown. The sampler starts at the center of the larger circle, and proceeds for 100 steps. It can never enter the smaller circle. Had it started in the smaller circle, it could have never entered the larger one.

$C = (1, 0)$. The single component transition matrices here are

$$P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and so

$$P_{1:2} = P_2 P_1 = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

Now $\pi(C)\widetilde{P}(C \to A) = 0$ but $\pi(A)\widetilde{P}(A \to C) = (1/3)(1/4) = 1/12$. Even though P_1 and P_2 both have detailed balance, their product P_2P_1 does not.

A lack of detailed balance is not of itself a critical flaw in a sampling method. Some central limit theorems for MCMC in §**??** require detailed balance but others do not.

We can easily change the systematic scan Gibbs sampler so that it does have detailed balance. We use the following Lemma.

Lemma 12.1. *If the transition matrices P and Q both have detailed balance for the distribution π, then so does $T = PQP$.*

Proof. Using sums of i and j over all states,

$$\begin{aligned}
\pi(\boldsymbol{x})T(\boldsymbol{x} \to \boldsymbol{y}) &= \sum_i \sum_j \pi(\boldsymbol{x})P(\boldsymbol{x} \to i)Q(i \to j)P(j \to \boldsymbol{y}) \\
&= \sum_i \sum_j P(i \to \boldsymbol{x})\pi(i)Q(i \to j)P(j \to \boldsymbol{y}) \\
&= \sum_i \sum_j P(i \to \boldsymbol{x})Q(j \to i)\pi(j)P(j \to \boldsymbol{y}) \\
&= \sum_i \sum_j P(i \to \boldsymbol{x})Q(j \to i)P(\boldsymbol{y} \to j)\pi(\boldsymbol{y}) \\
&= \pi(\boldsymbol{y})\sum_i \sum_j P(\boldsymbol{y} \to j)Q(j \to i)P(i \to \boldsymbol{x}) \\
&= \pi(\boldsymbol{y})T(\boldsymbol{y} \to \boldsymbol{x}). \qquad \square
\end{aligned}$$

Corollary 12.1. *In Gibbs sampling, the matrix*

$$P_{1:d:1} = P_1P_2 \dots P_{d-1}P_dP_{d-1} \dots P_2P_1 \tag{12.4}$$

has detailed balance with respect to the stationary distribution π.

Proof. We apply Lemma 12.1 $d-1$ times, starting with P_d on the inside and multiplying the result on the left and right by P_{d-i} at step $i = 1, \dots, d-1$. □

The detailed balance version of Gibbs sampling updates x_j for j increasing from 1 to d and then decreasing back to $j = 1$. By equation (11.31) we could also use

$$P_{1:d,d:1} = P_1P_2 \dots P_dP_d \dots P_2P_1. \tag{12.5}$$

Because P_d is a full conditional distribution sampling twice in a row gives the same distribution as sampling once. That is $P_d^2 = P_d$, and so (12.5) gives the same transition distribution as (12.4). The (12.5) version still has detailed balance even for non-Gibbs samplers where $P_d^2 \neq P_d$ (see Exercise 12.1). By the same argument $P_1^2 = P_1$. As a result, we do not need to both start and end the update with P_1. We can instead use

$$P_{2:d:1} = P_1P_2 \dots P_{d-1}P_dP_{d-1} \dots P_2$$

repeatedly.

There is another strategy to get detailed balance in a sampler based on full conditional distributions. Letting P_j update x_j from its full conditional distribution, we can make a block of d updates by first putting $P_1, \dots, P_d$ in one of $d!$ random orders (chosen from a uniform distribution ν of $1, \dots, d$), and then applying those updates in the chosen random order. The resulting transition matrix is

$$\frac{1}{d!}\sum_{\nu} P_\nu, \quad \text{where} \quad P_\nu = P_{\nu(d)}P_{\nu(d-1)}\cdots P_{\nu(2)}P_{\nu(1)},$$

where the sum is over all permutations ν of 1 through d. See Exercise 12.2.

12.5 Correlated components

We saw in Figure 12.3 that the Gibbs sampler can fail to be irreducible when the space is disconnected. A more common worry with the Gibbs sampler is that it can slow down when two or more of the component variables are strongly correlated.

The simplest example to show this problem is the bivariate Gaussian. Suppose that $(x_1, x_2) \sim N\bigl(\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right), \left(\begin{smallmatrix}1&\rho\\\rho&1\end{smallmatrix}\right)\bigr)$. The systematic scan Gibbs sampler update for odd i is

$$\begin{aligned}x_{i1} &= \rho x_{i-1,2} + \sqrt{1-\rho^2}\, z_i\\ x_{i2} &= x_{i-1,2}\end{aligned}$$

where $z_i \sim \mathcal{N}(0,1)$. When i is even then x_{i2} changes and x_{i1} remains the same. Combining two steps, we get

$$\begin{aligned}x_{i+2,1} &= \rho x_{i+1,2} + \sqrt{1-\rho^2}\, z_{i+2}\\ &= \rho\bigl(\rho x_{i,1} + \sqrt{1-\rho^2}\, z_{i+1}\bigr) + \sqrt{1-\rho^2}\, z_{i+2}\\ &= \rho^2 x_{i,1} + \rho\sqrt{1-\rho^2}\, z_{i+1} + \sqrt{1-\rho^2}\, z_{i+2}\\ &= \rho^2 x_{i,1} + \sqrt{1-\rho^4}\, \tilde{z}_{i+2}\end{aligned}$$

where $\tilde{z}_{i+2} \sim \mathcal{N}(0,1)$. If ρ is close to 1 then $x_{i+2,1}$ is close to $x_{i,1}$ and the chain does not move much. If ρ is close to -1 the picture is a bit more subtle, but still not good.

For this bivariate Gaussian, the joint distribution of two consecutive new values in component 1 is

$$\begin{pmatrix}x_{i,1}\\ x_{i+2,1}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho^2\\ \rho^2 & 1\end{pmatrix}\right).$$

If the correlation is extremely strong, then it will slow down the rate at which the chain mixes. On the other hand, a mild correlation will not be very detrimental.

As in the reducible case, this example would benefit from a reparametrization in terms of less correlated random variables. Here we know enough about the problem to see that once again encoding as sum and difference would be advantageous. In general, finding a good encoding is a bit of an art.

Consider for example a Bayesian approach to a simple linear regression model. We have pairs (z_i, y_i) for $i = 1, \dots, n$ with $z_i \in \mathbb{R}$ fixed and $y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. To focus on the correlation issue, let's suppose that σ is known. Exercise xxx asks you to incorporate unknown σ into the model. The unknown parameters can be bundled together as $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$. In this context we replace the state variable x by β and consider

$$\pi(\beta) \propto \exp\Bigl(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 z_i)^2\Bigr),$$

which is a Gaussian likelihood times a non-informative (constant) prior for β.

The least squares estimate of β is $\hat\beta$ with $\hat\beta_1 = \sum_{i=1}^n y_i(z_i - \bar z)/\sum_{i=1}^n (z_i - \bar z)^2$ and $\hat\beta_0 = \bar y - \hat\beta_1 \bar z$. Of course $\bar z$ and $\bar y$ are sample means of z_i and y_i respectively. We assume that the denominator of $\hat\beta_1$ is positive to rule out a trivial problem.

With a little linear algebra we find that the posterior distribution $\pi(\beta)$ corresponds to $\beta \sim \mathcal{N}(\hat\beta, \Sigma)$, where

$$\Sigma = \frac{1}{n\sigma^2}\begin{pmatrix} 1 & \bar z \\ \bar z & \overline{z^2}\end{pmatrix}^{-1},$$

for $\overline{z^2} = \sum_{i=1}^n z_i^2/n$. The posterior correlation between β_0 and β_1 is then $\rho = -\bar z/\sqrt{\overline{z^2}}$. Letting $\sigma_z^2 = \sum_{i=1}^n (z_i - \bar z)^2/n$ we find $\rho = -\bar z/\sqrt{\bar z^2 + \sigma_z^2}$. As a result if $|\bar z| \gg \sigma_z$ then the Gibbs sampler will mix very slowly, while if $|\bar z| \ll \sigma_z$ it will mix very quickly.

Seeing this, we decide to recode the model as $y_i = \alpha + \beta_1(z_i - \bar z)$ where of course $\alpha = \beta_0 + \beta_1 \bar z$. This is a standard centering often done for numerical reasons in linear regression. Here it results in a posterior correlation of 0 for α and β_1 regardless of $\bar z$. The lesson is that we can run the Gibbs sampler on the state space (α, β_1). Then if we really want the posterior distribution of β_0 we can estimate it by applying the transformation $\beta_0 = \alpha - \beta_1 \bar z$ to all of the sampled (α, β_1) vectors.

We do not need to use the Gibbs sampler when we have a normal posterior distribution for β. However, we may often find that a simple linear transformation of two or more variables in the parameter space reduces correlations and helps the Gibbs sampler mix faster.

12.6 Gibbs for mixture models

The Gibbs sampler is very useful in mixture models. Figure 12.4 shows some astronomical data on the speed of galaxies relative to our own. The speeds are given in units of the speed of light. The top histogram has velocities of 82

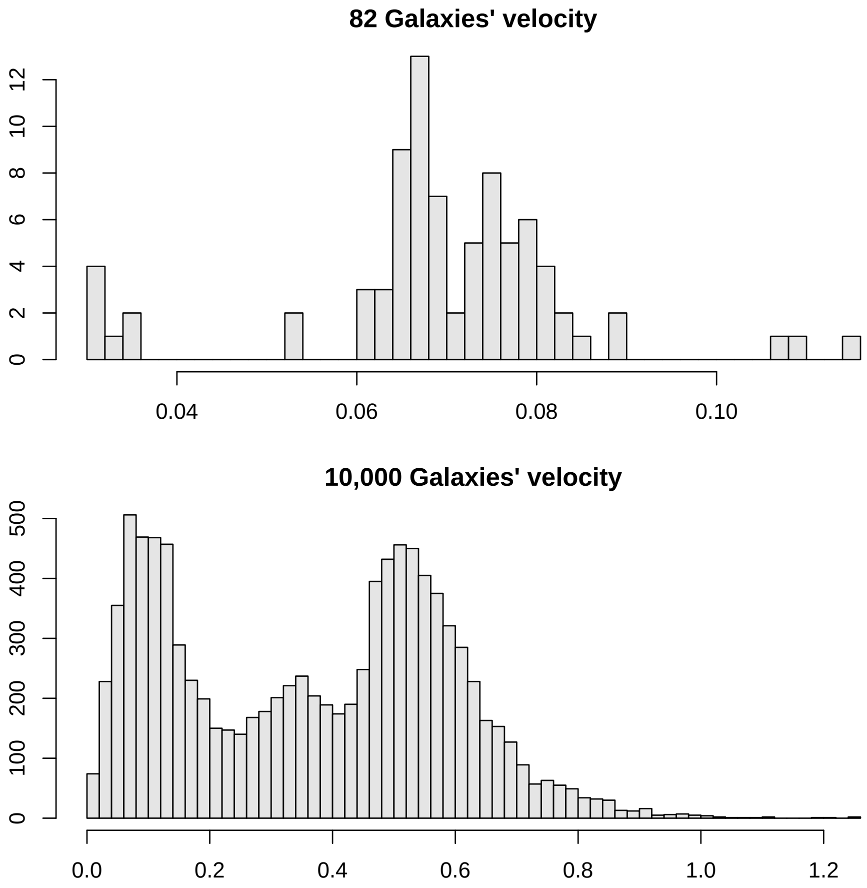

Figure 12.4: Galaxy velocity data.

galaxies from a famous data set used in mixture modeling (Roeder, 1990). The bottom histogram has 10,000 galaxies from the Sloan digital sky survey. Some of those speeds are greater than one, which is possible because space is expanding. The histograms show clear bumps. Perhaps the speeds are of several different kinds, each with its own nearly Gaussian distribution. Some comments on the scientific background are on page 72 of the end notes.

We investigate Gaussian mixture modeling of the galaxy velocity data in §12.7. First, we look at the Gibbs sampler for a mixture of binomials because it is a simpler problem. Suppose that $y_i \sim \mathrm{Bin}(m_i, p_i)$ for $i = 1, \dots, n$. Here y_i is the number of successes in m_i tries with success probability p_i. Now suppose that $p_i \in \{\gamma_1, \dots, \gamma_k\} \subset [0, 1]$, so there are k possible success probabilities. We don't know the γ_j values or which of them applies to any specific y_i. Instead, the p_i are independent random variables with $\mathbb{P}(p_i = \gamma_j) = \alpha_j > 0$ for $j = 1, \dots, k$

where $\sum_{j=1}^k \alpha_j = 1$. We write that as $\boldsymbol{\alpha} \equiv (\alpha_1, \dots, \alpha_k) \in \Delta^{k-1}$. Recall that $\Delta^{k-1} = \{\boldsymbol{x} \in [0,1]^k \mid \sum_{j=1}^k x_j = 1\}$.

In a Bayesian formulation we choose a prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_k) \in [0,1]^k$. Then we study the joint distribution of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ given $\boldsymbol{y}$ by sampling from

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = p(\boldsymbol{\alpha}, \boldsymbol{\gamma} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}) p(\boldsymbol{\alpha}, \boldsymbol{\gamma})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}) p(\boldsymbol{\alpha}, \boldsymbol{\gamma}).$$

Notice that $p(\boldsymbol{y}) = \iint p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}) p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) \,\mathrm{d}\boldsymbol{\alpha} \,\mathrm{d}\boldsymbol{\gamma}$ does not depend on $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ because they get integrated out. We will remove constants of proportionality from π without introducing new notation each time. The distinction between $\pi(\cdot)$ and $p(\cdot)$ is that π denotes a distribution that we will be interested in sampling from while p denotes distributions that we use to construct π.

The likelihood of this data is

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \prod_{i=1}^n \sum_{j=1}^k \alpha_j \binom{m_i}{y_i} \gamma_j^{y_i} (1 - \gamma_j)^{m_i - y_i}. \tag{12.6}$$

A simple and popular prior distribution for the parameters has $\boldsymbol{\gamma} \overset{\text{iid}}{\sim} \mathrm{Beta}(a, b)$ independently of $\boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{c})$, for $\boldsymbol{c} = (c_1, \dots, c_k)$. When there is no reason to use different c_j, we take $\boldsymbol{\alpha} \sim \mathrm{Dir}(c, c, \dots, c)$ for some $c > 0$. Under this model, $\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ is proportional to

$$\prod_{j=1}^k \alpha_j^{c-1} \times \prod_{j=1}^k \gamma_j^{a-1} (1 - \gamma_j)^{b-1} \times \prod_{i=1}^n \sum_{j=1}^k \alpha_j \binom{m_i}{y_i} \gamma_j^{y_i} (1 - \gamma_j)^{m_i - y_i}. \tag{12.7}$$

We will assume specified values for a, b and c, though they can also be made random in hierarchical models. The special case $a = b = c = 1$ is particularly simple, having $\gamma_j \sim \mathbf{U}(0, 1)$ and $\boldsymbol{\alpha} \sim \mathbf{U}(\Delta^{k-1})$.

Equation (12.7) includes a product of a sum and is awkward to sample from. It turns out that we can simplify the problem by data aumentatation, introducing latent categorical variables $C_i \in \{1, 2, \dots, k\}$ where $C_i = j$ if and only if y_i came from the group with success probability γ_j. In some formulas it is convenient to represent these categorical variable through z_{ij} where $z_{ij} = 1$ if $C_i = j$ and is 0 otherwise. Let Z represent all the z_{ij}. Under our model, the rows of Z are independent with $p(z_{ij} = 1 \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \alpha_j$. If we can sample from $p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, Z \mid \boldsymbol{y})$, then we get our desired sample of $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ along with some potentially useful information about Z.

Now $\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, Z) = p(\boldsymbol{\alpha}, \boldsymbol{\gamma}, Z \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, Z) p(\boldsymbol{\alpha}, \boldsymbol{\gamma}) p(Z \mid \boldsymbol{\alpha}, \boldsymbol{\gamma})$. By construction

$$p(Z \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \prod_{i=1}^n \prod_{j=1}^k \alpha_j^{z_{ij}}, \quad \text{and}$$

$$p(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, Z) = \prod_{i=1}^n \binom{m_i}{y_i} \gamma_{c_i}^{y_i} (1 - \gamma_{c_i})^{m_i - y_i} = \prod_{i=1}^n \binom{m_i}{y_i} \prod_{j=1}^k \Bigl[\gamma_j^{y_i} (1 - \gamma_j)^{m_i - y_i}\Bigr]^{z_{ij}}.$$

The product for observation i in the last step has $k-1$ ones in it along with the desired factor for $j = C_i$. Putting this together

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\gamma}, Z) \propto \prod_{j=1}^{k} \gamma_j^{a-1}(1-\gamma_j)^{b-1}\alpha_j^{c-1} \prod_{i=1}^{n}\prod_{j=1}^{k}\Bigl[\alpha_j\gamma_j^{y_i}(1-\gamma_j)^{m_i-y_i}\Bigr]^{z_{ij}}. \quad (12.8)$$

To get the full conditional distribution for $\boldsymbol{\alpha}$ given $\boldsymbol{\gamma}$ and Z, we freeze those values and find that $\pi(\boldsymbol{\alpha} \mid \boldsymbol{\gamma}, Z) \propto \prod_{j=1}^{k} \alpha_j^{c+z_{\bullet j}}$ where $z_{\bullet j} = \sum_{i=1}^{n} z_{ij}$. That is

$$\pi(\boldsymbol{\alpha} \mid \boldsymbol{\gamma}, Z) = \mathrm{Dir}(c + z_{\bullet 1}, \dots, c + z_{\bullet k}).$$

See §5.4 for a method to sample from the Dirichlet distribution. Letting $y_{\bullet j} = \sum_{i=1}^{n} z_{ij}y_i$ and $m_{\bullet j} = \sum_{i=1}^{n} z_{ij}m_i$ we find from (12.8) that γ_j given $\boldsymbol{\alpha}$ and Z are independent $\mathrm{Beta}(a + y_{\bullet j}, b + m_{\bullet j} - y_{\bullet j})$ distributions. Similarly the full conditional for Z has independent rows $(z_{i1}, \dots, z_{ik})$ with where $z_{ij} = 1$ (i.e., $C_i = j$) with probability proportional to $\alpha_j\gamma_j^{y_i}(1-\gamma_j)^{m_i-y_i}$. It can be sampled by the max Gumbel trick of §4.6. Exercise 12.7 asks you to work out these full conditional distributions with slightly more general prior distributions.

These updates are all easy to do because the full conditional distributions are in the same parametric family as the prior distributions, just with updated parameters. When the posterior distribution is in the same parameteric family as the prior distribution, then that prior is called a ***conjugate prior***. It is conjugate to the distribution defining the likelihood of the parameter.

We can generalize the setting as follows. Suppose that

$$y \sim \sum_{j=1}^{k} \alpha_j q(\cdot; \theta_j, \nu)$$

where q describes a probability density or mass function, θ_j is a parameter for mixture component j and ν is present if there are parameters common to all k mixture components. Then we want to sample from

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\theta}, \nu, Z) \propto p(\nu)\prod_{j=1}^{k}\alpha_j^{c-1}\prod_{j=1}^{k}p(\theta_j)\prod_{i=1}^{n}\prod_{j=1}^{k}\Bigl[\alpha_j q(y_i; \theta_j, \nu)\Bigr]^{z_{ij}}.$$

The distribution q could be more complicated, such as a regression model that relates some components of y to some other components. Depending on the functional forms of $p(\theta_j)$, $p(\nu)$, and q, we might find that some or all of the full conditional distributions are amenable to sampling.

12.7 Example: galaxy velocities

Now we turn to a mixture of Gaussians model for the galaxy data. A very natural model for this data has $y \sim \sum_{j=1}^{k} \alpha_j \mathcal{N}(\mu_j, \sigma_j^2)$, where $\alpha_j > 0$ with

$\sum_{j=1}^k \alpha_j = 1$ for $\mu_j \in \mathbb{R}$ and $\sigma_j > 0$. The model is a mixture of k different Gaussian distributions. The smaller data set in Figure 12.4 is commonly fit as a mixture of up to 6 Gaussians. The larger data set has two clear modes and it appears that there may be a third mode between them. We will consider $k = 3$ for it.

We consider data y_i generated as follows:

$$
\begin{aligned}
y_i &\sim \sum_{j=1}^k \alpha_j \mathcal{N}(\mu_j, \sigma_j^2), \quad \text{for} \quad \mu_j \sim \mathcal{N}(\mu_*, \tau^2),\\
\sigma_j^{-2} &\sim \mathrm{Gam}(\beta)/(\beta\sigma_*^2) \quad \text{and} \quad \boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m) \sim \mathrm{Dir}(\boldsymbol{c}).
\end{aligned}
$$

We will pick constant values for μ_*, τ^2, β, σ_*^2 and $\boldsymbol{c}$. Here μ_* can be thought of typical values of μ_j and because $\mathbb{E}(\mathrm{Gam}(\beta)) = \beta$, σ_*^2 is similarly a typical value of σ_j^2. It will prove simpler to work with $A_j = 1/\sigma_j^2$, $A = (A_1, \dots, A_k)$ and $B = 1/\tau^2$.

Now $\pi(\boldsymbol{\alpha}, \boldsymbol{\mu}, A, Z)$ is proportional to

$$
\prod_{j=1}^k \Bigl(\alpha_j^{c-1} e^{-A_j\beta\sigma_0^2} A_j^{\beta-1} e^{-(B/2)(\mu_j-\mu_*)^2}\Bigr) \times \prod_{i=1}^n \prod_{j=1}^k \Bigl[\alpha_j A_j^{1/2} e^{-(A_j/2)(y_i-\mu_j)^2}\Bigr]^{z_{ij}}.
$$

To find the full conditional distributions, let $z_{\bullet j} = \sum_{i=1}^n z_{ij}$ and $y_{\bullet j} = \sum_{i=1}^n z_{ij} y_{ij}$ as in the binomial example. It is natural to think in terms of $\bar{y}_{\bullet j} = y_{\bullet j}/z_{\bullet j}$ but that becomes an awkward 0/0 if $z_{\bullet j} = 0$, so we use $y_{\bullet j}$ and $z_{\bullet j}$ separately. We also use $\mathrm{SS}_j = \sum_{i=1}^n z_{ij}(y_i - \mu_j)^2$. The full conditional distributions are

$$
\begin{aligned}
\pi(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, A, Z) &\sim \mathrm{Dir}(c + z_{\bullet j})_{j=1}^k\\
\pi(z_{i1}, \dots, z_{ik} \mid \boldsymbol{\alpha}, \boldsymbol{\mu}, A) &\sim \mathrm{Cat}(\alpha_j e^{-A_j(y_i-\mu_j)^2/2})_{j=1}^k,\\
\pi(\mu_j \mid \boldsymbol{\alpha}, A, Z) &\sim \mathcal{N}\Bigl(\frac{A_j y_{\bullet j} + B\mu_*}{A_j z_{\bullet j} + B}, (A_j z_{\bullet j} + B)^{-1}\Bigr), \quad \text{and}\\
\pi(A_j \mid \boldsymbol{\alpha}, \boldsymbol{\mu}, Z) &\sim \mathrm{Gam}(\beta + z_{\bullet j}/2) \Big/ \Bigl(\beta\sigma_0^2 + \frac{1}{2}\mathrm{SS}_j\Bigr).
\end{aligned}
\tag{12.9}
$$

In the full conditional distributions, μ_j are independent of each other and so are the n rows of Z as well as the A_j.

The update for μ_j weights each of $z_{\bullet j}$ responses y_i with $Z_{ij} = 1$ by A_j and weights μ_* by B. It is as if we had seen B/A_j observations equal to μ_* in component j. A small positive value of B serves to keep μ_j well defined numerically.

Models that mix Gaussians are more tricky than mixtures of discrete distributions and also more tricky than single Gaussian models. A popular model for a single Gaussian distribution takes a non-informative flat prior on the mean. Doing that for the Gaussian mixture would involve having $\tau^2 = \infty$, or equivalently $B = 0$. It is possible to get $z_{\bullet j} = 0$ and in that case having $B = 0$ leaves the full conditional distribution for μ_j undefined: $\mathcal{N}(0/0, 1/0)$. Similarly, a popular choice for the variance of a Gaussian mixture component has $\beta = 0$. In the mixture context that would make a degenerate full conditional distribution of $\mathrm{Gam}(0)/0$ for A_j if $z_{\bullet j} = 0$. The full conditional for A_j is proportional to $A_j^{\beta-1/2}e^{-A_j(\beta\sigma_0^2+(1/2)\mathrm{SS}_j)}$. If $\beta > 0$ and $\sigma_0^2 > 0$ then this distribution is proper even if $z_{\bullet j} = 0$ (which makes $\mathrm{SS}_j = 0$). Neither of these problems come up for a model with just one Gaussian component and $n \geqslant 1$ observations. We can make a model with just one Gaussian by taking $k = 1$ in the mixture model. Then $z_{\bullet 1} = n > 0$ always holds.

A second difficulty with Gaussian mixtures is that the likelihood is unbounded. If $\mu_j = y_i$ for some i and j, then the likelihood is unbounded as $\sigma_j \to 0$. That poses a greater difficulty for maximum likelihood estimation than it does for Bayesian inference. Even if the posterior density is unbounded it can still be a proper distribution, and in this case all of the full conditional densities are bounded if B, β and σ_0^2 are all positive.

Figure 12.5 shows traces of $n = 5000$ iterations of the Gibbs sampler for a Gaussian mixture fit to the 10,000 Sloan survey galaxy velocities. There are $k = 3$ components. The parameters were set as follows: μ_0 was the average velocity, $1/B = \tau^2$ was the sample variance of the velocities, the mixture probabilities had a uniform distribution given by taking $c = 1$, The prior on σ_j^2 used $\beta = 1/10$. Taking $\beta = 1/1000$ was quite similar. The initial state was found by taking the lower, middle, and upper thirds of the velocities to be the three groups. That is $C_i = 1$ for the smallest third, $C_i = 2$ for the middle third and $C_i = 3$ for the largest third.

From the top trace we see that the curves for μ_2 and μ_3 crossed somewhere around the 500'th iteration. There is no astronomical distinction to be made between calling a set of stars group two or calling that same set group three. We revisit this point in §12.8 on the label switching problem.

Figure 12.6 shows in a solid curve the average of $\sum_{j=1}^{3} \alpha_j\varphi((x-\mu_j)/\sigma_j)/\sigma_j$ over iterations 1000 through 5000 for μ_j, σ_j and α_j, skipping the first 999 iterations. The result has a clear bimodal structure. The two components with larger values of μ_j are close enough together to form just one mode, so the estimate does not pick up the small third mode that the histogram in Figure 12.4 shows. Repeating the simulation with $k = 6$ produced the dashed curve in Figure 12.6. It has four modes including one roughly where that histogram had a third bump, but it does not resemble the histogram. In hindsight, using a Gaussian mixture model is a form of curve fitting, and with small k there are only $3k - 1$ effective parameters to fit it with (noting that α_k is determined by $\alpha_1, \dots, \alpha_{k-1}$). A fit that optimizes a numerical criterion like a likelihood or posterior mean might not correspond to what we see by eye. Among the reasons for this, a mixture of components from some special family (here Gaussian) might fit poorly if the data have mixture components outside that family. Mixtures remain a powerful technique for model building even if in this instance the result does not do what we might have wanted.

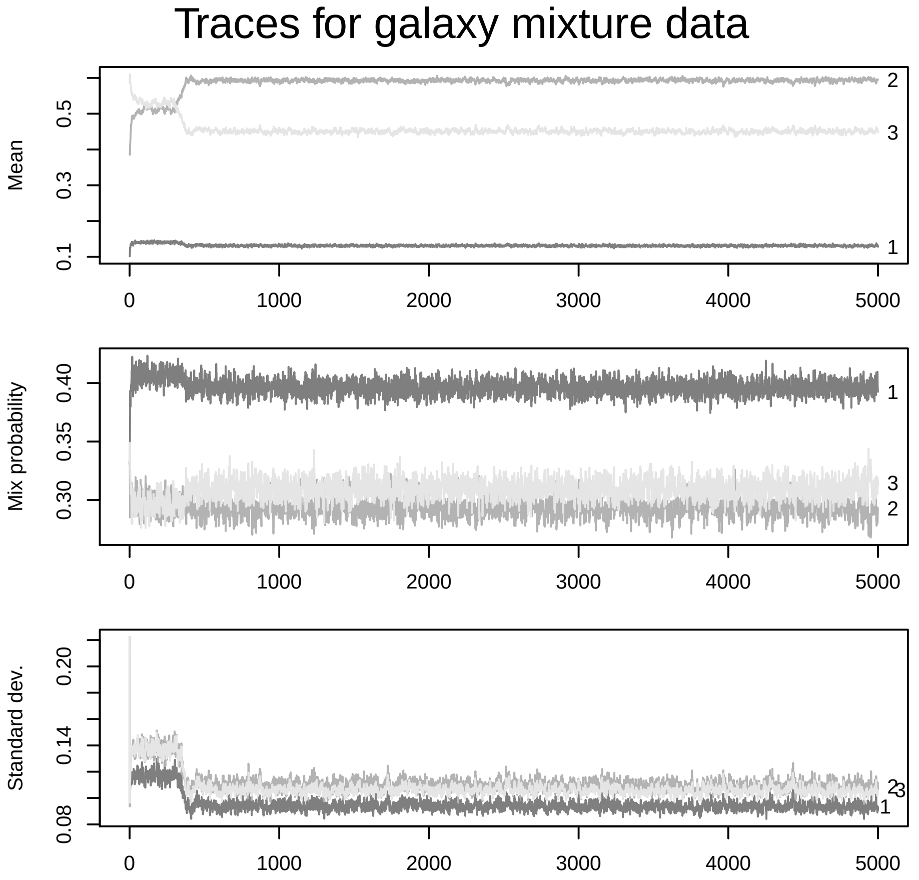

Figure 12.5: These are traces of μ_j, α_j, and $\sigma_j = A_j^{-1/2}$, for 5000 Gibbs steps fitting a Gaussian mixture model to the larger data set in Figure 12.4.

12.8 Label switching

Let $\tau(1), \tau(2), \dots, \tau(k)$ be a permutation of $1, 2, \dots, k$. There is no difference between mixture models

$$y_i \overset{\text{iid}}{\sim} \sum_{j=1}^{k} \alpha_j p(\cdot \,|\, \phi_j), \quad \text{and} \quad y_i \overset{\text{iid}}{\sim} \sum_{j=1}^{k} \alpha_{\tau(j)} p(\cdot \,|\, \phi_{\tau(j)}) \tag{12.10}$$

where we have simply switched the labels. Any mixture model for y_i must have a likelihood with $p(\boldsymbol{y} \,|\, \alpha, \phi)$ where applying the same permutation to the indices in both α and ϕ makes no difference. Equation (12.10) gives no special meaning to component 1 of the mixture. If the model has a prior $p(\alpha, \phi)$ that is also invariant to label permutations, then the posterior distribution $\pi(\alpha, \phi) \propto p(\alpha, \phi \,|\, \boldsymbol{y})$ will

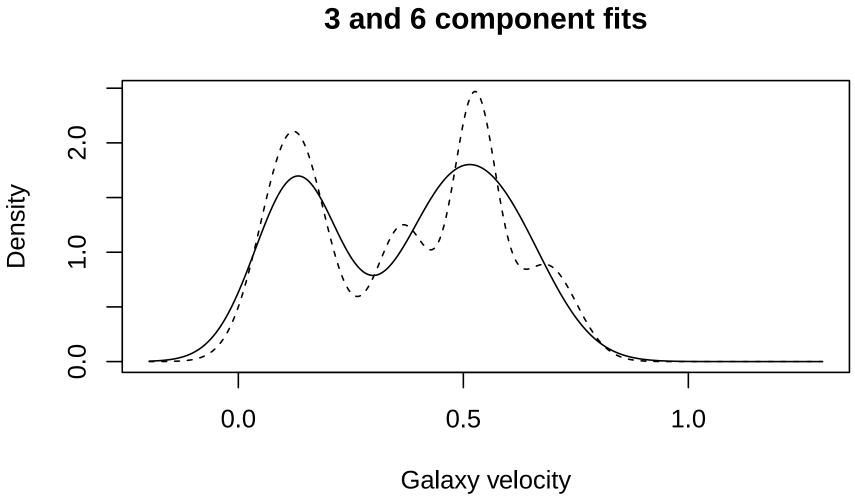

Figure 12.6: Estimated mixture density using $k = 3$ and $k = 6$ mixture components, with a burn-in of 999.

also be invariant. For the galaxy velocity problem, $\phi_j = (\mu_j, \sigma_j)$ has exactly the same posterior distribution as (μ_j, σ_j) for all $j = 2, \dots, k$. Similarly, all of the α_j must have mean $1/k$ regardless of the data. This ***label switching*** problem means that it is problematic to even give meaning to component $j = 1$, and that difficulty extends to interpreting sampling output related to that component.

Some aspects of the posterior distribution of the mixture model are not affected by label switching. For example, the estimated density for galaxy velocities at level y is

$$\sum_{j=1}^{k} \alpha_j \, \varphi\Bigl(\frac{y - \mu_j}{\sigma_j}\Bigr) \; / \; \sigma_j.$$

This quantity is invariant to permutation of the component labels, just like the posterior distribution is, and label switching is not a difficulty for it.

Some problems cannot be addressed by working only through invariant aspects of the posterior distribution. For instance, if one of the galaxies has a velocity of about 0.3 we might be interested to know which cluster it is likely to have come from. The same holds for a hypothetical galaxy that might later to be observed to have a velocity of 0.3. The first of these questions can be partially addressed by observing how often y_i was in the same cluster as $y_{i'}$, for each other galaxy i'. That is, we know a galaxy by the galaxies that co-clustered with it. In Gibbs sampling, that is how often $C_i = C_{i'}$ held. The number of pair equalities to track grows as the square of the number of galaxies, making it difficult to scale this approach. It is also quite indirect to answer a question about α, μ and σ in terms of co-clustering probabilities.

One way to break the label switching symmetry is to impose identifiability constraints. Let $\theta = (\alpha_1, \dots, \alpha_k, \phi_1, \dots, \phi_k)$ contain all parameters of the model. For a permutation τ of 1 through k let

$$\theta(\tau) = (\alpha_{\tau(1)}, \dots, \alpha_{\tau(k)}, \phi_{\tau(1)}, \dots, \phi_{\tau(k)})$$

be a reordering. To select one specific permutation we can impose a constraint like

$$\alpha_1 < \alpha_2 < \cdots < \alpha_k$$

on θ. We only lose a little generality in assuming that the mixture probabilities are all distinct. We could similarly impose a constraint on some aspect of θ_j assumed to be distinct. For instance, we might take

$$\mu_1 < \mu_2 < \cdots < \mu_k$$

in a mixture of univariate Gaussians.

To sample with an identification constraint is simple. After the MCMC algorithm has generated a parameter θ_i, inspect it to see if it obeys the constraint. If not, reorder it accordingly. It is extremely convenient that something so simple would work, and it almost seems too good to be true. See Stephens (1997, Section 3.1) for a proof.

Suppose that we reorder each θ_i so that corresponding $\mu_{i,1} < \mu_{i,2} < \cdots < \mu_{i,k}$. We might well obtain a multimodal histogram for one or more of the μ_j, even though a specific individual permutation has been selected. Jasra et al. (2005) show this taking place for μ_2 and μ_5 when the smaller galaxy data set is fit as a mixture of 6 Gaussians. The histogram of μ_5 values has a small secondary mode at the usual location for μ_6 and μ_2 has a small secondary mode at the usual location for μ_1.

In Figure 12.5, the parameters have settled into some mode of the posterior distribution after about 500 iterations. We know from invariance that there are $3! = 6$ modes corresponding to that mode that are equally important. If our Markov chain had mixed properly, it would have visited all of those different modes. If those modes are very well defined and separated by regions of extremely low posterior probability then the sampler might have no realistic chance of visiting two or more of them. It remains possible that the Markov chain has found one of those 6 modes and explored it thoroughly, though of course traces do not prove that. Mixing through all $k!$ versions of the posterior distribution is an exceedingly high bar. Celeux et al. (2000) write: “Although we may be somewhat presumptuous, we consider that almost the entirety of MCMC samplers implemented for mixture models has failed to converge!”

In the initial transient in Figure 12.5 the values of μ_2 and μ_3 have swapped positions. That suggests that one label switch has taken place there. The values of α_2 and α_3 also switch there. Inspecting σ_2 and σ_3 makes it clear that what happens instead is that the sampler is still moving from the starting position towards one of the six labellings during that apparent switch.

12.9 The slice sampler

In acceptance-rejection sampling, we sample a point uniformly in a region bounded above by the density function and bounded below by zero. Then discarding the 'height' component we obtain a sample from the target density. That same idea can be put to use in MCMC. In the ***slice sampler*** we run a Gibbs sampler for the uniform distribution on the region underneath $\pi(\boldsymbol{x})$. It is enough to sample the region underneath an unnormalized version $\pi_u(\boldsymbol{x})$. The region under π_u has $Z = \int \pi_u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ times the height of that under π.

The process is illustrated in Figure 12.7 for the one dimensional unnormalized density

$$\pi_u(x) = \frac{1}{\sigma}\max\Bigl(\varphi\Bigl(\frac{x+1}{\sigma}\Bigr),\ \frac{5}{2}\,\varphi\Bigl(\frac{x}{\sigma}\Bigr),\ 3\varphi\Bigl(\frac{x-1}{\sigma}\Bigr)\Bigr), \tag{12.11}$$

over the interval $[-2, 2]$, where $\sigma = 0.15$. The sampler starts at $(1, 4)$. The horizontal line at height 4 intersects the density in four places. The function π_u is larger than 4 on $[-0.151, 0.151] \cup [0.824, 1.176]$ and the next point is chosen uniformly from this union of two intervals. First one of the intervals is sampled with probability proportional to length. Then a uniformly distributed point is taken from within that interval. The vertical moves are easier, being sampled uniformly on $[0, \pi_u(x)]$.

The awkward step in the slice sampler is finding the intervals and their endpoints. In general this requires a numerical search. The simple example in equation (12.11) can be done without search, but is still tricky because the number of intervals can be 1 or 2 or 3.

A one dimensional example is easy for illustration but in one dimensional problems we may well be able to sample from π in a more direct way. Suppose now that π_u is nonzero on a region $\mathcal{X} \subseteq \mathbb{R}^d$. Now construct the $d+1$ dimensional region

$$\mathcal{R} = \{(\boldsymbol{x}, z) \mid \boldsymbol{x} \in \mathcal{X},\ 0 \leqslant z \leqslant \pi_u(\boldsymbol{x})\}.$$

If $(\boldsymbol{X}, Z) \sim \mathbf{U}(\mathcal{R})$ then $\boldsymbol{X} \sim \pi$. To run the Gibbs sampler within $\mathcal{R}$ we need $d+1$ full conditional distributions. The one for Z is simplest, it is $U[0, \pi_u(\boldsymbol{x})]$.

To describe the full conditional distribution for x_j let

$$\boldsymbol{x}_{-j} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$$

as before. Let $z\!:\!\boldsymbol{x}_{-j}$ be the point with k'th coordinate x_k for $k \neq j$ and j'th coordinate z. This is the point we get inserting z into the j'th position of x. The full conditional distribution for x_j is uniform on the slice

$$S_j(\boldsymbol{x}) = \{z \mid \pi_u(z\!:\!\boldsymbol{x}_{-j}) \geqslant \pi_u(\boldsymbol{x})\,\} \subset \mathbb{R}.$$

This set is ordinarily the union of one or more line segments. In real problems, we expect a small number of short intervals, but mathematically some pathological examples are possible. See Exercise 12.6.

The slice sampler has some good theoretical properties, as described in Chapter xxx, but it is very awkward to implement in general. If the function π_u is

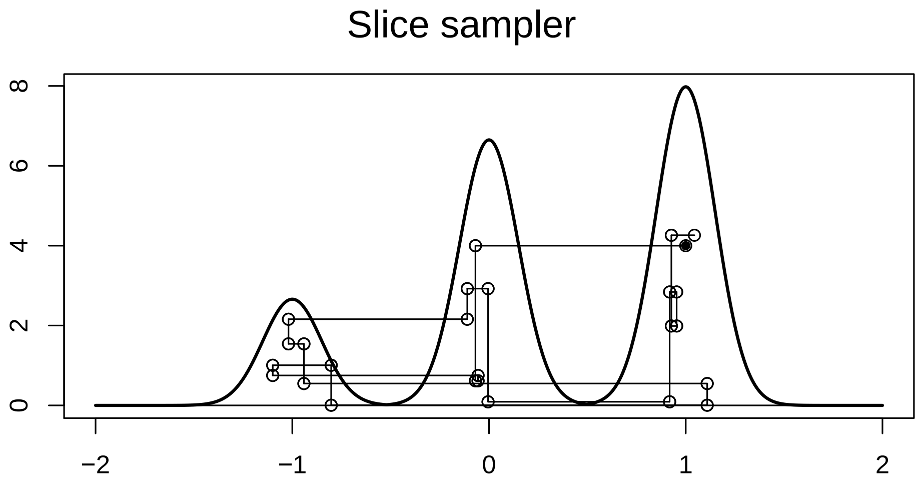

Figure 12.7: This figure illustrates the slice sampler. The unnormalized density π_u is given by the thick line. The slice sampler starts at the solid point $(1, 4)$ and from there executes 25 steps of the Gibbs sampler for the uniform distribution under π_u over the interval $[-2, 2]$.

unimodal in x_j for all j and all $\boldsymbol{x}_{-j}$, then we can run two bisection or other iterative searches to find the interval $[L, R]$ where $\pi_u(z : \boldsymbol{x}_{-j}) \geqslant \pi_u(\boldsymbol{x})$ if and only if $L \leqslant z \leqslant R$. It could take quite a few function evaluations to precisely determine L and R, and after all that work, only one value from within (L, R) will be used. Furthermore, for non-unimodal densities there may be no assurance that the desired region is just one solid interval.

It is not strictly necessary to identify the desired slice and sample within it. Another approach works with a randomly grown interval $[L, R]$ containing x_j that is somewhat longer than the slice to be sampled. That is π_u falls below $\pi_u(\boldsymbol{x})$ when x_j is either increased to R or decreased to L. Given such an interval we then randomly choose a point within $S_j(\boldsymbol{x}_{-j}) \cap [L, R]$. The stepping out algorithm Algorithm 12.2 generates one such random interval $[L, R]$.

Given the output $[L, R]$ of the stepping out algorithm, the simple way to sample within $[L, R] \cap S_j(\boldsymbol{x}_{-j})$ is to repeatedly draw $\mathbf{U}[L, R]$ variables until one of them is in $S_j(\boldsymbol{x}_{-j})$. This can be wasteful though if the intersection is small. For well behaved π_u we can expect that the region around x_{0j} at least belongs to the slice. Then if we have trouble sampling from the desired intersection, we can start to sample nearer to the starting point x_{0j} as in Algorithm 12.3.

Algorithms 12.2 and 12.3 are very carefully tuned so that taken together they preserve detailed balance with respect to π. They are compatible with each other, but if an alternative is chosen for one, then the other may need to be modified. See Chapter xxx.

The result of using these algorithms is to make the slice sampler much more

Algorithm 12.2 Stepping out procedure for the slice sampler

```
**Stepping out procedure** ( $x_0$, $\pi_u$, $j$, $w$, $m$ )

// $x_0 \in \mathbb{R}^d$ is the starting point
// $\pi_u$ is an unnormalized density
// $j \in \{1, \dots, d\}$ is the component to update
// $w > 0$ is an initial estimated slice size
// $m$ is a positive integer limit on iterations

$y \leftarrow \pi_u(x_0)$
$U \sim \mathbf{U}(0,1)$
$L \leftarrow x_{0j} - wU$
$R \leftarrow L + w$
$V \sim \mathbf{U}(0,1)$
$J \leftarrow \lfloor mV \rfloor$
$K \leftarrow m - 1 - J$

while $J > 0$ and $\pi_u(L\!:\!x_{-j}) > y$ do
   $L \leftarrow L - w$
   $J \leftarrow J - 1$
while $K > 0$ and $\pi_u(R\!:\!x_{-j}) > y$ do
   $R \leftarrow R + w$
   $K \leftarrow K - 1$
return  $[L, R]$
```

This algorithm finds an interval $[L, R]$ containing the j'the component x_{0j} of the point x_0 for use in the slice sampler. The interval it produces may be sampled from by Algorithm 12.2.

automatic. There is still the practical issue of picking the width $w = w_j$ to use for the j'th variable. As such the algorithm is similar to using Metropolis within Gibbs where we would have to decide on the scale of a proposal distribution for slice j. But where Metropolis within Gibbs would sometimes leave x_{0j} unchanged the modified slice sampler will change each component of $\boldsymbol{x}$ every time.

12.10 Variations

12.10.1 Grouped and collapsed Gibbs

The Gibbs sampler as presented updates each component from the full conditional distribution given the other components. Suppose that we can group the components into subsets $s_\ell \subset \{1, \dots, d\}$ for $\ell = 1, \dots, L$ with $s_\ell \cap s_k = \varnothing$ for $\ell \neq k$ and $\cup_{\ell=1}^{L} s_\ell = \{1, \dots, d\}$. Now if we are able to sample component ℓ from it's full conditional distribution given the other components, then we can base

Algorithm 12.3 Shrinkage sampling procedure for the slice sampler

```
**Shrinking slice sampler** ( $\boldsymbol{x}_0$, $\pi_u$, $j$, $L$, $R$ )

// $\boldsymbol{x}_0 \in \mathbb{R}^d$ is the starting point
// $(L, R)$ is an interval containing $\boldsymbol{x}_{0j}$ within which to sample
// $\pi_u$ is an unnormalized density
// $j \in \{1, \dots, d\}$ is the component to update

$y \leftarrow \pi_u(\boldsymbol{x}_0)$
$\widetilde{L} \leftarrow L$
$\widetilde{R} \leftarrow R$

**loop**
  $U \sim \mathbf{U}(0, 1)$
  $z \leftarrow \widetilde{L} + U \times (\widetilde{R} - \widetilde{L})$
  **if** $\pi_u(z\!:\!\boldsymbol{x}_{-j}) > y$ **then**
    **return** $z$
  **if** $z < x_{0j}$ **then**
    $\widetilde{L} \leftarrow z$
  **else**
    $\widetilde{L} \leftarrow z$
```

This algorithm samples a point within the intersection of $[L, R]$ and the slice $S_j(x_{-j})$. It is compatible with intervals (L, R) generated by Algorithm 12.2.

a Gibbs sampler on this decomposition.

12.10.2 Gibbs and Metropolis hybrids

Sometimes we can sample most but not all of the full conditional distributions needed for the Gibbs sampler. For example suppose that we can sample from $\pi_{j|-j}$ for every j from 1 to d inclusive except for one value j^*. The ***Metropolis within Gibbs*** algorithm uses a Metropolis-Hastings update in place of the missing j^*'th full conditional distribution. Specifically let

$$z \sim Q(\cdot \mid \boldsymbol{x}_{-j^*}), \qquad\qquad y_j = \begin{cases} x_j & j \neq j^* \\ z & j = j^*, \end{cases}$$

$$U \sim \mathbf{U}(0,1), \quad \text{and} \qquad\qquad A = \min\left(1, \frac{\pi(\boldsymbol{y})Q(z \mid \boldsymbol{x}_{-j^*})}{\pi(\boldsymbol{x})Q(x_{j^*} \mid \boldsymbol{x}_{-j^*})}\right)$$

and then accept $\boldsymbol{y}$ if and only if $U \leqslant A$.

Metropolis-within-Gibbs can be used for more than one of the components of $\boldsymbol{x}$. It can also be used for one or more of the components in the grouped Gibbs sampler. In fact, the particle sampling in the original Metropolis paper was a systematic scan Gibbs sampler (cycling over the 224 circular disks) using what we now call Metropolis within Gibbs (with uniform $[-\alpha,\alpha]^2$ proposals) to update the variables in groups of two (each group being the coordinates of one disk's center point).

The ***Metropolized Gibbs sampler*** is another such combination. Suppose that the full conditional distribution of x_j given x_{-j} is discrete. To be definite, suppose that this distribution takes the value z_k with probability $p_k > 0$ for $k = 1,\dots,M$. We can also suppose that no two of the z_k are equal to each other and that $\sum_{k=1}^M p_k = 1$ (i.e., our list is complete). One of the M values equals x_j, suppose to be definite that $z_1 = x_j$.

One way to view the Gibbs sampler is as a Metropolis sampler with $\mathbb{P}(y = z_k) = p_k$ and acceptance probability 1. Although the acceptance probability is 1 the probability that the chain moves to a new location is only $1 - p_1$ because the proposal to move to z_1 is really a proposal to remain in place. In the Metropolized Gibbs sampler we make a proposal that specifically excludes $y = x_j = z_1$. We take $Q(z_1 \to z_k) = p_k/(1-p_1)$. Then we use the Metropolis-Hastings acceptance probability

$$A(z_1 \to z_k) = \min\Bigl(1, \frac{p_k p_1 (1-p_k)^{-1}}{p_1 p_k (1-p_1)^{-1}}\Bigr) = \min\Bigl(1, \frac{1-p_1}{1-p_k}\Bigr).$$

The probability of accepting the proposal is

$$\begin{aligned}\sum_{k=2}^m \frac{p_k}{1-p_1}\min\Bigl(1,\frac{1-p_1}{1-p_k}\Bigr) &= \sum_{k=2}^m p_k \min\Bigl(\frac{1}{1-p_1},\frac{1}{1-p_k}\Bigr)\\ &> \sum_{k=2}^m p_k = 1-p_1.\end{aligned}$$

As a result the Metropolized Gibbs sampler has strictly greater chance of moving to a new value of x_j than the Gibbs sampler has.

Example 12.1. Here we replace one portion of a prior by a non-conjugate distribution.

12.10.3 Hit and run sampling

The Gibbs sampler works by sampling along coordinate directions. It failed to connect the two circular regions in the example of Chapter xxx because there was no 'stepping stone' between them. If we could sample in directions other than coordinate directions, we might well find a direct path from one circle to the other.

Exercise xxx considers rotating the space to a more favorable orientation and then running the Gibbs sampler in the new coordinate system. However in any given example we might not know which rotation to use. It then is natural to try random rotations in the hope that a good one will be found, or at least that a bad one will not be used every time.

In hit and run sampling we choose a random direction given by a unit vector $\theta \in \Omega = \{u \in \mathbb{R}^d \mid u^{\mathsf{T}}u = 1\}$. The simplest version has $\theta \sim \mathbf{U}(\Omega)$, but non-uniform distributions can also be used. Then we draw a sample from π along the line in direction θ.

Starting at $\boldsymbol{x}_0$, we repeat

$$\theta_i \sim \mathbf{U}(\Omega) \quad \text{and} \quad \boldsymbol{x}_i = \boldsymbol{x}_{i-1} + R_i\theta_i, \quad \text{where}$$
$$\mathbb{P}(R_i \leqslant r_0) = \frac{\int_{-\infty}^{r_0} \pi(\boldsymbol{x}_i + r\theta_i)\,\mathrm{d}r}{\int_{-\infty}^{\infty} \pi(\boldsymbol{x}_i + r\theta_i)\,\mathrm{d}r}.$$

It is clear that the hard part is sampling R_i along the desired line.

If the two circular disks were replaced by two balls in $\mathbb{R}^d$ for large d, then it would not matter what angle connected the two balls, eventually the hit and run sampler would connect them. It is also true however, that the $d-1$ dimensional angle of success between those two regions could be extremely small when d is large.

12.11 Example: volume of a polytope

Hit and run is the most competitive way to compute the volume of a high dimensional polytope.

Chapter end notes

Data augmentation methods increase the dimension of a problem while being designed to make the new higher dimensional problem easier to handle. Incorporating the new variables must not change the distribution of the original variables. Then from a sample of the enlarged ensemble, we may if we like, simply ignore the new variables. Or, should those variables have a useful interpretation, we can study them too. Tanner and Wong (1987) introduced data augmentation for sampling posterior distributions. It is similar to the expectation-maximization (EM) algorithm of Dempster et al. (1977) which is widely used for maximizing likelihoods. See van Dyk and Meng (2001) for a survey of data augmentation.

Galaxy speeds

The large data set of galaxy speeds is a random sample from about 500,000 galaxy speeds from the Sloan survey. I thank Eric Ford for sending it to me. Professor Ford points out that clustering in astronomical data can be due to differences in regions of the sky, differences in what was of inte rest to people composing the data set and even differences in the properties of the measuring devices. Finding or even counting bumps in a histogram is an interesting statistical and computational challenge but translating that into a conclusion about the cosmos should only be done in collaboration with an astronomer familiar with the specific sky surveys being used.

Mixtures and label switching

There are many nuanced issues in mixture modeling especially via Bayesian methods. See Frühwirth-Schnatter (2006) and McLachlan and Peel (2000) and references within these two books. Much has been written about the label switching problem and other aspects of Bayesian analysis of mixture models. See for instance Geweke (2007), Stephens (2000), McLachlan and Peel (2000, Chapter 4) and references in those works. Perhaps mixtures of log concave densities (Chang and Walther, 2007; Balabdaoui and Doss, 2018) could be used to counter the lack of fit we saw between the galaxy velocity data and the mixture of Gaussians model.

Slice sampler

The slice sampler was developed by Neal (2003). That article also includes the stepping out algorithms. In addition to the stepping out algorithm, there is another one that repeatedly doubles the interval width, instead of adding length w each time.

Additional references

Metropolis within Gibbs was in a Bayesian approach to generalized linear models by Zeger and K. (1991). The Metropolized Gibbs sampler is due to Liu (1996). The example of centering regression is from Gilks and Roberts (1996).

Exercises

12.1. Let P_m for $m = 1, \dots, M$ preserve detailed balance with respect to π. Show that $P_{1:M,M:1} = P_1 P_2 \dots P_M P_M \dots P_2 P_1$ of equation (11.31) preserves detailed balance too.

12.2. For $M \geqslant 2$ let $P_1, \dots, P_M$ be transition matrices that preserve detailed balance with respect to π. Let P_* be a transition matrix formed as $P_{\nu(1)} P_{\nu(2)} \cdots P_{\nu(M)}$ where ν is a uniform random permutation of $1, \dots, M$. The randomness in ν is a part of P_* and in usage a different random ν would be repeatedly generated. **Hint:** proving it first for $M = 2$ might be useful.

12.3. The example in §12.4 used the two disks

$$D_1 = \{(x_1, x_2) \mid (x_1 - 1)^2 + (x_2 - 1)^2 \leqslant 1\}, \quad \text{and}$$
$$D_2 = \{(x_1, x_2) \mid (x_1 - 2.3)^2 + (x_2 - 2.3)^2 \leqslant 0.27^2\}.$$

The Gibbs sampler for the distribution $\pi = U\bigl(D_1 \cup D_2\bigr)$ is reducible. Now let

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{for} \quad A = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

- **a)** Give a concise expression for the distribution of y.
- **b)** Find an expression for x in terms of y.
- **c)** Run the Gibbs sampler for 10,000 steps on the distribution of y. Start at the value $y_0 = Ax_0$ where $x_0 = (1, 1)$. Translate the sampled values y_i into corresponding values x_i. Plot the trajectory taken by the first 100 points. Report the sample mean of x over the first 10,000 points. Show the histogram of the first component of the sampled x points.

12.4. For $i = 1, \dots, n$, let $y_i = \beta_0 + \beta_1 z_i + \varepsilon_i$ where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Suppose that the prior distribution on (β_0, β_1) is non-informative. Let σ^{-2} have prior distribution Gamma(3) independent of (β_0, β_1).

- **a)** Write down the unnormalized posterior distribution of $\theta = (\beta_0, \beta_1, \sigma)$.
- **b)** Exhibit the three full conditional distributions for the parameter vector θ.
- **c)** For the xxx data run the Gibbs sampler and plot the acf for β_1.
- **d)** Recode the model, replacing $(\beta_0, \beta_1, \sigma)$ by $(\alpha, \beta_1, \sigma)$ where $\alpha = \beta_0 - \beta_1 \bar{z}$. Now show the likelihood, the full conditional distributions and the acf for β_1.

12.5. For the Gaussian probit model of §12.3, find the conditional distribution of β given $\boldsymbol{w}$ when the prior for β is $\mathcal{N}(\beta_0, \Sigma)$ for known β_0 and known full rank matrix Σ. Extend this to a mixture prior

$$p(\beta) \sim \sum_{j=1}^{k} \alpha_j \mathcal{N}(\beta_{0j}, \Sigma_j)$$

for known $\alpha_j > 0$ that sum to 1 and known β_{0j} and Σ_j.

12.6. Here we construct some awkward cases for the slice sampler.

- **a)** Find a density $\pi(x)$ on $x \in \mathbb{R}^2$ where $\{x_1 \mid \pi(x_1, 0) \geqslant \pi(0, 0)\} = \mathbb{R}$.
- **b)** Find a density $\pi(x)$ on $x \in \mathbb{R}^2$ where $\{x_1 \mid \pi(x_1, 0) \geqslant \pi(0, 0)\}$ is

$$\bigcup_{p \in \mathcal{P}} [p, p + 1)$$

where $\mathcal{P} = \{2, 3, 5, 7, \dots\}$ is the set of prime numbers.

12.7. For the mixture of binomials problem determine the full conditional distributions when the prior for $\boldsymbol{\gamma}$ has independent $\gamma_j \sim \text{Beta}(a_j, b_j)$ these are independent of $\boldsymbol{\alpha} \sim \text{Dir}(c_1, \dots, c_k)$.

12.8. Find a Gibbs sampler for a mixture of $\text{Poi}(\lambda_j)$ distributions. For the prior on λ_j use a Gamma distribution with shape α and rate β.

Bibliography

Balabdaoui, F. and Doss, C. R. (2018). Inference for a two-component mixture of symmetric distributions under log-concavity. *Bernoulli*, 24(2):1053–1071.

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electon plasma. *Australian Journal of Physics*, 18(2):119–123.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Chang, G. T. and Walther, G. (2007). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, 51(12):6242–6251.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Durrett, R. (1999). *Essentials of Stochastic Processes.* Springer, New York.

Feller, W. F. (1968). *Introduction to Probability Theory and it's Applications*, volume I. Wiley, New York, 3rd edition.

Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics*, 38(2):1034–1070.

Flegal, J. M. and Jones, G. L. (2011). Implementing MCMC: estimating with confidence. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov chain Monte Carlo*, pages 175–197. Chapman & Hall, Boca Raton, FL.

Fosdick, L. D. (1959). Calculation of order parameters in a binary alloy by the Monte Carlo method. *Physical Review*, 116(3):565.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media, New York.

Gelman, A., Roberts, G. O., and Gilks, W. O. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics*, volume 5, pages 599–607. Oxford University Press, Oxford.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical science*, 7(4):457–472.

Gelman, A. and Shirley, K. (2011). Inference from simulations and monitoring convergence. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov chain Monte Carlo*, pages 163–174. Chapman and Hall/CRC, Boca Raton, FL.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis*, 51(7):3529–3550.

Geyer, C. (2011). Introduction to Markov chain Monte Carlo. In Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov chain Monte Carlo*, pages 3–48. Chapman and Hall/CRC, Boca Raton, FL.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In Keramides, E. M., editor, *Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation of North America.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science*, 7(2):473–483.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving mcmc. In Gilks, W. R., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, volume 6, pages 89–114. Chapman & Hall/CRC, Boca Raton, FL.

Glynn, P. W. and Whitt, W. (1991). Estimating the asymptotic variance with batch means. *Operations Research Letters*, 10(8):431–435.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234.

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

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, pages 50–67.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics*, 25(5):2084–2102.

Levin, D. A., Peres, Y., and Wilmer, E. L. (2009). *Markov chains and mixing times*. American Mathematical Society, Providence, RI.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and computing*, 6(2):113–119.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, 48(3):188–190.

MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, New York.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1091.

Neal, R. M. (2003). Slice sampling. *The annals of statistics*, 31(3):705–767.

Newman, M. E. J. and Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics*. Oxford University Press, New York.

Norris, J. R. (1998). *Markov Chains*. Cambridge University Press, Cambridge.

Onsager, L. (1944). Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Reviews*, 65(3–4):117–149.

Owen, A. B. (2017). Statistically efficient thinning of a Markov chain sampler. *Journal of Computational and Graphical Statistics*, 26(3):738–744.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612.

Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624.

Schmeiser, B. (1982). Batch size effects in the analysis of simulation output. *Operations Research*, 30(3):556–568.

Serre, D. (2002). *Matrices: theory and applications*. Springer, New York.

Snell, J. L. and Kindermann, R. (1980). *Markov Random Fields and Their Applications*. American Mathematical Society, Providence, RI.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. PhD thesis, University of Oxford.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, 62(4):795–809.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728.

van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of computational and graphical statistics*, 10(1):1–50.

Vats, D., Flegal, J. M., and Jones, G. L. (2015). Multivariate output analysis for Markov chain Monte Carlo. Technical Report arXiv:1512.07713, University of Minnesota.

Zeger, S. L. and K., M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American statistical association*, 86(413):79–86.