ST911 Fundamentals of Statistical Inference
Part III

Gareth Roberts
Department of Statistics,
University of Warwick

Contact details:
Room D1.04 (Statistics Department)
e-mail: Gareth.O.Roberts@warwick.ac.uk
webpage: www2.warwick.ac.uk/fac/sci/statistics/staff/academic-research/roberts/

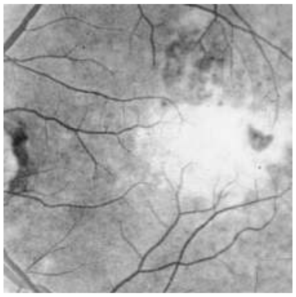
Disclaimer: These lecture notes may cover material that is not covered in the lecture. They may also omit some material that is discussed in the lectures! Further, these notes may contain typos. If you are in doubt about anything, please feel free to ask me. If you find any typos please let me know!

1 Introduction

Let's start of with some examples in which Monte Carlo methods are being used.

1. **Image restoration**:
   Consider a binary (black-and-white) image x with 128×128 pixels, that is x is a 128^2 long array with entries $x_i \in \{0, 1\}, i \in \{1, \ldots, 128^2\}$. We receive y, a noisy version of x corrupted by multiplicative noise: the value at each pixel is flipped with probability p independent of any other pixel. Further, we know how binary pixel images tend to look like: reasonably large clusters of white or black pixels. Monte Carlo methods can be used to estimate the true image x. For example, we can determine the mode of the posterior distribution, that is the image $\hat{x}$ for which the noisy observation x is most likely.

2. **Image interpretation**:
   Image restoration is a low-level task, but often we are interested in high-level information such as characterisation of the objects visible in an image. For example, consider the image below which shows vascular structure in a retina. Researchers at Warwick used Monte Carlo methods to extract the vascular structure and quantify the uncertainty associated with this extraction.



3. **Bioinformatics:**
   Genome sequencing projects produce sequences of letters from the 4-letter alphabet A,T, C and G (the so-called nucleotides). But how do we interpret these sequences? Because of the sheer amount of data but also because of the vast subject knowledge, sophisticated methods that incorporate this knowledge are needed to analyze these sequences. For example, of interest are common pattern or "words" of a certain length w in multiple sequences. These can give use clues about functionality such as gene regulation. However, we don't know the exact location of these words and, worse, the words may contain "typos" due to mutation!

4. **Bayesian Statistics:**
   Suppose we have a likelihood model $f(y|\theta)$ and a prior density $\pi(\theta)$, then the posterior density is given by

   $$\pi(\theta|y) \quad \propto \quad f(y|\theta)\pi(\theta).$$

   However, it may be difficult to compute the appopriate normalizing constant of the posterior density. In this module we discuss simulation methods that can sample from the posterior model without explicit computation of the normalizing constant.

2 Foundations of stochastic simulation

2.1 Monte Carlo inference

Suppose we would like to compute the probability of winning Minesweeper using a pre-specified play strategy. To do this analytically we would have to enumerate all possible ways of planting bombs in a given grid. Given an $n \times n$-grid and k bombs that would make $(n^2)!/k!(n^2-k)!$ configurations! Then for all possible starting conditions we would have to determine which of these bomb configurations would lead to winning or losing the game if we adopt the given strategy. Sounds like a lot of work, doesn't it? And it gets worse as the grid gets bigger!

Well, a savvy statistician would adopt the following approach. Instead of looking at all possible bomb configurations, he would randomly choose some configurations and then estimate the winning probability. In more detail, he would choose some large number N (which would still be smaller than all possible bomb configurations) and then would sample N bomb configurations $X_1, \ldots, X_N$ and determine which would lead to a loss and which to a win. He then would estimate the probability of winning as $\hat{p} = W/N$ where W is the number of games that were won. How would he sample the bomb configurations? Well, he could repeatedly call the minesweeper program or write a computer program to do so.

In this first section of the module we will learn a bit about techniques for sampling standard distributions. But before we do this, let's first look at the scientific method called ”Monte Carlo estimation”. Most numerical problems fall into one of two categories:

1. optimisation, or
2. integration.

In this module we will focus mainly on the latter as statistical inference is usually related to integration. Just recall that both expectations and probabilities are derived from integrals (or sums)! So consider the following integral

$$I = \int_0^1 h(x)dx$$

How would you solve this problem numerically without resorting to simulation methods? Well, one approach would be Riemann summation. We evaluate the function $h(x)$ at n points $(x^{(1)}, \ldots, x^{(n)})$ in a regular grid and then compute

$$I \approx \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)}).$$

Monte Carlo estimation proceeds differently. We start by re-writing the integral as

$$I = \int_0^1 \frac{h(x)}{f(x)} f(x)dx$$

where $f(x)$ is a density on $[0, 1]$ such that if $h(x) \neq 0$ then $f(x) > 0$. But this means that

$$I = \mathbb{E}_f(h(X)/f(X)),$$

where $\mathbb{E}_f$ stands for expectation with respect to the distribution specified by f. We now produce an iid sample $(x^{(1)}, \ldots, x^{(n)})$ from the distribution specified by the density f and set

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})/f(x^{(i)}).$$

The law of large numbers tells us that $\hat{I}_n$ converges with probability 1 to the integral I as n tends to infinity. Moreover, the Central Limit Theorem states that

$$(\hat{I}_n - I)/\sqrt{\mathrm{Var}(\hat{I}_n)}$$

is approximately Standard Normal. So the variance $\mathrm{Var}(\hat{I}_n)$ tell us about the accuracy of our estimate and it can be estimated as

$$v_n \quad = \quad \frac{1}{n(n-1)} \sum_{j=1}^{n} (h(x_j)/f(x_j) - \hat{I}_n)^2$$

2.2 Basic principles

2.2.1 Random numbers

Starting point of any stochastic simulation is a random number generator. Random number generators start with some initial value x_0, the so-called *seed*, and repeatedly apply a deterministic function f to it to obtain a sequence $x_i = f^i(x_0), i = 1, 2, \ldots$. This sequence cannot be truly ”random” because it is produced by a deterministic function. However, it imitates the behaviour of a sample from a Uniform(0,1) random variable in the following sense. Consider the sequence $(x_1, \ldots, x_n) \in (0, 1)$ and apply a statistical test for the departure from independence and/or the uniform distribution.

Definition 1 *The sequence $(x_1, \ldots, x_n)$ is called a sequence of* (pseudo-) random numbers *if statistical tests for the departure from independence and the uniform distribution are not rejected more often than expected by chance.*

There is vast literature about the use and misuse of random numbers as well as a multitude of different random number generators. In the following we assume we have a “good” random number generator available.

2.2.2 Inverse Transformation method

Theorem 1 *Consider the cumulative distribution function (cdf) $F(x)$. Let F^{-1} be the generalized inverse of F, that is*

$$F^{-1}(u) \quad = \quad \min\{x \in S : \ F(x) \geq u\} \qquad u \in (0, 1]$$

Let U be a Uniform(0,1) random variable and set $X = F^{-1}(U)$, then the distribution of X has cdf $F(x)$.
(Note that for a continuous cdf the generalized inverse is just the usual inverse).

Sketch Proof: By the definition of the generalized inverse and monotonicity of F we have

$$P(X \leq x) \quad = \quad \mathbb{P}(F^{-1}(U) \leq x) \quad = \quad P(U \leq F(x)) \quad = \quad F(x).$$

Example 1 Simulating an Exponential random variable with parameter λ
An $Exponential(\lambda)$ distributed random variable has cdf

$$F(x) \quad = \quad 1 - \exp(-\lambda x) \qquad \textit{for } x \geq 0.$$

Let $U \sim \mathrm{Uniform}(0,1)$ and set

$$Y \quad = \quad -\frac{1}{\lambda}\log(1-U).$$

Then Y has an Exponential distribution with parameter λ. This can be further simplified by noting that $1-U$ is also Uniform(0,1) and so

$$Y \quad = \quad -\frac{1}{\lambda}\log(U)$$

has an Exponential(λ) distribution.

Example 2 Bernoulli(p) trial and Binomial B(n,p) random variable
Let U be a Uniform(0,1) random variable. If we set

$$X \quad = \quad \begin{cases} 1 & \text{if } U < p \\ 0 & \text{otherwise} \end{cases}$$

then X is a Bernoulli trial with success probability p.

Let $X_1, \ldots, X_n$ be an iid sample of Bernoulli(p) trials, then $Y = \sum_{i=1}^{n} X_i$ has a Binomial(n,p) distribution.

Example 3 Geometric(p) random variable
Suppose X takes values in $\mathbb{N}$ and $\mathbb{P}(X = j) = p_j$. Then

$$F^{-1}(u) \quad = \quad \min\{j \in \mathbb{N} : u \leq \sum_{i=1}^{j} p_i\}.$$

Now, if $X \sim$ Geometric(p), then $\mathbb{P}(X > j) = (1-p)^j$. Thus,

$$\sum_{i=1}^{j} p_i \quad = \quad 1-(1-p)^j \quad \geq \quad u$$

if and only if

$$j \quad \geq \quad \frac{\log(1-u)}{\log(1-p)}$$

Let $\lceil a \rceil$ denote the ceiling of a, then $X = \left\lceil \frac{\log(U)}{\log(1-p)} \right\rceil$ has a Geometric(p) distribution.

2.2.3 Rejection sampling

Suppose we would like to sample X which is a continuous random variable with density function $f(x)$. We do not know how to sample from X but we do know how to sample from a similar random variable Y with density function $g(y)$. If we have that support(f) $\subseteq$ support(g) and that

$$f(x)/g(x) \quad \leq \quad M \qquad \text{for all } x$$

then we can use sampling from Y to produce a sample from X. We repeat the following steps until a sample is returned.

- Step 1: Sample $Y = y$ from $g(x)$ and $U = u$ from a Uniform(0,1) distribution. Go to Step 2.
- Step 2: If $u \leq \frac{f(y)}{M\ g(y)}$ return $X = y$. Otherwise repeat Step 1.

Proposition 1 *The distribution of the random variable X sampled in the above rejection sampling algorithm has density $f(x)$.*

Sketch Proof: We have

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}\Big(Y \leq x \mid U \leq \frac{f(Y)}{Mg(Y)}\Big) \\ &= \frac{P\Big(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\Big)}{\mathbb{P}\Big(U \leq \frac{f(Y)}{Mg(Y)}\Big)}. \end{aligned}$$

To compute the above we need the joint density of Y and U. By independence this is

$$h(y,u) \quad = \quad g(y)\ \mathbf{1}_{[0\leq u \leq 1]}.$$

Thus,

$$\begin{aligned} \mathbb{P}\Big(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\Big) &= \int_{-\infty}^{x} g(y) \int_{0}^{f(y)/Mg(y)} 1 du dy \\ &= \int_{-\infty}^{x} g(y)\frac{f(y)}{Mg(y)} dy \quad = \quad \frac{1}{M}\int_{-\infty}^{x} f(y)dy \end{aligned}$$

It follows that

$$\mathbb{P}\Big(U \leq \frac{f(Y)}{Mg(Y)}\Big) \quad = \quad \frac{1}{M}\int_{-\infty}^{\infty} f(y)dy \quad = \quad \frac{1}{M}$$

and so our claim follows as

$$\mathbb{P}(X \leq x) \quad = \quad \frac{P\Big(Y \leq x, U \leq \frac{f(Y)}{Mg(Y)}\Big)}{\mathbb{P}\Big(U \leq \frac{f(Y)}{Mg(Y)}\Big)} \quad = \quad \int_{-\infty}^{x} f(y)dy.$$

How many iterations of the algorithm do we need on average? Well, in each iteration we produce a sample with probability $\mathbb{P}(U \leq \frac{f(Y)}{Mg(Y)}) = \frac{1}{M}$ so the total number of iterations is Geometrically distributed with parameter $1/M$. Thus the mean number of iterations needed is M. Note the following:

1. The smaller the bound M the more efficient the algorithm in terms of number of iterations. Thus we should look for a density g close to f.

2. If the support of f is unbounded, then in order to be able to find a bound M the density g needs to have fatter tails than f.

Example 4 *Suppose we would like to sample $|X|$ where X is a standard Normal random variable. The density of $|X|$ is given by*

$$f(x) \quad = \quad \sqrt{\frac{2}{\pi}} \exp\Big(-\frac{x^2}{2}\Big) \qquad \textit{for } x \in \mathbb{R}^+.$$

We already know how to sample an Exponential random variable so let us choose as density g the density of an Exponential distribution with parameter 1. Then

$$\begin{aligned} \frac{f(x)}{g(x)} \quad &= \quad \sqrt{\frac{2}{\pi}} \exp\Big(-\frac{x^2-2x}{2}\Big) \quad = \quad \sqrt{\frac{2e}{\pi}} \exp\Big(-\frac{(x-1)^2}{2}\Big) \\ &\leq \quad \sqrt{\frac{2e}{\pi}}. \end{aligned}$$

So we set $M = \sqrt{\frac{2e}{\pi}}$ and hence

$$\frac{f(x)}{Mg(x)} \quad = \quad \exp\Big(-\frac{(x-1)^2}{2}\Big).$$

So the rejection sampling algorithm proceeds as follows:

- *Step 1: Sample $Y = y$ from an Exponential(1) distribution and $U = u$ from a Uniform(0,1) distribution. Go to Step 2.*
- *Step 2: If $u \leq \exp(-\frac{(y-1)^2}{2})$ return $X = y$. Otherwise repeat Step 1.*

Example 5 *Consider a random variable Y with density $g(x)$ defined on state space S. Now suppose $A \subset S$ and we would like to sample the conditional random variable $X = (Y|Y \in A)$ with state space A. In this case rejection sampling can be done by repeatedly sampling X until our sample lies in A. More formally, X has density $f(x) = \frac{g(x)}{\mathbb{P}(Y \in A)}$ for $x \in A$. Thus*

$$\frac{f(x)}{g(x)} \quad \leq \quad \frac{1}{\mathbb{P}(Y \in A)} = M \qquad \textit{and} \qquad \frac{f(x)}{Mg(x)} \quad = \quad \mathbf{1}_{[x \in A]} \quad \textit{for } x \in S,$$

Suppose now U is uniformly distributed on the unit interval. Then

$$\mathbb{P}(U \leq f(Y)/Mg(Y)) \quad = \quad \begin{cases} 1 & \textit{if } Y \in A \\ 0 & \textit{if } Y \notin A \end{cases}$$

Thus in the standard rejection sampling algorithm, we accept if $Y \in A$ and otherwise we reject. We don't need to sample U to make this decision.

If the evaluation of the target density f is very expensive, rejection sampling can be made computationally less expensive if additional to the upper bound $Mg(x)$ on the target density $f(x)$ we also have an easily evaluated lower bound $h(x)$. The so-called *envelope rejection sampling algorithm* proceeds as follows.

1. Sample $Y = y$ from $g(y)$ and $U = u$ from a Uniform(0,1) distribution.
2. Accept if $u \leq h(y)/Mg(y)$ and return $X = y$ as a sample. Otherwise go to Step 3.
3. Accept if $u \leq f(y)/Mg(y)$ and return $X = y$ as a sample. Otherwise go to Step 1.

This is more efficient because on average we need $1/M \int h(x)dx$ times less evaluations of f which are replaced by evaluations of h. The function h can be found for example by a Taylor expansion.

3 Importance sampling

In the previous section we encountered rejection sampling where we use a proposal density to produce samples from the target density. In this section we retain the samples of the proposal density but alter the estimation procedure to get unbiased estimates of the characteristics of the target density.

3.1 Standard importance sampling

Recall what we said when we discussed Monte Carlo inference! We are interested in the integral

$$I = \mathbb{E}_f(h(X)) = \int_S h(x)f(x)dx$$

where f is a density. Then we re-write the integral as

$$I = \int_S \frac{f(x)}{g(x)} h(x)g(x)dx$$

where g is a density such that $g(x) > 0$ whenever $f(x)h(x) \neq 0$. We now produce an iid sample $(x_1, \ldots, x_n)$ from g and estimate I as

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} h(x_i) = \frac{1}{n}\sum_{i=1}^{n} w(x_i)h(x_i).$$

We call this procedure importance sampling. The density g is called the proposal or instrumental density and the weights $w(x_i) = \frac{f(x_i)}{g(x_i)}$ are called *importance weights*. Note that $\hat{I}$ is an unbiased estimator of I.

There are two reasons why we might be interested in performing importance sampling:

1. sampling from $f(x)$ is not possible or too expensive;
2. $h(X)$, where $X \sim f$, has a large variance, so the conventional unbiased estimator has large Monte Carlo error.

The variance of an importance estimator will only be finite if the estimator is square integrable, that is

$$\mathbb{E}_g\Big(h^2(X)\frac{f^2(X)}{g^2(X)}\Big) \quad = \quad \mathbb{E}_f\Big(h^2(X)\frac{f(X)}{g(X)}\Big) \quad < \quad \infty$$

Thus the variance will often be infinite if the ratio $f(x)/g(x)$ is unbounded. Hence, if possible, we should choose a proposal density g that has thicker tails than f. Generally, if $f(x)/g(x)$ is unbounded, then even if the variance of the estimator is finite, the procedure is inefficient as the variance of importance weights is large. The importance weights of the sample will vary widely giving much relative weight to a few sample values.

Example 6 Normal tail probabilities
Let $p = \mathbb{P}(Z > 4)$ where Z is a standard Normal random variable. Using a naive approach we might produce an iid standard Normal sample $Z^{(1)}, \dots, Z^{(M)}$ and set

$$\hat{p} \quad = \quad \frac{1}{M}\sum_{i=1}^{M}\mathbf{1}_{[Z^{(i)}>4]}$$

However, M would have to be VERY large to get an estimate that differs from zero. An alternative is to produce an iid sample from an Exponential random variable of rate 1 translated to the right by 4. Then

$$w(x) \quad = \quad \frac{f(x)}{g(x)} \quad = \quad \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2 + (x-4)).$$

Instead of the importance estimator $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} w(x_i)h(x_i)$ the following ratio estimator is often used:

$$\tilde{I} \quad = \quad \frac{\sum_{j=1}^{n} h(x_j)w(x_j)}{\sum_{j=1}^{n} w(x_j)}.$$

This estimator has two advantages:

1. While it is biased, it usually has a smaller variance than the importance estimator introduced earlier. But note that this estimator is still consistent as for $x_1, \dots, x_n$ iid with density g we have

$$\frac{1}{n}\sum_{j=1}^{n} f(x_j)/g(x_j) \overset{n\to\infty}{\to} 1.$$

2. We can apply importance sampling even if we know $f(x)$ and thus $w(x)$ only up to a constant of proportionality.

If we cannot find an importance density that leads to a reasonably small variance of importance weights there are several procedures that may be adopted to reduce the variance.

1. The first approach is called *sequential importance resampling* and proceeds as follows.

   (a) Produce an importance sample $Y^{(1)}, \dots, Y^{(n)}$ with importance weights $w_i = f(Y^{(i)})/g(Y^{(i)}), i = 1, \dots, n$.

   (b) Produce a new sample $X^{(1)}, \dots, X^{(n)}$ by sampling from $Y^{(1)}, \dots, Y^{(n)}$ where $Y^{(j)}$ is sampled with probability $w_j / \sum_{i=1}^{n} w_i$.

2. The second procedure is called *rejection control* and considers discarding any sample points that have an importance weight below a given threshold c. Discarding sample points will introduce a bias, but by changing the importance weights appropriately this bias can be avoided. Given the importance sample $Y^{(1)}, \dots, Y^{(n)}$ with importance weights $w_1, \dots, w_n$ rejection control proceeds as follows:

   (a) For $j = 1, \dots, n$ accept $Y^{(j)}$ with probability $p_j = \min\{1, w_j/c\}$. Otherwise discard $Y^{(j)}$.

   (b) If $Y^{(j)}$ is accepted recompute its importance weight as $\tilde{w}_j = q w_j / p_j$ where $q = \int \min\{1, w(x)/c\} g(x) dx$.

   Note that because q is the same for all sample points, we do not need to compute it explicitly if we use the ratio estimator. Further note that rejection control produces an importance sample according to the proposal density

$$g^*(x) \quad = \quad \frac{\min\{g(x), f(x)/c\}}{q}.$$

4 Markov chain theory

In this section we review some of the concepts in Markov chain theory which are important for MCMC. Let us start with the basics and first define a Markov chain.

Definition 2 Markov chain. A random process $X = \{X_n, n = 0, 1, 2, \dots\}$ taking values in $\mathcal{S}$ is a *Markov chain*, if

$$\mathbb{P}\Big(X_{n+1} \in A \mid X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_0 = x_0\Big) \\ = \mathbb{P}\Big(X_{n+1} \in A \mid X_n = x_n\Big)$$

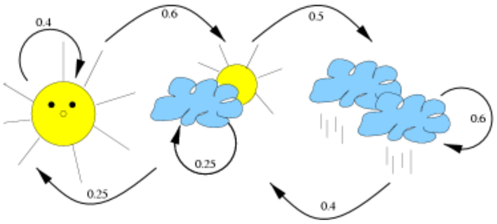
for all $n \geq 0$, $A \subseteq \mathcal{S}$ and $x_0, \dots, x_n \in \mathcal{S}$.

Thus, given we know the present value of the chain (at time n), the value the chain takes in the *future* (at time $n+1$) does not depend on values in the past (at times before n).

Example 7 **Weather.** Suppose X_n is the weather on day n. We set

$$X_n = \begin{cases} 0 & \text{if it is sunny on day } n \\ 1 & \text{if it is cloudy on day } n \\ 2 & \text{if it rains on day } n. \end{cases}$$

The figure below shows the transition probabilities for the weather chain.



By modelling the weather as a Markov chain, we assume that the weather tomorrow given the weather today does not depend on the weather of yesterday or any earlier day.

In this course we mostly consider time-homogeneous Markov chains, that is Markov chains whose transition probabilities do not depend on the time.

Definition 3 **Transition probabilities, time-homogeneity.** A Markov chain X is *time-homogeneous*, if its transition probabilities

$$\mathbb{P}\Big(X_{n+1} \in A \mid X_n = x\Big) \quad = \quad P(x, A) \quad = \quad \int_A p(x,y)dy$$

do not depend on n. We call $P(x, A)$ the transition kernel. Within this module we assume that the transition kernel is absolutely continuous for every $x \in \mathcal{S}$ that is, it has an associated density or probability mass function. Thus for fixed $x \in \mathcal{S}$ the function $p(x, y)$ is a density (or pmf). To ease terminology in the following we will refer to $p(x, y)$ as a density, but keep in mind that on a discrete state space it would be a probability mass function.
The n-step transition densities $p^{(n)}(x, y)$ of X are defined as

$$\mathbb{P}\Big(X_n \in A \mid X_0 = x\Big) \quad = \quad P^n(x, A) \quad = \quad \int_A p^{(n)}(x,y)dy$$

In the weather example, if we assume that the Markov chain is time-homogeneous then we assume that changes in the weather do not depend on the current month. This may not be very sensible as the weather behaves differently in different seasons. However, we make this common assumption because time-homogeneous Markov chains are much easier to deal with.

If the state space $\mathcal{S}$ of X is finite, then we can collect the transition probabilities in a transition matrix.

Definition 4 **Transition matrix.** Let $\mathbb{P}(X_{n+1} = j | X_n = i) = p_{ij}$. Then the transition matrix of X is given by

$$\mathbf{P} \quad = \quad (p_{ij})_{i,j \in \mathcal{S}}.$$

Task 1 *Determine the transition matrix for the weather chain.*
The transition matrix for the weather Markov chain is

$$\mathbf{P} = \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.25 & 0.25 & 0.5 \\ 0 & 0.4 & 0.6 \end{pmatrix}$$

The n-step transition probabilities can be computed from the transition matrix as follows

$$p_{ij}^{(n)} = \mathbf{P}^n(i,j).$$

Task 2 *Compute the 2-step transition matrix for the weather chain.*

$$\mathbf{P}^2 = \begin{pmatrix} 0.31 & 0.39 & 0.3 \\ 0.1625 & 0.4125 & 0.425 \\ 0.1 & 0.34 & 0.56 \end{pmatrix}$$

Lemma 1 Distribution at time n. Suppose the initial distribution of X, that is the distribution of X_0 is given by the density $q^{(0)}(x)$. Then we can compute the density of X at time n as follows

$$q^{(n)}(x) = \int_{\mathcal{S}} q^{(0)}(y)\, p^{(n)}(y,x)dy.$$

For a finite state space we may use matrix algebra to compute the distribution of X at time n. If $q^{(n)}$ is the vector of the distribution at time n and $\mathbf{P}^n$ the n-step transition matrix, then

$$q^{(n)} = q^{(0)}\, \mathbf{P}^n.$$

Task 3 *Suppose on day 0 it is sunny. Thus $q^{(0)} = (1,0,0)$. Then, the distribution of the weather on the day 2 is*

$$\begin{aligned} q^{(2)} &= q^{(0)}\mathbf{P}^2 \\ &= (1,0,0) \begin{pmatrix} 0.31 & 0.39 & 0.3 \\ 0.1625 & 0.4125 & 0.425 \\ 0.1 & 0.34 & 0.56 \end{pmatrix} \\ &= (0.31, 0.39, 0.3) \end{aligned}$$

Thus, if on day 0 it is sunny, we have a 31% chance of having sunny weather on day 2.

If a Markov chain satisfies certain regularity conditions then the distribution of the chain converges to a limit distribution which is also called the invariant, stationary or equilibrium distribution. Such a chain is called an *ergodic* Markov chain.

Task 4 *List possible reasons why a Markov chain may fail to be ergodic and thus will not have a limit distribution.*

1. *Reducibility*

2. *Periodicity*

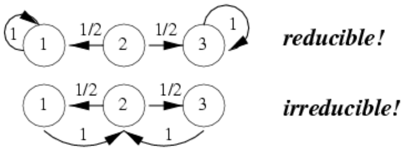
3. *Transience.*

A discrete time Markov chain on a discrete state space is ergodic, if it is irreducible, aperiodic and positive recurrent. We first review these concepts for countable (discrete) state spaces and then discuss the analogues for general state spaces.

Definition 5 Irreducibility
A Markov chain is *irreducible* if all states intercommunicate, that is for all $i, j \in \mathcal{S}$ there is an $n \geq 0$ such that

$$\mathbb{P}\Big(X_n = i \mid X_0 = j\Big) \quad > \quad 0.$$

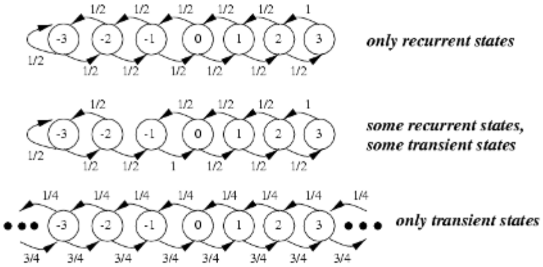
Task 5 *Are the following two Markov chains irreducible?*



Definition 6 Recurrence
A Markov chain X is recurrent, if

$$\mathbb{P}\Big(X \text{ visits } i \text{ eventually } \mid X_0 = i) \quad = \quad 1 \qquad \text{for all } i \in \mathcal{S}.$$

Task 6 *Which of the following Markov chains is recurrent?*



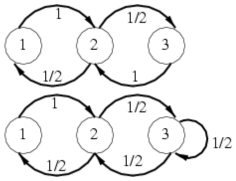
Definition 7 Positive Recurrence
A recurrent chain is *positive recurrent*, if $\mathbb{E}(T_{ii}) < \infty$ for all $i \in \mathcal{S}$, where T_{ii} is the time of the first return to state i. If the Markov chain is ergodic with stationary distribution π, then

$$\pi(i) \quad = \quad 1/\mathbb{E}(T_{ii}).$$

Definition 8 Aperiodicity
A Markov chain is aperiodic if there do not exist $d \geq 2$ and disjoint subsets $\mathcal{S}_1, \ldots, \mathcal{S}_d \subset \mathcal{S}$ such that

$$\begin{aligned} P(x, \mathcal{S}_{i+1}) &= 1 \quad \textit{for all } x \in \mathcal{S}_i, \quad i \in \{1, \ldots, d-1\} \\ P(x, \mathcal{S}_1) &= 1 \quad \textit{for all } x \in \mathcal{S}_d. \end{aligned}$$

Task 7 *Which of the following Markov chain is periodic and which aperiodic?*



all states have period 2.

all states are aperiodic.

Remark 1

1. Recurrence and aperiodicity are class properties. If a Markov chain is irreducible, then the whole state space is one communicating class. Thus an irreducible chain is recurrent if one of its states is recurrent. The same is true for positive recurrence and aperiodicity.

2. Irreducibility essentially ensures that the state space does not split into subsets such that the chain cannot move from one subset to the other.

3. (Positive) Recurrence ensures that the chain eventually visits every subset of the state space of positive mass (sufficiently often).

4. Perodicity causes the state space to split into subsets (cyclically moving subclasses) which are visited by the chain in sequential order.

Now let's look at a continuous state space $\mathcal{X}$. Because the probability of a continuous random variable taking a fixed value is zero, we need to revise our concept of irreducibility.

Definition 9 ϕ-irreducibility
A Markov chain is ϕ-irreducible if there exists a non-zero measure ϕ on $\mathcal{X}$ such that for all $A \subseteq \mathcal{X}$ with $\phi(A) > 0$ and for all $x \in \mathcal{X}$, there exists a positive integer $n = n(x)$ such that

$$P^n(x, A) \quad > \quad 0.$$

For example, if $\phi(A) = \delta_{x_0}(A)$ then this requires that the state x_0 can be reached from any other state with positive probability. Thus, irreducibility is a more stringent condition than ϕ-irreducibility. For continuous state spaces, $\phi(\cdot)$ might be the Lebesgue measure.

The concept of aperiodicity as defined earlier also applies to continuous Markov chains.

A Markov chain that is ϕ-irreducible and aperiodic has a limit distribution. To measure the distance between two probability measures we use the total variation distance:

Definition 10 *The total variation distance between two probability measures P_1 and P_2 is defined as*

$$||P_1(\cdot) - P_2(\cdot)|| \quad = \quad \sup_A |P_1(A) - P_2(A)|.$$

Theorem 2 Equilibrium distribution The distribution of an aperiodic, ϕ-irreducible Markov chain converges to a limit distribution π, that is

$$\lim_{n\to\infty} ||P^n(x,\cdot) - \pi(\cdot)|| \quad = \quad 0 \text{ for } \pi - \text{almost all } x \in \mathcal{X}.$$

We call the limit distribution π the *equilibrium* distribution or the *stationary* distribution.

The above limit theorem holds for almost all starting values. To make it hold for all starting values and thus for all initial distributions $q^{(0)}$ we need an additional property of the Markov chain, the so-called Harris recurrence.

Definition 11 Harris recurrence: *A Markov chain X is Harris recurrent if for all $B \subseteq \mathcal{X}$ with $\pi(B) > 0$ and all $x \in \mathcal{X}$ we have*

$$P(X_n \in B \textit{ for some } n > 0 | X_0 = x) \quad = \quad 1.$$

Theorem 3 The distribution of an aperiodic, Harris recurrent Markov chain converges to a limit distribution π, that is

$$\lim_{n\to\infty} ||P^n(x,\cdot) - \pi(\cdot)|| \quad = \quad 0 \quad \textit{for all } x \in \mathcal{X}.$$

Note that because

$$q^n(A) \quad = \quad \mathbb{P}(X_n \in A) \quad = \quad \int q^{(0)}(x) P^n(x,A) dx$$

it follows that $\lim_{n\to\infty} \mathbb{P}(X_n \in A) = \pi(A)$ for all $A \subseteq \mathcal{X}$ and all initial distributions $q^{(0)}$.

Standing Assumption: For the rest of these notes we assume that the Markov chain X is ergodic. In practice this needs to be check for each individual case!

Because Theorem 3 holds for any initial distribution $q^{(0)}$ if we run an ergodic Markov chain for a long time, then it settles down to a statistical equilibrium, regardless of its starting point. Furthermore, if we start a chain in equilibrium then it remains in equilibrium. This can be easily shown using induction as follows. For $n = 1$

$$q^{(1)}(x) \quad = \quad \int_{\mathcal{X}} q^{(0)}(y)\; p(y,x) dy \quad = \quad \int_{\mathcal{X}} \pi(y)\; p(y,x) dy \quad = \quad \pi(x)$$

Now assume $q^{(n)}(\cdot) = \pi(\cdot)$, then

$$q^{(n+1)}(x) \quad = \quad \int_{\mathcal{X}} q^{(n)}(y)\; p(y,x) dy \quad = \quad \int_{\mathcal{X}} \pi(y)\; p(y,x) dy \quad = \quad \pi(x)$$

Note that using the dominated convergence theorem we have

$$\lim_{n\to\infty} q^{(n+1)}(x) \quad = \quad \lim_{n\to\infty} \int_{\mathcal{S}} q^{(n)}(y)\; p(y,x)dy \quad = \quad \int_{\mathcal{S}} \lim_{n\to\infty} q^{(n)}(y)\; p(y,x)dy$$

It follows that

$$\pi(x) \quad = \quad \int_{\mathcal{S}} \pi(y)\; p(y,x)dy.$$

We call the last equation the *general balance* equation. It shows that the transition probabilities of X preserve the equilibrium. This means that once the chain is in equilibrium it remains in equilibrium. That is why we call a distribution which satisfies the general balance equation the *invariant* distribution of X. If X is ergodic, then we can use the general balance equations to show that π is the equilibrium distribution for the chain X. However, often it is difficult to find transition probabilities which solve the equations given by general balance. Much easier are the *detailed balance* equations.

Lemma 2 Detailed balance. Suppose π is a distribution on $\mathcal{S}$ which satisfies

$$\pi(x)\; p(x,y) \quad = \quad \pi(y)\; p(y,x)$$

for all $x, y \in \mathcal{S}$, where $p(x,y)$ is the transition density/pmf of an ergodic Markov chain X. Then π is the stationary distribution of X.

Proof:
The distribution π satisfies general balance because

$$\int_{\mathcal{S}} \pi(x)p(x,y)dx \quad = \quad \int_{\mathcal{S}} \pi(y)p(y,x)dx \quad = \quad \pi(y)\int_{\mathcal{S}} p(y,x)dx \quad = \quad \pi(y).$$

Note that detailed balance is not necessary for general balance! If detailed balance holds, then the chain is time-reversible. This means in equilibrium the chain statistically behaves the same whether it is run forwards or backwards in time. However, there are many ergodic Markov chains which are not time-reversible.

Example 8 Weather chain.
The weather chain has transition matrix:

$$\mathbf{P} \quad = \quad \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.25 & 0.25 & 0.5 \\ 0 & 0.4 & 0.6 \end{pmatrix}$$

Its general balance equations are:

$$(\pi(0), \pi(1), \pi(2)) \quad = \quad (\pi(0), \pi(1), \pi(2)) \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.25 & 0.25 & 0.5 \\ 0 & 0.4 & 0.6 \end{pmatrix}$$

$$\begin{aligned} \pi(0) &= 0.4\pi(0) + 0.25\pi(1) \\ \pi(1) &= 0.6\pi(0) + 0.25\pi(1) + 0.4\pi(2) \\ \pi(2) &= 0.5\pi(1) + 0.6\pi(2) \end{aligned}$$

$$\Rightarrow \quad \pi(0) \quad = \quad \frac{5}{12}\pi(1), \qquad \pi(2) \quad = \quad \frac{5}{4}\pi(1)$$

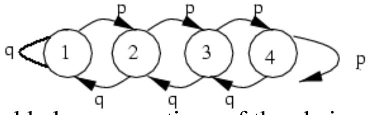
Thus

$$1 = \pi(0) + \pi(1) + \pi(2) = (\frac{5}{12} + 1 + \frac{5}{4})\pi(1) = \frac{8}{3}\pi(1)$$

and so $\pi(1) = \frac{3}{8}$.
The detailed balance equations of the weather chain are:

$$\begin{array}{rcl} \pi(0)p(0,1) & = & \pi(1)p(1,0) \qquad 0.6\pi(0) = 0.25\pi(1) \\ \pi(1)p(1,2) & = & \pi(2)p(2,1) \qquad 0.5\pi(1) = 0.4\pi(2) \\ \Rightarrow \quad \pi(0) & = & \frac{5}{12}\pi(1), \quad \pi(2) \quad = \quad \frac{5}{4}\pi(1) \end{array}$$

Example 9 Consider the Markov chain with the following state-flow diagram where $0 < p < 1$.



The general balance equations of the chain are:

$$\begin{array}{rcl} \pi(1) & = & q\pi(1) + q\pi(2) \\ \pi(2) & = & p\pi(1) + q\pi(3) \\ \pi(3) & = & p\pi(2) + q\pi(4) \\ \pi(4) & = & p\pi(3) + p\pi(4) \end{array}$$

The detailed balance equations are given by

$$\begin{array}{rcl} p\pi(1) & = & q\pi(2) \\ p\pi(2) & = & q\pi(3) \\ p\pi(3) & = & q\pi(4) \end{array}$$

It follows that

$$\begin{array}{rcccl} \pi(2) & = & \frac{p}{q}\pi(1) & & \\ \pi(3) & = & \frac{p}{q}\pi(2) & = & \left(\frac{p}{q}\right)^2\pi(1) \\ \pi(4) & = & \frac{p}{q}\pi(3) & = & \left(\frac{p}{q}\right)^3\pi(1) \end{array}$$

As $1 = \pi(1) + \pi(2) + \pi(3) + \pi(4)$ *we have that*

$$\pi(1) \quad = \quad \frac{1 - \frac{p}{q}}{1 - (\frac{p}{q})^4}.$$

Note that detailed balance equations are usually easier to solve than general balance equations.

The usefulness of MCMC is based on the following important theorem for ergodic Markov chains.

Theorem 4 Ergodic theorem: *Let h be some real function and X an ergodic Markov chain with stationary distribution π. Consider the ergodic average*

$$\overline{h}_N = \frac{1}{N}\sum_{n=1}^{N} h(X_n).$$

Now, suppose that Y has distribution π. If $\mathbb{E}_\pi(|h(Y)|) < \infty$ then, as $N \to \infty$ the ergodic average $\overline{h}_N$ converges to $\mathbb{E}_\pi(h(Y))$ with probability one.

We also have a Central limit theorem. It requires a certain condition on the speed of convergence known as geometric convergence. We use the same notation as in the previous theorem.

Theorem 5 Central Limit Theorem:
If X is geometrically ergodic and $\mathbb{E}_\pi\Big(h(Y)^{2+\epsilon}\Big) < \infty$ for some $\epsilon > 0$ then

$$\hat{h}_N \xrightarrow{\mathcal{D}} \mathcal{N}\Big(\mathbb{E}_\pi(h(X)),\ \frac{\tau^2}{N}\Big)$$

Here the convergence is convergence in distribution. The term τ^2 is related to the integrated autocorrelation time of X, which quantifies how accurated estimates derived from the MCMC sample will be.

In the following chapter we see how we can construct a Markov chain whose equilibrium distribution is the target distribution given by π.

5 Markov chain Monte Carlo

5.1 Introduction

Let's recap from the last section. The ergodic theorem is a law of large numbers and we can exploit it as follows. Suppose we would like to know the expectation of the random variable $h(Y)$ where Y has the distribution given by the density (pmf) π. However, we cannot compute $\mathbb{E}(h(Y)) = \int h(y)\ \pi(y)dy$. Fortunately, we can construct an ergodic Markov chain X whose stationary distribution has density π. Then we run X up to some large time N and estimate $\mathbb{E}(h(Y))$ by $\frac{1}{N}\sum_{n=1}^{N} h(X_n)$. The ergodic theorem tells us that for sufficiently large N, our estimate will be close to $\mathbb{E}(h(Y))$. This is the main idea behind MCMC.

So we start with a distribution π and then try to find an ergodic Markov chain whose stationary distribution is π. For any given distribution there are usually many Markov chains which are suitable. Thus there is a variety of ways in which to construct a Markov chain whose distribution converges to the target distribution.

It is actually not very difficult to find a Markov chain whose invariant distribution is the desired distribution. There is a set of methods, so-called "samplers", which we can use to define such a Markov chain. If the constructed chain is ergodic then we can proceed by simulating that chain and estimating the quantities of interest. In the following two sections we will learn about the two most common samplers.

5.2 Metropolis-Hastings Sampler

Let $\mathcal{S}$ be the state space of the target distribution. The transitions of a Metropolis-Hastings chain are produced as follows. First, we choose for each $x \in \mathcal{S}$ a density $q(x, \cdot)$ on $\mathcal{S}$ (or a pmf if $\mathcal{S}$ is discrete). Thus $q(x, \cdot), x \in \mathcal{S}$, specify the transition probabilities/densities of a Markov chain on the state space $\mathcal{S}$ given the current state is x. These transition probabilities/densities $q(x, \cdot)$ should be such that they are relatively easy to sample.

Now suppose the current state of our Markov chain is $X_n = x$. Then we sample a state z according to $q(x, \cdot)$. We propose this state z as the new state of the chain and accept it with probability

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

If the proposed state z is accepted then the Markov chain moves to z, that is $X_{n+1} = z$. Otherwise the chain remains in x, that is $X_{n+1} = x$. We summarize this procedure in the following definition.

Definition 12 Metropolis-Hastings Sampler: *Choose transition probabilities/densities $q(x, y), x, y \in \mathcal{S}$. These define the proposal distributions. Now suppose $X_n = x \in \mathcal{S}$. Proceed as follows:*

1. *Sample $Z = z$ from $q(x, z)$, $z \in \mathcal{S}$.*
2. *Accept $Z = z$ with probability*

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

If $Z = z$ is accepted set $X_{n+1} = z$. If $Z = z$ is not accepted set $X_{n+1} = x$.

Let us look at some examples. The first example is about a mixture distribution. Continuous mixture distributions with two components have densities of the form

$$f(x) \quad = \quad pf_1(x) + (1 - p)f_2(x)$$

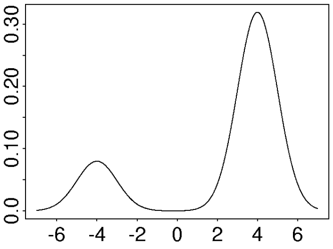
where $0 < p < 1$ and $f_i(x)$ is a density. We can sample mixtures by sampling x from $f_1(\cdot)$ with probability p and from $f_2(\cdot)$ with probability $1 - p$. In the following example we show how to sample from a mixture distribution using the Metropolis-Hastings sampler. The density in the example could be sampled directly but we discuss the example here as we will later use it to discuss some implementational issues.

Example 10 Bimodal Normal mixture distribution:

- *The target density is*

$$\pi(x) \;=\; p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2)+(1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp(-\frac{1}{2\sigma_2^2}(x-\mu_2)^2)$$

where $0 < p < 1$. The figure below shows the density above for $\sigma_1 = \sigma_2 = 1$, $\mu_1 = 4$, $\mu_2 = -4$ and $p = 0.8$.



- *The proposal density:*
We sample w from a standard Normal density and propose $z = x + w$ as our new state. Thus $z \sim \mathcal{N}(x, 1)$ and our proposal density is

$$q(x,z) \quad = \quad \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}(z-x)^2\Big).$$

- *The acceptance probability:*

$$\begin{aligned}\alpha(x,z) &= \min\Big\{1, \frac{\pi(z)q(z,x)}{\pi(x)q(x,z)}\Big\} \\ &= \min\Big\{1, \frac{\pi(z)\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(x-z)^2)}{\pi(x)\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(z-x)^2)}\Big\} \\ &= \min\Big\{1, \frac{\pi(z)}{\pi(x)}\Big\}\end{aligned}$$

- *The Metropolis-Hastings sampler proceeds as follows:*

  1. *Choose $X_0 = x_0 \in \mathbb{R}$.*
  2. *Suppose $X_n = x$. Sample $z \sim \mathcal{N}(0,1)$ and set $y = x + z$. Accept y with probability $\min\{1, \frac{\pi(y)}{\pi(x)}\}$. If accepted set $X_{n+1} = y$, else set $X_{n+1} = x$.*

Example 11 Points on a unit circle: (adapted from Ross, 2002)
Suppose $\underline{x} = \{x^{(1)}, \ldots, x^{(m)}\}$ are the positions of m points on the unit circle. Let $\pi(x^{(1)}, \ldots, x^{(m)})$ be the density that distributes the m points iid uniformly on the circle conditional on no points being within distance d of each other. (Distributions of this type often occur in chemical settings where the points are centres of spherical molecules of diameter d). Let A be the event that the minimum distance between m

points iid uniformly distributed on the unit circle is greater than d and set $p = \mathbb{P}(A)$. Let $\mathcal{S}$ be the space of any configuration of m points in $(0, 2\pi)$ such that the mimimum distance between points is greater than d. Then our target distribution is

$$\pi(\underline{x}) \quad = \quad \frac{1}{2\pi p}\mathbf{1}_{[\underline{\mathbf{x}}\in\mathcal{S}]}.$$

In one dimension we are just about able to compute p, but already in 2 dimensions this becomes infeasible. So similar to the previous example we have a simple form for the target distribution but we do not have a closed form expression for its normalizing constant.

Let's think about an easy way of moving from one $\underline{x} \in \mathcal{S}$ to another $\underline{x}' \in \mathcal{S}$. One way is to choose a point $x \in \underline{x}$ at random and delete it and then sample a new location z uniformly in $(0, 2\pi)$ and setting $\underline{x}' = \underline{x} \cup \{z\} \backslash \{x\}$. (This may produce a configuration $\underline{x}'$ that does not lie in $\mathcal{S}$ but, as we will see later on, this does not really matter.) This approach is described by the transition density

$$q(\underline{x}, \underline{x}') \quad = \quad \frac{1}{2\pi m}\mathbf{1}_{[z\in(0,1)]} \qquad \text{where } \underline{x}' = \underline{x} \cup \{z\}\backslash x.$$

Then for $\underline{x} \in \mathcal{S}$ and $\underline{x}' = \underline{x} \cup \{z\}\backslash\{x\}$ we have

$$\begin{aligned} \alpha(\underline{x}, \underline{x}') &= \min\left\{1, \frac{\pi(\underline{x}')q(\underline{x}', \underline{x})}{\pi(\underline{x})q(\underline{x}, \underline{x}')}\right\} \\ &= \min\left\{1, \frac{\mathbf{1}_{[\underline{x}'\in\mathcal{S}]}\mathbf{1}_{[x\in(0,2\pi)]}}{\mathbf{1}_{[\underline{x}\in\mathcal{S}]}\mathbf{1}_{[z\in(0,2\pi)]}}\right\} \\ &= \begin{cases} 1 & \text{if } \underline{x}' \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Thus as long as we start in $\mathcal{S}$ any state that we move to will also lie in $\mathcal{S}$! In summary, the Metropolis algorithm works as follows: Choose $X_0 \in \mathcal{S}$, for example by placing the points successively a distance $d + \epsilon$ apart from each other (where ϵ is sufficiently small). Now suppose $X_n = \underline{x} \in \mathcal{S}$. Proceed as follows:

1. *Choose $i \in \{1, \dots, m\}$ at random and sample z uniformly on $(0, 2\pi)$. Set $\underline{z} = \underline{x} \cup \{z\}\backslash\{x^{(i)}\}$.*

2. *If $\underline{z} \in \mathcal{S}$, then accept $\underline{z}$ and set $X_{n+1} = \underline{z}$. If $\underline{z} \notin \mathcal{S}$, reject $\underline{z}$ and set $X_{n+1} = \underline{x}$.*

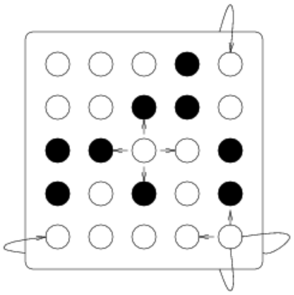
Example 12 *The Ising model describes a lattice which at each site has a small dipole or spin which is directed upwards or downwards. Thus each site j may take a value $x^{(j)} \in \{-1, 1\}$ representing a downward respectively an upward spin. The figure below represents a configuration of an Ising model, where sites are displayed as circles. If the site carries an upwards spin then it is filled in black. If, on the other hand, the site has a downwards spin, it is filled in white. The spin at each site is influenced by the spins of its neighbour sites. The figure below shows a possible choice of neighbourhood. This particular choice is called a 4-neighbourhood with periodic boundary conditions. The ferro-magnetic Ising model gives a high probability to spin configurations in which many neighbouring sites have the same spin. Physicist tend to speak* *of energies rather than probabilities. A state that has small energy is easy to maintain and thus has a high probability. The energy function of a simple ferro-magnetic Ising model is given by*

$$H(x) \quad = \quad -\beta \sum_{(i,j)\in\mathcal{E}} x^{(i)} x^{(j)}.$$

Here x is a spin configuration on a lattice. Assuming the lattice has m sites, then $x = (x^{(1)}, \ldots, x^{(m)})$ is an m-dimensional vector of 1's and -1's. Let "$i \sim j$" mean that site i is a neighbour of site j then we are summing over $\mathcal{E} = \{(i,j) : i \sim j\}$ the set of all site pairs that are neighbours of each other. The constant β is positive for the ferro-magnetic Ising model. (If β is negative, then we call it the anti-ferromagnetic Ising model.) Based on the energy function H we can now define the probability mass function:

$$\pi(x) \quad = \quad \frac{1}{Z} \exp(-H(x)).$$

We call Z the partition function, but it is nothing else but the inverse of the normalizing constant of π.



A Metropolis-Hastings type algorithm can be used to sample (approximately) from an Ising model. To determine a new state, we run through the sites in a sequential order and propose to flip its spin. As an exercise, determine the proposal probabilities and work out the acceptance probability! Here is a pseudo-code description of the algorithm:

```
Initialize x
for $t=1$ to $N$
    for $i=1$ to $m$
        $d = \beta \sum_{j:i\sim j} x^{(i)} x^{(j)}$
        $U \sim$ Uniform(0,1)
        if $\log(U) < \min\{0, -2d\}$ then
            $x^{(i)} = -x^{(i)}$
```

Note that the resulting chain $(X_n)_{n\in\mathcal{N}}$ is no longer time-reversible (as we can spot the order in which sites are updated from a realisation of the chain)! However, each step satisfies detailed balance and what is more, the chain $(\tilde{X}_k)_{k\in\mathcal{N}} = (X_{|\mathcal{S}|k})_{k\in\mathcal{N}}$ is time-reversible.

Let's look at some of the more theoretical properties of the Metropolis-Hastings algorithm. First notice, that there is a lot of freedom in how you choose the proposal mechanism $q(x,y)$. A priori, the only necessary condition is that the support of the target density π is a subset of the support of the appropriate proposal densities. More precisely, we need

$$\mathcal{S} = \text{support}(\pi) \subseteq \bigcup_{x\in\mathcal{S}} \text{support}(q(x,\cdot)).$$

You may have noticed in the preceding examples that the acceptance probability $\alpha(x,y)$ is based on ratios of $\pi(\cdot)$, thus we do not need to know the normalizing constant of $\pi(\cdot)$ to be able to compute this probability. You may also have noticed that the acceptance probability contains terms that are similar to the terms in the detailed balance equations. This is no coincidence, the acceptance probability is chosen such that detailed balance holds! Thus if the resulting chain is ergodic then it has $\pi(\cdot)$ as stationary distribution. Let us have a look at the detailed balance equations of the Metropolis-Hastings chain. For this, we first need to determine the transition kernel of the MH-chain.

Lemma 3 *The transition kernel $p(x,y)$ for the Metropolis-Hastings sampler is given by*

$$p(x,y) \quad = \quad q(x,y)\alpha(x,y) + \mathbf{1}_{[x=y]}r(x).$$

where

$$r(x) \quad = \quad \begin{cases} \sum_{y\in\mathcal{S}} q(x,y)\Big(1-\alpha(x,y)\Big) & \textit{if } \mathcal{S} \textit{ is discrete} \\ \int_{\mathcal{S}} q(x,y)\Big(1-\alpha(x,y)\Big)dy & \textit{if } \mathcal{S} \textit{ is continuous} \end{cases}$$

(Note that this transition kernel is not continuous with respect to the Lebesgue measure.)

Proof: Suppose $\mathcal{S}$ is discrete. Recall that the chain moves to a new state y if this state was proposed and accepted. This happens with probability $q(x,y)\alpha(x,y)$. This is the probability of going from x to y when $y \neq x$. Now consider the probability of going from x to x. This can happen in two ways. Firstly we may propose x as the new state and then accept it, which happens with probability $q(x,x)\alpha(x,x)$. Secondly, we may propose some state y and reject it, in which case the chain remains in x. This occurs with probability

$$r(x) \quad = \quad \sum_{y\in\mathcal{S}} q(x,y)\Big(1-\alpha(x,y)\Big).$$

Thus, in summary, the transition probabilities of the Metropolis-Hastings chain are given by

$$p(x,y) \quad = \quad q(x,y)\alpha(x,y) + \mathbf{1}_{[x=y]}r(x).$$

The proof for continuous $\mathcal{S}$ follows analogously.

We are now in a position to check the detailed balance equations.

Lemma 4 *The Metropolis-Hastings chain satisfies detailed balance with respect to π.*

Proof: For $x \neq y$ we have

$$\begin{aligned}
\pi(x)p(x,y) &= \pi(x)q(x,y)\alpha(x,y) \\
&= \min\Big\{\pi(x)q(x,y), \pi(y)q(y,x)\Big\} \\
&= \pi(y)q(y,x)\min\Big\{\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}, 1\Big\} = \pi(y)p(y,x).
\end{aligned}$$

Detailed balance holds trivially for $x = y$ and so our claim follows.

5.3 Proposal distributions

Metropolis-Hastings Samplers are often classified according to their proposal distributions.

1. **Gibbs Sampler**
   The Gibbs Sampler is a popular choice that uses full conditional distributions as proposal distributions. Let $x_t = (x_t^{(1)}, \ldots, x_t^{(d)})$ and

   $$x_t^{(-i)} = (x^{(1)}, \ldots, x^{(i-1)}, x^{(i+1)}, \ldots, x^{(d)}).$$

   We choose a component $i \in \{1, \ldots, d\}$ and propose as a new state

   $$z = (x^{(1)}, \ldots, x^{(i-1)}, y, x^{(i+1)}, \ldots, x^{(d)})$$

   where y is sampled from the full conditional density

   $$\pi(y|x_t^{(-i)}) \quad = \quad \frac{\pi(z)}{\int \pi(x_t^{(1)}, \ldots, x_t^{(i-1)}, w, x_t^{(i+1)}, \ldots, x_t^{(d)})dw}.$$

   One can show that for this choice of proposal distribution the acceptance probability is equal to one. If the full conditional distributions are standard and thus easily sampled then the Gibbs sampler is a very popular choice. We will devote a whole section on the Gibbs sampler, but for now we look at a simple example.

   Example **13 Bivariate Normal distribution**
   This is just a toy example as we can sample bivariate normal distributions directly. But it illustrates well how the Gibbs Sampler works. We would like to sample X and Y which have joint density

   $$\pi(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\Big(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\Big)$$

This density specifies a bivariate Normal distribution with mean (0,0) and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

As an exercise show that $(X|Y = y)$ has distribution $\mathcal{N}(\rho y, 1 - \rho^2)$ and that $(Y|X = x)$ has distribution $\mathcal{N}(\rho x, 1 - \rho^2)$. In computer practical 5 you will implement a sequential scan Gibbs Sampler that alternatingly updates the x-coordinate and then the y-coordinate. Suppose $X_n = (x_n, y_n)$ then we proceed as follows. We first sample $X = x$ from the conditional distribution of $(X|Y = y_n)$. Then we sample $Y = y$ from the conditional distribution of $(Y|X = x)$ and set $X_{n+1} = (x, y)$.

2. **Independence Sampler**

As the name suggests the independence sampler proposes states which are independent of the current state of the chain, that is $q(x, y) = f(y)$ for all $x \in \mathcal{S}$, where f is a pmf or density. The acceptance probability for the independence sampler reduces to

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

Note that this is just the ratio of importance weights $w(y)/w(x)$ for target density π and instrumental density f.

Example 14 *Consider target density*

$$\pi(x) = \frac{1}{\pi(1 + x^2)}, \qquad x \in \mathbb{R}.$$

If we use standard Normal proposals with zero mean and standard deviation 4, then the proposal density is

$$q(x, y) \propto \exp(-y^2/32),$$

and so the acceptance probability is given by

$$\alpha(x, y) = \min\left\{1, \frac{\exp(-x^2/32)(1 + x^2)}{\exp(-y^2/32)(1 + y^2)}\right\}.$$

While the independence sampler may not work so well in practice its theoretical properties are well understood (using tools from renewal theory). For example we can show that an independence sampler is ergodic as long as the support of π is a subset of the support of f.

The independence sampler is also very similar to rejection sampling. Let's compare the acceptance probability for rejection sampling with the expected acceptance probability of the independence sampler in stationarity. For rejection sampling to apply we assume that $\pi(x) \leq Mf(x)$. Then if Y has distribution f and

X has distribution π we have

$$\begin{aligned}
\mathbb{E}\Big(\min\{1, \frac{\pi(Y)f(X)}{\pi(X)f(Y)}\}\Big) &= \int\int \mathbf{1}_{[\pi(y)f(x)\geq\pi(x)f(y)]}\pi(x)f(y)dxdy \\
&+ \int\int \frac{\pi(y)f(x)}{\pi(x)f(y)}\mathbf{1}_{[\pi(y)f(x)<\pi(x)f(y)]}\pi(x)f(y)dxdy \\
&= 2\int\int \mathbf{1}_{[\pi(y)/f(y)\geq\pi(x)/f(x)]}\pi(x)f(y)dxdy \\
&\geq 2\int\int \mathbf{1}_{[\pi(y)/f(y)\geq\pi(x)/f(x)]}\pi(x)\frac{\pi(y)}{M}dxdy \\
&= \frac{2}{M}\,\mathbb{P}(\pi(X_1)/f(X_1)\geq\pi(X_2)/f(X_2)) \;=\; \frac{1}{M}
\end{aligned}$$

where X_1 and X_2 are iid with distribution π. Thus, in stationarity, the acceptance probability of an independence sampler is larger than the acceptance probability of the rejection sampling algorithm. This of course comes at the expense of producing a dependent sample with only asymptotically the correct distribution. Similar to rejection sampling it makes sense to choose an independence sampler whose proposal distribution f is as close as possible to the target π. (To see this note that if $f = \pi$ the chain immediately reaches stationarity.) In practice the proposal distribution f_θ usually depends on some parameter θ and we tune the parameter empirically to get a good average acceptance rate. We may use trial runs to estimate the expected acceptance rate.

If $\pi(x) \leq Mf(x)$ then we can even compute the rate of convergence of the transition kernel to the stationary distribution as follows. For $y \neq x$:

$$\begin{aligned}
p(x,y) &= f(y)\min\Big\{\frac{\pi(y)\ f(x)}{\pi(x)\ f(y)}, 1\Big\} \\
&= \min\Big\{\frac{\pi(y)\ f(x)}{\pi(x)}, f(y)\Big\} \quad\geq\quad \frac{\pi(y)}{M} \qquad (1)
\end{aligned}$$

$$\begin{aligned}
||P(x,\cdot)-\pi|| &= \sup_A |\int_A p(x,y)-\pi(y)dy| \\
&= \int_{\{y:\pi(y)>p(x,y)\}} \pi(y)-p(x,y)dy \\
&\leq (1-\frac{1}{M})\int_{\{y:\pi(y)>p(x,y)\}} \pi(y)dy \quad\leq\quad (1-\frac{1}{M})
\end{aligned}$$

where the first inequality follows from equation (1). Now, analogously,

$$\int_A p^2(x,y)-\pi(y)dy = \int_A \Big(\int_A p(u,y)-\pi(y)dy\Big)(p(x,u)-\pi(u))du \leq (1-\frac{1}{M})^2.$$

Using induction we can now show that

$$||P^n(x,\cdot)-\pi|| \quad\leq\quad (1-\frac{1}{M})^n.$$

The above means that the independence sampler is uniformly ergodic if $\pi(x) \leq Mf(x)$, see the definition below.

Definition 13 *An ergodic Markov chain with invariant distribution π is* geometrically ergodic *if there exist a non-negative function M such that $\mathbb{E}_\pi(M(X)) < \infty$ and a positive constant $r < 1$ such that*

$$||P^n(x, \cdot) - \pi(\cdot)|| \quad \leq \quad M(x)r^n$$

for all x and all n. If the function M is bounded above, that is there exists $K > 0$ such that $M(x) < K$ for all x then the chain is called uniformly ergodic.

3. **Random walk Metropolis-Hastings sampler**
   Here we choose $q(x, y) = f(y-x)$ for some probability mass function or density f. The random walk Metropolis-Hastings sampler derives its name from the fact that the proposals are made according to a random walk, that is

   $$y \quad = \quad x + z$$

   where z is drawn from f. The acceptance probability for this proposal distribution is

   $$\alpha(x, y) \quad = \quad \min\{1, \frac{\pi(y)f(x-y)}{\pi(x)f(y-x)}\}.$$

   Note that if f is symmetric about 0, then this is a Metropolis Sampler. The example on the bi-modal mixture distribution was an example for a Metropolis Sampler and also of the random walk MH sampler.

   Common choices for f are multivariate Normal densities, t-densities or uniform densities.

4. **The Metropolis sampler**
   Metropolis et al. originally proposed to use symmetric proposal pmf's or densities, that is $q(x, y) = q(y, x)$. The acceptance probability then simplifies to

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

5.4 Mixtures and Cycles

The description of Markov chain Monte Carlo may suggest that we have to do either Gibbs Sampling or a random walk Metropolis-Hastings sampler, etc. We can, in fact, combine different types of updates into one MCMC algorithm, a so-called hybrid MCMC algorithm. We can do this by either choosing the type of update randomly or systematically. Let $p_1(x, \cdot), \ldots, p_n(x, \cdot)$ be the transition densities/pmfs of the different update types that we are considering. A random choice of update corresponds to a mixture of kernels, that is our Markov chain is updated according to

$$p(x, y) \quad = \quad \sum_{i=1}^{n} \alpha_i p_i(x, y)$$

where α_i is the probability with which we choose an update of type i. Alternatively we may systematically cycle through our selection of update mechanisms leading to a chain with transition kernel defined by

$$p(x,y) \quad = \quad \int \dots \int \prod_{i=1}^{n} p_i(z_{i-1}, z_i) dz_1 \dots dz_{n-1},$$

where $z_0 = x$ and $z_n = y$.

Lemma 5 *Let $p_1(x,y)$ and $p_2(x,y)$ be two transition densities/pmfs with the same stationary distribution π. If $p_1(x,y)$ is uniformly ergodic then the mixture kernel*

$$p(x,y) = \alpha p_1(x,y) + (1-\alpha) p_2(x,y), \qquad 0 < \alpha < 1$$

is also uniformly ergodic.

The above lemma is the reason why we often include an independence sampler as an update kernel as under the appropriate conditions it ensures uniform ergodicity.

6 Implementational issues in MCMC

6.1 Assessing convergence

How to assess convergence has been a fiercely discussed topic in the literature. Some of the disagreement is due to the fact that there are different types of convergence. Firstly, there is the issue of whether the distribution of the chain is close to the stationary distribution. Secondly, there is the issue of how well the chain explores the state space. Generally, chains that do not explore the state space well (we say, mix slowly) tend to converge slower. But note that even if the chain is in stationarity then a slow mixing chain will lead to very inaccurate estimates, a third issue that is discussed using the general keyword of convergence. All three issues are related to each other but are not exactly the same and this has lead to a lot of confusion. For example, a chain with a bi-modal stationary distribution may not have reached equilibrium yet as it only ever has visited the region around one of the modes. Now suppose we are interested in estimating a function that is equal for both modes. Then it can happen that the corresponding ergodic average converges much earlier than the chain itself.

6.2 Initialisation bias and Burn-In

The Markov chain of interest is usually started in a state that does not have the stationary distribution (otherwise we would not be doing MCMC)! In practical 3 you explored the effect the initial state can have on the states visited by the Markov chain. To reduce the possibility of a bias, the so-called initialisation bias caused by the effect of the starting value, an initial M steps of the chain are discarded and estimation is based on the states visited after time M, ie. we use ergodic average

$$\overline{h}_N \quad = \quad \frac{1}{N-M} \sum_{n=M+1}^{N} h(X_n).$$

The initial phase up to time M is called the transient phase or burn-in period. How do we decide on the length of the burn-in period? A first step would be to examine the output of the chain by eye. This is a very crude method but is very quick and cheap. However, this should be followed up by more sophisticated methods.

To assess convergence to stationarity there are essentially three potential approaches.

1. **Convergence rate computations:**
   When we looked at the speed of convergence for the independence sampler we essentially performed a convergence rate calculation. The advantage of this approach is, of course, that it is exact. In general, these calculations can be difficult and will only apply to specific cases. While they are exact methods, they can be overly pessimistic as they take into account any worst-case scenario even if it has a very small probability of occurring. At times they can be so pessimistic that the suggested burn-in period simply becomes impractical.

2. **Perfect simulation:**
   This is a method that augments conventional MCMC methods in such a way that the chain is automatically run until it has reached convergence. Again, it only works in specific cases and can be slow. But it is again an exact method (as far as any simulation method can be exact). Researchers at the Warwick Statistics department were actively involved in extending the first perfect simulation algorithm introduced by Propp and Wilson.

3. **Convergence diagnostics:**
   This is the most widely used method. Convergence diagnostics examine the output of the chain and try to detect a feature that may suggest divergence from stationarity. They are usually relatively easy to implement. However, while they may indicate that convergence has not been reached they are not able to guarantee convergence. We will examine the most commonly used diagnostics in the next section. A collection of convergence diagnostics is implemented in CODA, which is freeware that runs on certain versions of SPlus, and BOA which runs on `R`.

6.3 Convergence diagnostics

In the following we introduce a small collection of common convergence diagnostics. There are many more available, if you are interested then check the MCMC literature.

6.3.1 Geweke's method

Geweke's method is based on spectral analysis methods from time series analysis. Geweke considers two subsequences of a run of a Markov chain, one consisting of the states immediately after burn-in and the other consisting of states at the very end of the run. Suppose our run is of length n and we want to examine whether a burn-in of M time steps is sufficient. Consider the ergodic averages

$$\bar{h}_A = \frac{1}{n_A}\sum_{t=M+1}^{M+n_A} h(X_t) \quad \text{and} \quad \bar{h}_B = \frac{1}{n_B}\sum_{t=n-n_B}^{n} h(X_t)$$

where $n_A + M < n - n_B$. If they chain has converged then the two ergodic averages should be similar. How similar they should be can be quantified using the appropriate Central Limit Theorem. For this, we need estimates of the variance of the ergodic averages (under stationarity) and we can use the spectral density to derive these. The spectral density for the time series $(h(X_t))_{t\in\mathbb{Z}}$ is defined as

$$S(\omega) \quad = \quad \frac{1}{2\pi}\sum_{t=-\infty}^{\infty}\mathrm{Cov}(h(X_0), h(X_t))\exp(it\omega)$$

where i is the square root of -1. We can estimate the spectral density using kernel estimation methods. We now have the following asymptotic result: If the ratios n_A/n, n_B/n are fixed with $(n_A + n_B)/(n - M) < 1$ then under stationarity we have

$$Z_n = \frac{\bar{h}_A - \bar{h}_B}{\sqrt{\frac{1}{n_A}\hat{S}_A(0) + \frac{1}{n_B}\hat{S}_B(0)}} \to \mathcal{N}(0, 1)$$

as $n \to \infty$ and convergence is convergence in distribution. Here $\hat{S}_A(0)$ and $\hat{S}_B(0)$ are spectral estimates of the appropriate variances. We can use the above asymptotics to approximately test whether the means of the two sequences are equal subject to variation. Note that this is testing for a sufficient, not a necessary condition for convergence. By default BOA implements this statistic by choosing n_A to be $n/10$ and $n_B = n/2$ which was originally suggested by Geweke.

6.3.2 Gelman and Rubin's method

This method requires several chains started in initial states that are overdispersed compared to the stationary distribution. The idea is that if there is initialisation bias then the chains will be close to different modes while under the stationary regime the chains should behave in a similar manner. If the chain tends to get stuck in a mode then the path appears to be experiencing stable fluctuations but the chain has not converged yet. In this case we speak of *"meta-stability"* which is often caused by multi-modality of the stationary distribution. Gelman and Rubin apply concepts from ANOVA techniques (and thus rely on Normal asymptotics). We assume each of a total of m chains is run for $2n$ iterations where the first n iterations are classified as burn-in. We first compute the variance of the means of the chains:

$$B \quad = \quad \frac{1}{m-1}\sum_{j=1}^{m}(\bar{h}_{.j} - \bar{h}_{..})^2$$

Here we run m chains and X_{ij} is the state of chain j at time i. $\bar{h}_{.j}$ is the ergodic average of the jth chain based on the last n iterations and $\bar{h}_{..}$ is the average of the ergodic averages of the m chains. This is interpreted as the between chain variance. As $n \to \infty$ the between chain variance will tend to zero.

Then we compute the within chain variance W, which is the mean of the variance of each chain, that is

$$W \quad = \quad \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n-1}\sum_{i=n+1}^{2n}(h(X_{ij}) - \bar{h}_{.j})^2.$$

Note that this quantity initially will tend to be an underestimate of the posterior variance. We calculate the weighted average of both variance estimates

$$V = \frac{n-1}{n}W + B,$$

which is an estimate of the posterior variance. We then monitor $R = \sqrt{\frac{V}{W}}$ which is called the potential scale reduction factor (PSRF). R tends to be bigger than one and will converge towards one as stationarity is reached. Using Normal asymptotics, one can show that R has an F-distribution and the hypothesis of $R = 1$ can be tested. As a rule of thumb if the 0.975-quantile is less than 1.2, no lack of convergence is being detected.

Brooks and Gelman suggested a correction to the above scale reduction factor to take account of the sampling variability in the variance estimates. They also extended the above approach using techniques of multivariate ANOVA in order to make it applicable to multivariate chains.

BOA computes the uncorrected and the corrected, univariate PSRF as well as the 0.975-quantile. Moreover, it computes the multivariate PSRF.

6.3.3 The method by Raftery and Lewis

Set $p = \mathbb{P}_\pi(h(X) \leq u)$ for some u, then the method of Raftery and Lewis estimates the number of iterations needed to produce an $\alpha\%$-confidence interval of length $2r$ for p. In detail, it computes the following quantities:

$$\begin{aligned} M &= \text{the burn-in length} \\ N &= \text{the total number of iterations} \\ k &= \text{the thinning factor, ie we only use every } k\text{th value for estimation} \end{aligned}$$

The method starts with the binary process $Z_t = \mathbf{1}_{[h(X_t) \leq u]}$ which in general is not Markov. However, if we consider $Z_t^{(k)} = Z_{tk}$ then if is reasonable to assume that this is approximately Markov for large k. First the method determines k by comparing the fit of a first-order Markov chain, ie. $Z_t^{(k)}$ is conditionally independent of the past given $Z_{t-1}^{(k)}$, with the fit of a second-order Markov chain, ie. $Z_t^{(k)}$ is conditionally independent of the past given both $Z_{t-1}^{(k)}$ and $Z_{t-2}^{(k)}$. We then choose the smallest k for which the first-order model is preferred.
Now, let the transition matrix of $Z_t^{(k)}$ be given by the matrix

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

which has stationary distribution $\pi = (\alpha+\beta)^{-1}(\beta, \alpha)$. Then for $\lambda = 1 - \alpha - \beta$ we have

$$Q^n = \begin{pmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{pmatrix} + \frac{\lambda^n}{\alpha+\beta}\begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}$$

Let $e_0 = (1, 0)$, $e_1 = (0, 1)$ and consider

$$\mathbb{P}(Z_n^{(k)} = i | Z_0^{(k)} = j) \quad = \quad e_j Q^n e_i^T$$

which we would like to estimate within ϵ of $\pi(i)$. This can be done if

$$|\lambda^n| \quad \leq \quad \frac{(\alpha + \beta)\epsilon}{\max(\alpha, \beta)}$$

which holds for

$$n \geq n_0 = \log\Big(\frac{(\alpha + \beta)\epsilon}{\max(\alpha, \beta)}\Big) / \log(|\lambda|)$$

and so a suitable burn-in of $M = n_0 k$ is suggested.

To determine the total runtime N one can show that asymptotically

$$\bar{Z}_n^{(k)} = \frac{1}{n}\sum_{t=1}^{n} Z_t^{(k)} \quad \sim \quad \mathcal{N}\Big(p, \frac{(2 - \alpha - \beta)\alpha\beta}{n(\alpha + \beta)^3}\Big)$$

and so to produce a $\gamma\%$-confidence interval of length at most $2r$ we need to choose

$$n \geq n_1 = \frac{(2 - \alpha - \beta)\alpha\beta}{n(\alpha + \beta)^3} \frac{1}{r^2} \Big(\Phi^{-1}\Big(0.5(\gamma + 1)\Big)\Big)^2,$$

where Φ is the cdf of a standard Normal distribution, and so the total run time suggested is $N = n_1 k$.

This convergence diagnostic is implemented in BOA. BOA also outputs a lower bound on the total run time which is the number of sample points needed to estimate the specified quantile to the desired accuracy using an iid sample. The dependence factor measures the multiplicative increase in the number of iterations needed compared to an iid sample. This will usually be greater than one due to the autocorrelation of the chain, see the next section on the Monte Carlo error.

6.4 Monte Carlo error

Once we have collected samples from our run of the Markov chain, we use ergodic averages to estimate the quantities of interest. But how accurate are these estimates and how large should we choose the sample size?

Definition 14 (Autocovariance/Autocorrelation): *In stationarity, the autocovariance of lag k for the time series $(h(X_k))_{k \geq 0}$ is defined as*

$$\begin{aligned} \gamma_k &= \mathrm{Cov}\Big(h(X_n), h(X_{n+k})\Big) \qquad n, k \geq 0, \\ \gamma_0 &= \mathrm{Cov}\Big(h(X_n), h(X_n)\Big) \quad = \quad \mathrm{Var}\Big(h(X_n)\Big) \end{aligned}$$

The autocorrelation of lag k is defined as

$$\rho_k \quad = \quad \frac{\gamma_k}{\gamma_0}, \qquad k \geq 0.$$

Note that if the chain X is in stationarity and h is the identity function then $\gamma_0 = \sigma^2$ where σ^2 is the variance of the stationary distribution π. But what we are really interested in is the variance of ergodic averages. Define

$$\frac{\tau_n^2}{n} \quad = \quad \mathrm{Var}(\bar{h}_n) \quad = \quad \mathrm{Var}\Big(\frac{1}{n}\sum_{k=1}^{n} h(X_k)\Big).$$

Note that to keep notation simple we omit the dependence of τ_n^2 on h.

Lemma 6 *In stationarity*

$$\tau_n^2 = \sigma^2\Big[1 + 2\sum_{k=1}^{n-1}\frac{n-k}{n}\rho_k\Big].$$

Proof:

$$\begin{aligned}
\frac{\tau_n^2}{n} &= \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} h(X_i)\Big) \\
&= \frac{1}{n^2}\Big[\sum_{i=1}^{n}\mathrm{Var}\Big(h(X_i)\Big) + 2\sum_{i=1}^{n-1}\sum_{j>i}\mathrm{Cov}\Big(h(X_i), h(X_j)\Big)\Big] \\
&= \frac{1}{n^2}\Big[n\sigma^2 + 2\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\mathrm{Cov}\Big(h(X_i), h(X_{i+k})\Big)\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n\sigma^2}\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\gamma_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n}\sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\rho_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + \frac{2}{n}\sum_{k=1}^{n-1}\sum_{i=1}^{n-k}\rho_k\Big] \\
&= \frac{\sigma^2}{n}\Big[1 + 2\sum_{k=1}^{n-1}\frac{(n-k)}{n}\rho_k\Big]
\end{aligned}$$

One can show that as $n \to \infty$

$$\tau_n^2 \quad \longrightarrow \quad \tau^2 = \sigma^2\Big[1 + 2\sum_{k=1}^{\infty}\rho_k\Big].$$

The quantity τ^2/σ^2 is called the integrated autocorrelation time and we have already encountered τ^2 in the statement of the Central Limit Theorem for Markov chains.

From the above it follows that the accuracy of MCMC estimates depends on the autocorrelation. Suppose $Y_1, \dots, Y_n$ are independent and identically distributed according to π. Then the mean of the sample $\frac{1}{n}\sum_{i=1}^n Y_i$ has variance

$$\mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^n Y_i\Big) \quad = \quad \frac{\sigma^2}{n}.$$

Now compare this with the variance of the ergodic average $\frac{1}{n}\sum_{i=1}^n X_i$ which is $\frac{\tau_n^2}{n}$. From the above lemma it follows that if the autocorrelations of the chain X are positive then the ergodic average will be less accurate than an estimate based on an independent sample. Informally, this may be explained by the fact that positively correlated variables carry redundant information and so are less informative than independent variables. On the other hand, if we can introduce negative autocorrelations into our Markov chain, then this will make our estimation procedure more accurate.

The estimation of the integrated autocorrelation time is difficult and there are various approaches one might to consider.

1. Naive covariance estimator:

$$\hat{\gamma}_k \quad = \quad \frac{1}{n}\sum_{i=1}^{n-k}\Big(h(X_i) - \bar{h}_n\Big)\Big(h(X_{i+k}) - \bar{h}_n\Big)$$

   However the resulting estimator

$$\hat{\tau}^2 \quad = \hat{\gamma}_0 + 2\sum_{k=1}^{n-1}\hat{\gamma}_k$$

   is not consistent.

2. Blocking or batch estimation:
   Divide the run of your chain into b blocks of length k. The idea is that averages of different blocks are asymptotically independent and we can use them in a standard variance estimator. Note that $\mathrm{Var}(\bar{h}_n) \approx \frac{1}{b}\mathrm{Var}(\bar{h}_k^{(j)})$ where $\bar{h}_k^{(j)}$ is the ergodic average in the jth block. Thus for suitably large b and k we have

$$\mathrm{Var}(\bar{h}_n) \quad \approx \quad \frac{1}{b(b-1)}\sum_{j=1}^{b}\Big(\bar{h}_k^{(j)} - \bar{h}_n\Big)^2.$$

3. Window estimates:
   Use $\hat{\tau}^2 = \hat{\gamma}_0 + 2\sum_{k=1}^{L}\hat{\gamma}_k$ for suitably chosen L. There are different ways of choosing L and here are some of the most common suggestions:

   - truncated periodogram: choose L to be the smallest integer such that $L > 3\sum_{k=1}^{L}\hat{\rho}_k$.
   - initial series estimator: choose L to be the largest integer such that for $k \le L$ the series $\hat{\Gamma}_k = \hat{\gamma}_{2k} + \hat{\gamma}_{2k+1}$ remains positive, non-decreasing and convex.

6.5 Scaling

Consider a Metropolis-Hastings algorithm and suppose we have decided on the parametric proposal distributions that we are going to use. We now have to decide which values to choose for the parameter(s) θ of our proposals. For example, we have decided to do a random walk Metropolis algorithm with Normal proposals. We then need to choose a variance for the Normal proposal distribution. This is commonly known under the keyword of scaling.

First, let us consider the independence sampler and suppose the proposal distribution f is such that $\pi(x)/f(x) \leq M$ for all x. We have shown that in this case the expected acceptance probability in stationarity is bounded below by $1/M$. We also have shown that we have uniform ergodicity with rate $(1 - 1/M)$. So it makes sense to choose θ which minimizes the bound M. However, we may not be able to do this analytically. A more empirical approach is to estimate the average acceptance rate in trial runs and then choose the parameter θ that gives us the largest acceptance rate.

While for the independence sampler it is advantageous to try to maximise the average acceptance rate, this is not the case for the random walk Metropolis-Hastings sampler. Suppose we have a Normal proposal distribution centred at the current value of the chain and with variance σ. If we choose σ very small, then we get a high acceptance rate but it takes a long time to explore the state space. The performance is particularly bad when we have a multi-modal target distribution and the chain has to traverse low probability regions in order to move from one mode to another. This is a typical situation in which meta-stability may arise. On the other hand if we choose a target distribution that proposes very large steps, then a lot of proposals will lie in the tails of the target distribution and thus are likely to be rejected. As a result the chain does not move at all. It is clear that we have to find an appropriate middle ground.

For a standard Normal target distribution with Normally distributed random walk proposals, researchers derived an optimal value for the variance σ of the proposal distribution by minimizing the following efficiency criterion:

$$\Big(1 + 2\sum_{k>0} \rho_k\Big).$$

From the discussion of the Monte Carlo error you will recognize this as the ratio of the asymptotic variance of the ergodic average $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$ in stationarity and the variance of the average based on an iid sample. The optimal value for the variance σ of the Normal proposal distribution turned out to be $\sigma = 2.4$ with the corresponding average acceptance rate of

$$\alpha \quad = \quad \frac{2}{\pi}\arctan(2/\sigma) \quad = \quad 0.44.$$

The authors did further investigations that lead to the recommendation of trying to achieve an acceptance rate close to $1/2$ for low dimensional problems and an acceptance rate close to $1/4$ for high dimensional problems. Please note that these are only rough guides. Really, you need to carefully explore the behaviour of your chain in trial runs. But it is worthwhile to keep in mind that for a random walk Metropolis-Hastings algorithm aiming at too high an acceptance probability is counter-productive.

One more word of warning: adaptive algorithms in which parameters are adjusted again and again during the simulation based on the past of the chain are becoming very popular. However, they tend to have complicated convergence properties and constant adaptation can indeed destroy convergence properties or alter the associated stationary distribution. A safe (but maybe not the best) option is to first calibrate your proposal distribution in trial runs and then fix the parameters and keep them fixed.

7 Gibbs Sampler

7.1 Introduction

Example **15** *Suppose $\underline{y} = (y_1, \dots, y_n)$ are iid observations from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Thus a likelihood function is given by*

$$L(\mu, \sigma^2|\underline{y}) \quad = \quad \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \Big)$$

We take a Bayesian approach and model prior information on μ and σ^2. Rather than specifying a prior on the variance σ^2, we model the precision $\tau = 1/\sigma^2$. We assume μ and τ are independent and that μ has a $\mathcal{N}(m, s^2)$ prior distribution and τ a Gamma(α, β) distribution. The posterior is then given by

$$\begin{aligned} \pi(\mu, \tau|\underline{y}) &\propto L(\mu, \tau^{-1}|\underline{y}) \exp\Big(-\frac{1}{2s^2}(\mu - m)^2 \Big) \exp(-\beta\tau)\tau^{\alpha-1} \\ &\propto \exp\Big(-\frac{1}{2s^2}(\mu - m)^2 \Big) \exp\Big(-\frac{\tau}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \Big) \exp(-\beta\tau)\tau^{\alpha-1+n/2} \end{aligned}$$

There is no closed form expression for the normalising constant of this posterior distribution and so we may consider sampling this distribution using some indirect sampling method. But notice the following. If we condition on τ, then the conditional distribution of μ has a standard form, namely Normal with mean $(n\bar{y}\tau + ms^{-2})/(n\tau + s^{-2})$ and precision $n\tau + s^{-2}$ where $\bar{y}$ is the sample average. If we condition on the other hand on μ then τ has a Gamma($\alpha + n/2, \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2$) distribution. Because these are standard distributions it seems a natural approach to use these as proposal distributions. The algorithm would proceed as follows. Choose an initial state $X_0 = (\mu_0, \tau_0)$ within the state space. Now suppose $X_n = (\mu_n, \tau_n)$ are given, then we produce X_{n+1} as follows:

1. *Sample $\mu \sim \mathcal{N}(\frac{n\bar{y}\tau_n + ms^{-2}}{n\tau_n + s^{-2}}, (n\tau_n + s^{-2})^{-1})$ and $U \sim$ Uniform(0,1).*
2. *If $U \leq \alpha((\tau_n, \mu_n), (\tau_n, \mu))$ then set $\mu_{n+1} = \mu$. Else set $\mu_{n+1} = \mu_n$.*
3. *Sample $\tau \sim$ Gamma$\Big(\alpha + n/2, \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_n)^2\Big)$ and $U \sim$ Uniform(0,1).*
4. *If $U \leq \alpha((\tau_n, \mu_{n+1}), (\tau, \mu_{n+1}))$ then set $\tau_{n+1} = \tau$. Else set $\tau_{n+1} = \tau_n$.*
5. *Set $X_{n=1} = (\mu_{n+1}, \tau_{n+1})$.*

Note the following:

- The fact that the normalizing constant of the posterior distribution is unknown does not matter, as it cancels out when computing the conditional distribution.
- As we will show a little later the acceptance probability $\alpha((\tau_n, \mu_n), (\tau_{n+1}, \mu_{n+1}))$ is constant one, and so we can simplify the algorithm and produce X_{n+1} as follows:

  1. Sample $\mu_{n+1} \sim \mathcal{N}(\frac{n\bar{y}\tau_n + ms^{-2}}{n\tau_n + s^{-2}}, (n\tau_n + s^{-2})^{-1})$.
  2. Sample $\tau_{n+1} \sim$ Gamma$(\alpha + n/2, \beta + \frac{1}{2}\sum_{i=1}^{n}(y_i - \mu_{n+1})^2$.
  3. Set $X_{n+1} = (\mu_{n+1}, \tau_{n+1})$.

Suppose we want to sample a random vector $(Y^{(1)}, \ldots, Y^{(d)})$ with probability mass function or density π. The *Gibbs Sampler* uses full conditional distribution to produce a d-dimensional Markov chain $\{(X_n^{(1)}, \ldots, X_n^{(d)}), n = 0, 1, \ldots\}$. Let us first define the full conditional distributions.

Definition 15 Full conditional distributions

(a) If $\mathcal{S}$ is discrete let $\pi(y^{(1)}, \ldots, y^{(d)})$ be the p.m.f. of the random vector $Y = (Y^{(1)}, \ldots, Y^{(d)})$. Set

$$\pi\left(y^{(-j)}\right) = \sum_{z \in \mathcal{S}} \pi(y^{(1)}, \ldots, y^{(j-1)}, z, y^{(j+1)}, \ldots, y^{(d)})$$

then the jth full conditional p.m.f. is given by

$$\begin{aligned} \pi_j\left(x \mid y^{(i)}, i \neq j\right) &= \mathbb{P}\left(Y^{(j)} = x \,\middle|\, Y^{(i)} = y^{(i)}, i \neq j\right) \\ &= \frac{\mathbb{P}\left(Y^{(j)} = x, Y^{(i)} = y^{(i)}, i \neq j\right)}{\mathbb{P}\left(Y^{(i)} = y^{(i)}, i \neq j\right)} \\ &= \frac{\pi(y^{(1)}, \ldots, y^{(j-1)}, x, y^{(j+1)}, \ldots, y^{(d)})}{\pi\left(y^{(-j)}\right)} \end{aligned}$$

(b) Now let $\mathcal{S}$ be continuous and let $\pi(y^{(1)}, \ldots, y^{(d)})$ be the density of the random vector $Y = (Y^{(1)}, \ldots, Y^{(d)})$. Define

$$\pi\left(y^{(-j)}\right) = \int_{\mathcal{S}} \pi(y^{(1)}, \ldots, y^{(j-1)}, z, y^{(j+1)}, \ldots, y^{(d)})dz$$

then the jth full conditional density is defined as

$$\pi_j\left(x \mid y^{(i)}, i \neq j\right) = \frac{\pi(y^{(1)}, \ldots, y^{(j-1)}, x, y^{(j+1)}, \ldots, y^{(d)})}{\pi\left(y^{(-j)}\right)}$$

We call the distribution defined by $\pi_j(x|y^{(i)}, i \neq j)$ the jth full conditional distribution of π. Note that we only need to determine the jth full conditional distribution up to a normalizing constant and so terms that do not depend on x can be ignored.

We can now define the general Gibbs Sampler: Assume $X_n = \underline{x} = (x^{(1)}, \ldots, x^{(d)}) \in \mathcal{S}$. Then proceed as follows:

1. Choose $j \in \{1, \ldots, d\}$ at random (or sequentially).
2. Sample x from $\pi_j(x|x^{(i)}, i \neq j)$ and set

$$X_{n+1} \quad = \quad \Big(x^{(1)}, \ldots, x^{(j-1)}, x, x^{(j+1)}, \ldots, x^{(d)}\Big).$$

The Gibbs sampler is a Metropolis-Hastings sampler that uses the full conditional distributions as proposal distributions. The proposal for the Gibbs sampler is always accepted as the following computations show. Suppose that the vector $\underline{y}$ is such that $x^{(i)} = y^{(i)}$ for $i \neq j$. Then the acceptance probability is

$$\begin{aligned} \alpha(\underline{x}, \underline{y}) &= \min\Big\{1, \frac{\pi(\underline{y})\pi_j(x^{(j)}|x^{(i)}, i \neq j)}{\pi(\underline{x})\pi_j(y^{(j)}|x^{(i)}, i \neq j)}\Big\} \\ &= \min\Big\{1, \frac{\pi(\underline{y})}{\pi(\underline{x})}\frac{\pi(\underline{x})/\pi(x^{(-j)})}{\pi(\underline{y})/\pi(y^{(-j)})}\Big\} \\ &= \min\Big\{1, \frac{\pi(\underline{y})}{\pi(\underline{x})}\frac{\pi(\underline{x})/\pi(x^{(-j)})}{\pi(\underline{y})/\pi(x^{(-j)})}\Big\} \quad = \quad 1. \end{aligned}$$

The Gibbs sampler is based on full conditional distributions. It is remarkable that full conditional distributions unlike marginal distributions can uniquely specify the joint distribution. This is due to the Hammersley-Clifford theorem which we state for the two-dimensional case.

Theorem 6 Hammersley-Clifford
If $\int \frac{\pi(y|x)}{\pi(x|y)}dy$ exists, then the joint density associated with the conditional densities $\pi(y|x)$ and $\pi(x|y)$ is given by

$$\pi(x, y) \quad = \quad \frac{\pi(y|x)}{\int \pi(y|x)/\pi(x|y)dy}.$$

Proof: We have $\pi(y|x)\pi(x) = \pi(x|y)\pi(y)$ and so

$$\int \frac{\pi(y|x)}{\pi(x|y)}dy = \int \frac{\pi(y)}{\pi(x)}dy = \frac{1}{\pi(x)}$$

and so the claim follows if the above integral exists. In fact the Gibbs sampler was originally developed for complex models that are defined through their full conditional distributions, the so-called local characteristics of Gibbs distributions. These are distributions from statistical mechanics like the Ising model who are defined through an energy function. Let's have a look at how Gibbs sampling works in the Ising model.

Example 16 Gibbs Sampler for Ising model
Suppose we want to sample the Ising model using the Gibbs Sampler. We first need to determine the full conditional distributions. Recall that

$$\pi(x^{(1)},\dots,x^{(m)}) \quad = \quad \frac{1}{Z}\exp\Big(J\sum_{(i,k)\in\mathcal{E}} x^{(i)}x^{(k)}\Big),$$

where $\mathcal{E} = \{(i,k): \ i\sim k\}$ the set of all neighbour pairs of sites. Let $\mathcal{E}_j = \{(i,k): i\sim k, \ j\in\{i,k\}\}$ be the set of neighbour pairs in which one of the sites is the site j. Then we have

$$\begin{aligned}
\pi_j(x|x^{(i)}, i\neq j) &= \frac{\pi(x^{(1)},\dots,x^{(j-1)},x,x^{(j+1)},\dots,x^{(m)})}{\sum_{z\in\{-1,1\}}\pi(x^{(1)},\dots,x^{(j-1)},z,x^{(j+1)},\dots,x^{(m)})}\\
&= \frac{\frac{1}{Z}\exp\Big(J\sum_{i\sim j}x^{(i)}x + J\sum_{(i,k)\in\mathcal{E}\setminus\mathcal{E}_j}x^{(i)}x^{(k)}\Big)}{\sum_{z\in\{-1,1\}}\frac{1}{Z}\exp\Big(J\sum_{i\sim j}x^{(i)}z + J\sum_{(i,k)\in\mathcal{E}\setminus\mathcal{E}_j}x^{(i)}x^{(k)}\Big)}\\
&= \frac{\exp\Big(J\sum_{i\sim j}x^{(i)}x\Big)\exp\Big(J\sum_{(i,k)\in\mathcal{E}\setminus\mathcal{E}_j}x^{(i)}x^{(k)}\Big)}{\sum_{z\in\{-1,1\}}\exp\Big(J\sum_{i\sim j}x^{(i)}z\Big)\exp\Big(J\sum_{(i,k)\in\mathcal{E}\setminus\mathcal{E}_j}x^{(i)}x^{(k)}\Big)}\\
&= \frac{\exp\Big(J\sum_{i\sim j}x^{(i)}x\Big)}{\sum_{z\in\{-1,1\}}\exp\Big(J\sum_{i\sim j}x^{(i)}z\Big)}\\
&= \frac{\exp\Big(J\sum_{i\sim j}x^{(i)}x\Big)}{\exp\Big(-J\sum_{i\sim j}x^{(i)}\Big)+\exp\Big(J\sum_{i\sim j}x^{(i)}\Big)}.
\end{aligned}$$

It follows that

$$\pi_j(1|x^{(i)}, i\neq j) \quad = \quad \frac{1}{\exp\Big(-2J\sum_{i\sim j}x^{(i)}\Big)+1}.$$

The Gibbs sampler now cycles through the sites (at random or sequentially) and updates the spin at site j to an upward spin $(x^{(j)}=1)$ with probability $\frac{1}{1+\exp(-2J\sum_{i\sim j}x^{(i)})}$ and to a downward spin $(x^{(j)}=-1)$ with probability $1-\frac{1}{1+\exp(-2J\sum_{i\sim j}x^{(i)})}$. Note that these probabilities do not depend on the normalising constant $1/Z$. Moreover, notice that the update probabilities for each site only depend on the values of its neighbour sites. In the statistical physics literature the above Gibbs Sampler is called the heat bath algorithm, because a heat bath causes spins to flip.

We can use detailed balance to show that the invariant distribution of the Gibbs Sampler is π.

Lemma 7 *Each transition of the Gibbs Sampler satisfies detailed balance with respect to π.*

Proof: Let $\underline{x} = (x^{(1)}, \ldots, x^{(d)})$ and $\underline{y} = (x^{(1)}, \ldots, x^{(j-1)}, z, x^{(j+1)}, \ldots, x^{(d)})$. Now, let $p(\underline{x}, \underline{y})$ denote the transition densities for the Gibbs Sampler. Then we have

$$\begin{aligned} \pi(\underline{x})p(\underline{x}, \underline{y}) &= \pi(\underline{x})\frac{\pi(\underline{y})}{\pi(x^{(-j)})} \quad = \quad \pi(\underline{y})\frac{\pi(\underline{x})}{\pi(x^{(-j)})} \\ &= \pi(\underline{y})\pi_j(x^{(j)}|x^{(i)}, i \neq j) \quad = \quad \pi(\underline{y})p(\underline{y}, \underline{x}). \end{aligned}$$

and so detailed balance holds.

The above theorem shows that the Gibbs Sampler chain has the desired invariant distribution. Thus, if the chain is ergodic then its distribution converges towards π. The following lemma provides a condition that ensures irreducibility:

Lemma 8 *Let $(X^{(1)}, \ldots, X^{(d)})$ be a random vector with joint density $\pi(x^{(1)}, \ldots, x^{(d)})$. Let $\pi_i(x)$ be the marginal density of $X^{(i)}$. If the fact that $\pi_i(x^{(i)}) > 0$ for all $i \in \{1, \ldots, d\}$ implies that $\pi(x^{(1)}, \ldots, x^{(d)}) > 0$ then the Gibbs sampler is π-irreducible.*

Example 17 *Consider the following bivariate distribution*

$$\pi(x, y) = \frac{1}{2}\mathbf{1}_{[0 \leq x \leq 1, 0 \leq y \leq 1]} + \frac{1}{2}\mathbf{1}_{[-1 \leq x \leq 0, -1 \leq y \leq 0]}$$

The conditional densities are given by

$$\begin{aligned} \pi(x|y) &= \begin{cases} \mathbf{1}_{[x \in (0,1)]} & \text{if } y \in (0,1) \\ \mathbf{1}_{[x \in (-1,0)]} & \text{if } y \in (-1,0) \end{cases} \\ \pi(y|x) &= \begin{cases} \mathbf{1}_{[y \in (0,1)]} & \text{if } x \in (0,1) \\ \mathbf{1}_{[y \in (-1,0)]} & \text{if } x \in (-1,0) \end{cases} \end{aligned}$$

so in principle we could run a Gibbs Sampler. However it would be reducible because if we started it say in the square $(0, 1) \times (0, 1)$ it would never be able to reach the other part of the state space which is the square $(-1, 0) \times (-1, 0)$. If we look at the marginal densities we find that

$$\begin{aligned} \pi_x(x) &= \frac{1}{2}\mathbf{1}_{[0 \leq x \leq 1]} + \frac{1}{2}\mathbf{1}_{[-1 \leq x \leq 0]} \text{ and} \\ \pi_y(y) &= \frac{1}{2}\mathbf{1}_{[0 \leq y \leq 1]} + \frac{1}{2}\mathbf{1}_{[-1 \leq y \leq 0]} \end{aligned}$$

which does not satisfy the positivity condition as $\pi_x(-0.5) > 0$ and $\pi_y(0.5) > 0$ but $\pi(-0.5, 0.5) = 0$.

It can be shown that Harris recurrence holds if the transition kernel of a π-irreducible Gibbs chain is absolutely continuous (with respect to an appropriate dominating measure).

7.2 Blocking and Re-parametrisation

The performance of the Gibbs Sampler is often poor when we have strong correlation between the individual components of the multi-dimensional target distribution. One way of avoiding this is by re-parametrising the distribution so that we sample variables with little correlation.

Example **18** **Bivariate Normal distribution** *Recall example 13. If the correlation ρ is close to one, then the performance of the Gibbs Sampler deteriorates. If we look at the full conditional distributions then this becomes very apparent. For example, x given y is updated according to a $\mathcal{N}(\rho y, 1-\rho^2)$ distribution, so, for ρ very close to one, x will be updated to a value close to y. Similarly, the y-component will be updated to a value that lies close to the value of the current x-component. The problem can be solved easily by sampling from the principal components of the bivariate Normal distribution. The principle components are also Normally distributed but are uncorrelated. Thus we can develop a Gibbs Sampler for the principle components and then simply derive our original variables from the appropriate linear combination of principle components. See Practical 3 for more details. (Note that because the principle components are Normal and are uncorrelated the re-parameterized Gibbs Sampler does actually sample the principle components exactly, not only approximately. So in this case we could sample from the target bivariate Normal using a simple transformation. However, this example illustrates very well the powerful effect that re-parametrisation can have.)*

Another approach to avoid poor performance due to high correlation is to block variables. This means that variables that are highly correlated are updated simultaneously. So for example if we have $(X^{(1)}, X^{(2)}, X^{(3)})$ and $X^{(1)}$ and $X^{(2)}$ have a strong correlation, then instead of sampling from the univariate full conditional distributions we may instead sample from the conditional distribution of $(X^{(1)}, X^{(2)}|X^{(3)})$ to update both the first and the second component simultaneously. After having updated the first two components, we update the third component using the usual univariate full conditional distribution.

7.3 Hierarchical models and graphical models

A set of distributions for which Gibbs Sampling is particular well adapted is that of hierarchical models. Here we have

$$\pi(x) = \int \pi_1(x|z_1)\pi_2(z_1|z_2)\dots\pi_k(z_{k-1}|z_k)\pi_{k+1}(z_k)dz_1\cdots dz_k,$$

where $z_1, \dots, z_k$ are hyper-parameters. Such models occur often when we are modelling complex data that requires the introduction of several layers of prior distribution.

Example 19 A normal hierarchical model:
Consider the following model:

$$\begin{aligned} Y_{ij} &\sim \mathcal{N}\Big(\alpha_i + \beta_i(x_j - \bar{x}), \frac{1}{\tau}\Big) \\ \alpha_i &\sim \mathcal{N}\Big(\alpha, \frac{1}{\tau_\alpha}\Big) \\ \beta_i &\sim \mathcal{N}\Big(\beta, \frac{1}{\tau_\beta}\Big) \\ \tau &\sim \mathrm{Gamma}(10^{-3}, 10^3), \qquad i \in \{1, \dots, m\}, j \in \{1, \dots, k\}. \end{aligned}$$

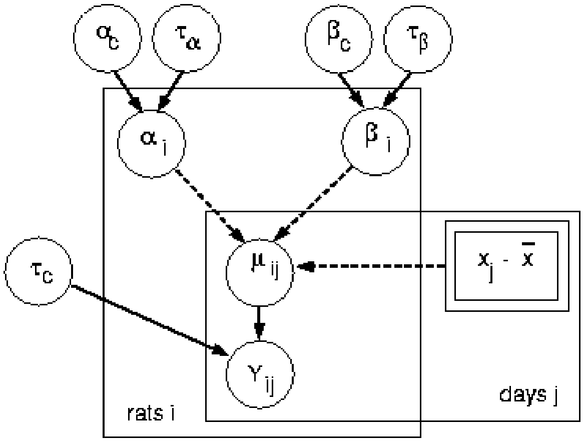
We have encountered a similar model when we introduced the Gibbs Sampler. But suppose now we would like to model the uncertainty we have in the prior parameters $\alpha, \tau_\alpha, \beta$ and τ_β. To do this we extend the model as follows:

$$\begin{aligned} \alpha &\sim \mathcal{N}(0, 10^6) \\ \tau_\alpha &\sim \mathrm{Gamma}(10^{-3}, 10^3) \\ \beta &\sim \mathcal{N}(0, 10^6) \\ \tau_\beta &\sim \mathrm{Gamma}(10^{-3}, 10^3) \end{aligned}$$

Note that these are very "flat" priors. You might want to work out the full conditional distributions for this example, but you may consider letting WinBUGS do the work. The above model has been used to analyse the growth in rats.

A convenient way of specifying joint distributions are graphical models. A graphical model uses a directed acyclic graph (DAG) to indicate the conditional independence within a joint distribution. We say X is conditionally independent of Z given Y if the conditional random variable $X|Y$ is independent of Z. The DAG consists of nodes and of directed edges (arrows) which are acyclic. This means that if you follow the directed edges you cannot return to a node. The nodes represent the variables of the model, we use squares to denote fixed variables and circles to denote random variables. The edges indicate the relationship between variables. Solid arrows denote a probabilistic relationship and dashed edges a deterministic one. DAGs are useful because they identify relevant variables when computing full conditional distributions. In WinBUGS you can specify a target distribution using a DAG with the Doodle Tool.

Example 20 Normal hierarchical model *The posterior distribution specified in the previous example is given by the following DAG. There is a slight change in notation, but it is a good exercise to work out which is which from the above model description.*



7.4 Data augmentation

A common application area for data augmentation is Bayesian missing data problems. Suppose the (complete data) likelihood for the parameter θ given data z is $L(\theta|z)$. However, some of the data is incomplete, that is $z = (z_{\text{obs}}, z_{\text{miss}})$ where z_{miss} has not been observed. Instead of using the complete data likelihood we now need to use the incomplete data likelihood $L(\theta|z_{\text{obs}})$ where

$$L(\theta|z_{\text{obs}}) \quad = \quad \int L(\theta|z_{\text{obs}}, z_{\text{miss}}) dz_{\text{miss}}.$$

Then the appropriate posterior is given by

$$\pi(\theta|z_{\text{obs}}) \quad \propto \quad L(\theta|z_{\text{obs}})\pi(\theta) \quad = \quad \int L(\theta|z_{\text{obs}}, z_{\text{miss}})\pi(\theta) dz_{\text{miss}}$$

which often will be more complex than the posterior model based on the complete data likelihood. We can avoid the problem by sampling from the joint posterior for the missing data and the parameter θ defined by

$$\pi(\theta, z_{\text{miss}}|z_{\text{obs}}) \quad \propto \quad \pi(\theta)L(\theta|z_{\text{obs}}, z_{\text{miss}})$$

and ignoring the values imputed for the missing data. We can use Gibbs sampling which alternatingly samples from the conditional distribution of z_{miss} given θ and z_{obs} and from the conditional distribution of θ given z_{miss} and y_{obs}.

Example 21 *(Robert, 2002)*
Consider data $\underline{y} = (y_1, \ldots, y_n)$ which are iid Poisson(λ) distributed. We do not receive this raw data, but the count data $\underline{x} = (x_0, x_1, x_2, x_3, x_4)$. For $i = 0, \ldots, 3$ the number x_i is the number of times i occurs in $\underline{y}$. However, x_4 is the number of $j \in \underline{y}$ such that $j \geq 4$. Thus for x_4 experiments we do not have the exact outcome. The incomplete data likelihood for this data is then

$$\exp(-\lambda \sum_{i=0}^{3} x_i) \prod_{i=0}^{3} (\frac{\lambda^i}{i!})^{xi} \Big(1 - e^{-\lambda} \sum_{i=0}^{3} \frac{\lambda^i}{i!}\Big)^{x_4}$$

Assuming $\underline{y}$ is ordered in descending order, we have that $y_{\text{mis}} = y_1, \ldots, y_{x_4}$. If we knew the missing data then the (complete data) likelihood would have the much simpler form:

$$\exp(-\lambda n) \prod_{i=0}^{3} (\frac{\lambda^i}{i!})^{x_i} \prod_{j=1}^{x_4} \frac{\lambda^{y_j}}{y_j!}$$

As prior for λ we choose $\pi(\lambda) = 1/\lambda$. Data Augmentation now produces a sample from the joint density of λ and y_{mis} given by

$$\pi(\lambda, y_{\text{mis}}|y_{\text{obs}}) \propto \frac{1}{\lambda} \exp(-\lambda n) \lambda^m \prod_{j=1}^{x_4} \frac{\lambda^{y_j}}{y_j!},$$

where $y_1, \ldots, y_{x_4} \geq 4$ and $m = \sum_{i=0}^{3} i x_i$. Suppose the current state of the Gibbs Sampler is $X_t = (\lambda, y_1, \ldots, y_{x_4})$. Then we update X_t as follows. Sample

$$\begin{aligned} \tilde{y}_j &\sim \text{Poisson}(\lambda) \textit{ conditional on } \tilde{y}_j \geq 4 \qquad \textit{for } j = 1, \ldots, x_4 \\ \tilde{\lambda} &\sim \text{Gamma}(m + \sum_{j=1}^{x_4} \tilde{y}_j, n) \end{aligned}$$

Data augmentation and thus Gibbs sampling may also be applied when introducing auxiliary variables. Auxiliary variables are not strictly missing data, but are introduced for example to make a likelihood function more tractable. (Recall that we have seen earlier how auxiliary variables and thus increasing the state space can make sampling more amenable!) A common area where we might consider using such an approach is mixture distributions.

Example 22 Bayesian mixture distribution: *Consider the following Normal mixture model with two components:*

$$f(y|\mu_1, \mu_2) \quad = \quad \frac{1}{2} \exp\Big(-\frac{1}{2}(y-\mu_1)^2\Big) + \frac{1}{2} \exp\Big(-\frac{1}{2}(y-\mu_2)^2\Big).$$

Now consider the data $\underline{y} = (y^{(1)}, \ldots, y^{(k)})$ which is iid each with density f. Then the joint density for data $\underline{y}$ is

$$f(\underline{y}|\mu_1, \mu_2) \quad = \quad \prod_{i=1}^{k} \Big[\frac{1}{2} \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big) + \frac{1}{2} \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big)\Big]$$

Now, if $p(\mu_1, \mu_2)$ is the prior density for μ_1 and μ_2 then the posterior density for μ_1 and μ_2 is given by

$$\pi(\mu_1, \mu_2|\underline{y}) \quad \propto p(\mu_1, \mu_2) \prod_{i=1}^{k} \Big[\exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big) + \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big)\Big].$$

Suppose we observed the vector $I = (I_1, \ldots, I_k)$ which indicates which data point was produced by which mixture component. Thus if $I_j = 1$ then $y_j \sim \mathcal{N}(\mu_1, 1)$ and if

$I_j = 2$ then $y_j \sim \mathcal{N}(\mu_2, 1)$. Then the density of $\underline{y}$ given I is

$$f(\underline{y}|\mu_1, \mu_2, I) \quad \propto \quad \prod_{i:I_i=1} \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big) \prod_{i:I_i=2} \exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big).$$

Now suppose the prior for the means is a bivariate Normal distribution with mean $\theta = (\theta_1, \theta_2)$ and covariance matrix $\Sigma = \mathbb{I}$, the identity matrix. Then

$$p(\mu_1, \mu_2) \quad = \quad p(\mu_1)p(\mu_2) \quad \propto \quad \exp(-\frac{1}{2}(\mu_1-\theta_1)^2)\exp(-\frac{1}{2}(\mu_2-\theta_2)^2).$$

Assume the uniform distribution on $\{0,1\}^k$ for I then the joint posterior of μ_1, μ_2 and I is given by

$$\pi(\mu_1, \mu_2, I|\underline{y}) \;\propto\; p(\mu_1)p(\mu_2)p(I)\exp\Big(-\frac{1}{2}(\sum_{i:I_i=1}(y^{(i)}-\mu_1)^2 + \sum_{i:I_i=2}(y^{(i)}-\mu_2)^2)\Big).$$

Note how much simpler the above expression is than the original posterior. We now develop a Gibbs Sampler to sample from the joint posterior distribution. First we need to derive the full conditional distributions. To ease notation, let H be the set of indices $i \in \{1, \ldots, k\}$ such that $I_i = 1$ and T the set of indices i such that $I_i = 2$. The first full conditional distribution is given by

$$\begin{aligned}
\pi_1(\mu_1|\mu_2, I, \underline{y}) \quad &\propto \quad p(\mu_1)p(\mu_2)\exp\Big(-\frac{1}{2}(\sum_{i\in H}(y^{(i)}-\mu_1)^2 + \sum_{i\in T}(y^{(i)}-\mu_2)^2\Big) \\
&\propto \quad p(\mu_1)\exp\Big(-\frac{1}{2}\sum_{i\in H}(y^{(i)}-\mu_1)^2\Big) \\
&\propto \quad \exp\Big(-\frac{1}{2}((\mu_1-\theta_1)^2 + \sum_{i\in H}(y^{(i)}-\mu_1)^2\Big) \\
&\propto \quad \exp\Big(-\frac{1}{2}(|H|+1)\Big[\mu_1 - \frac{\theta_1 + \sum_{i\in H} y^{(i)}}{(|H|+1)}\Big]^2\Big) \\
&\sim \quad \mathcal{N}\Big(\frac{\theta_1 + \sum_{i\in H} y^{(i)}}{(|H|+1)}, \frac{1}{(|H|+1)}\Big).
\end{aligned}$$

Analogously the second full conditional density $\pi_2(\mu_2|\mu_1, I, \underline{y})$ is the Normal density $\mathcal{N}\Big(\frac{\theta_2+\sum_{i\in T} y^{(i)}}{(|T|+1)}, \frac{1}{(|T|+1)}\Big)$. Finally the third full conditional p.m.f. is given by

$$\begin{aligned}
\pi_3(I|\mu_1, \mu_2, \underline{y}) \quad &\propto \quad \exp\Big(-\frac{1}{2}(\sum_{i\in H}(y^{(i)}-\mu_1)^2 + \sum_{i\in T}(y^{(i)}-\mu_2)^2)\Big) \\
&\propto \quad \prod_{i\in H}\exp\Big(-\frac{1}{2}(y^{(i)}-\mu_1)^2\Big)\prod_{i\in T}\exp\Big(-\frac{1}{2}(y^{(i)}-\mu_2)^2\Big).
\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{P}(I_i = 1|\mu_1, \mu_2, \underline{y}) &\propto \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_1)^2\Big) \quad \textit{and} \\ \mathbb{P}(I_i = 2|\mu_1, \mu_2, \underline{y}) &\propto \exp\Big(-\frac{1}{2}(y^{(i)} - \mu_2)^2\Big) \quad \textit{and so} \\ \mathbb{P}(I_i = 1|\mu_1, \mu_2, \underline{y}) &= \frac{\exp(-\frac{1}{2}(y^{(i)} - \mu_1)^2)}{\exp(-\frac{1}{2}(y^{(i)} - \mu_1)^2) + \exp(-\frac{1}{2}(y^{(i)} - \mu_2)^2)} \\ &= 1 - \mathbb{P}(I_i = 2|\mu_1\mu_2, \underline{y}).\end{aligned}$$

7.5 Swendsen-Wang algorithm

An example of data augmentation that has become very famous is the so-called Swendsen-Wang algorithm that allows us to sample the Ising model even close to the critical temperature. For each pair of neighbouring sites (i, j) it introduces an auxiliary variable $u_{ij} \in \{0, 1\}$, the so-called bonds. The joint distribution of the Ising model and these auxiliary variables is the so-called Fortuin-Kasteleyn model given by

$$\pi(\underline{x}, \underline{u}) \quad \propto \quad \prod_{<i,j>\in\mathcal{E}} \Big(\mathbf{1}_{[u_{ij}=0]}q + (1-q)\mathbf{1}_{[u_{ij}=1]}\mathbf{1}_{[x_i=x_j]}\Big),$$

where $q \in (0, 1)$ is a parameter of the model. Note that $\pi(\underline{x}, \underline{u}) > 0$ only for pairs $(\underline{x}, \underline{u})$ such that for any i, j: if $x_i \neq x_j$ then $u_{ij} = 0$.

Let N be the total number of edges and $k(\underline{x}) = |\{(< i, j >\in \mathcal{E} : x_i \neq x_j\}|$. Now choose $\beta = -\log(q)/2 > 0$. Then the marginal distribution of $\underline{X}$ can be shown to be an Ising model:

$$\begin{aligned}\pi(\underline{x}) &= \sum_{\underline{u}} \pi(\underline{x}, \underline{u}) \quad \propto \quad \sum_{\underline{u}} \prod_{<i,j>\in\mathcal{E}} \Big(\mathbf{1}_{[u_{ij}=0]}q + (1-q)\mathbf{1}_{[u_{ij}=1]}\mathbf{1}_{[x_i=x_j]}\Big) \\ &= \sum_{\underline{u}} q^{k(\underline{x})} \prod_{<i,j>\in\mathcal{E}:x_i=x_j} \Big(q\mathbf{1}_{[u_{ij}=0]} + (1-q)\mathbf{1}_{[u_{ij}=1]}\Big) \\ &\propto q^{k(\underline{x})} \sum_{s=0}^{N-k(\underline{x})} \left(\begin{array}{c} N-k(\underline{x}) \\ s \end{array}\right) q^s(1-q)^{N-k(\underline{x})-s} \\ &= \exp\Big(-2\beta k(\underline{x})\Big) \\ &= \exp\Big(\beta[(N-k(\underline{x})) - k(\underline{x})]\Big) \\ &= \exp\Big(\beta \sum_{<i,j>\in\mathcal{E}} x_i x_j\Big).\end{aligned}$$

To compute the marginal distribution of $\underline{U}$ first define

$$A(\underline{u}) \quad = \quad \{\underline{x} : \text{if } u_{ij} = 1 \text{ then } x_i = x_j\}$$

and note that if $\underline{x} \notin A(\underline{u})$ then $\pi(\underline{x}, \underline{u}) = 0$. Now,

$$\begin{aligned}\pi(\underline{u}) &= \sum_{\underline{x} \in A(\underline{u})} \pi(\underline{x}, \underline{u}) \quad \propto \quad \sum_{\underline{x} \in A(\underline{u})} \prod_{<i,j>:u_{ij}=0} q + \prod_{<i,j>:u_{ij}=1} (1-q)\mathbf{1}_{[x_i = x_j]}\Big) \\ &= \prod_{u_{ij}=0} q \prod_{u_{ij}=1} (1-q) 2^{c(\underline{u})}\end{aligned}$$

We call neighbour sites i and j connected if $u_{ij} = 1$. In the above $c(\underline{u})$ is the number of connected components in the configuration $\underline{u}$. The above model is the so-called random cluster model.

The remarkable thing is that the conditional distributions are very simple indeed. Given the values of the sites $\underline{X} = \underline{x}$, the distribution of bonds is given by

$$\pi(u_{ij}|x_i, x_j) \quad \propto \quad q\mathbf{1}_{[u_{ij}=0]} + (1-q)\mathbf{1}_{[u_{ij}=1]}\mathbf{1}_{[x_i=x_j]}$$

Thus if two neighbouring sites i and j have the same spin, that is $x_i = x_j$ then we set $u_{ij} = 0$ with probability q and $u_{ij} = 1$ otherwise. If i and j have different spin then we set $u_{ij} = 0$ deterministically.

The conditional distribution of $\underline{x}$ given $\underline{U} = \underline{u}$ is also easy. The distribution of spins is uniform conditional on no connected sites having different spins as the following computation shows. Note that we only need to consider $x \in A$ as defined above.

$$\pi(\underline{x}|\underline{u}) \quad \propto \prod_{<i,j)>:u_{ij}=1} \mathbf{1}_{[x_i=x_j]}. \tag{2}$$

Thus we simply assign each connected cluster of sites the same spin, which is upwards with probability 1/2 and downwards otherwise. The Swendsen-Wang algorithm is simply a Gibbs Sampler that samples $\underline{x}$ given $\underline{u}$ and $\underline{u}$ given $\underline{x}$.

8 Reversible jump MCMC

This is a method that is applicable to problems when "one of the things we don't know is how many things we don't know" (Prof. Green). There is a large collection of problems for which we need to define models of variable dimension. Such problems include model selection, model comparison and object recognition. For example we have discussed mixture models. Often we don't know how many components the mixture contains and part of the inference should be to determine the most appropriate number.

Consider the following general set-up. We have a collection of models:

$$\mathcal{M}_k \quad = \quad \{f_k(\cdot|\theta_k), \theta_k \in \Theta_k\}.$$

For example for mixture distributions the model $\mathcal{M}_k$ would be a mixture distribution with k components. For each k we have a prior $\pi_k(\theta_k)$ on the model parameters. This defines a collection of posterior distributions conditional on the model index k

$$\pi(\theta_k|k, \underline{x}) \quad \propto \quad \pi_k(\theta_k) f_k(\underline{x}|\theta_k).$$

We also have a prior $\rho(k)$ on the model index k which can either be bounded above ($k \in \{1, \ldots, K\}$) or there may be an infinite number of potential models ($k \in \mathbb{N}$). Given data $\underline{x}$ our target distribution is given by the posterior density

$$\pi(k, \theta_k | \underline{x}) \quad \propto \quad \rho(k)\pi(\theta_k | k, \underline{x}).$$

This allows us to evaluate quantities such as the posterior probability for model $\mathcal{M}_k$ given by

$$p(\mathcal{M}_k | \underline{x}) \quad \propto \quad \int_{\Theta_k} \rho(k)\pi(\theta_k | k, \underline{x}) d\theta_k.$$

To be able to sample a model of variable dimension the sampler needs to be able to move within and between models. Moves within a model can be done using any of the techniques discussed before, but moves between models of different dimension need a new approach.

As with any MCMC algorithm discussed so far, we need to ensure detailed balance (and thus time-reversibility). If $p(x, \cdot)$ is our transition kernel and π our target distribution this boils down to the equation:

$$\int_A \int_B p(x, dy)\pi(x) dx \quad = \quad \int_A \int_B p(y, dx)\pi(y) dy.$$

If $q_k(x, dy)$ is the proposal transition kernel that we use if the current state is a model of type $\mathcal{M}_k$ and $\alpha_k(x, y)$ is the appropriate acceptance probability, then

$$p(x, B) \quad = \quad \sum_k \int_B q_k(x, dy')\alpha(x, y') + r(x)\mathbf{1}_B(x).$$

Here $r(x)$ is the probability of a move from x being rejected. (Note that this is the usual expression for the transition kernel of a Metropolis-Hastings chain, just that now we also have to sum over all possible model indices k.) Now comes the technical (measure-theoretical) bit. We need $\pi(dx)q_k(x, dy)$ to be absolutely continuous with respect to a symmetric measure $\xi_k(dx, dy)$ on $\Theta \times \Theta$ where $\Theta = \sum_k \{k\} \cup \Theta_k$. Let $g_k(x, y)$ be the density of $q_k(x, dy)\pi(dx)$ with respect to $\xi_k(dx, dy)$ then the acceptance probability

$$\alpha_k(x, y) \quad = \quad \min\{1, \frac{g_k(y, x)}{g_k(x, y)}\}$$

ensures detailed balance as the following computations show:

$$\begin{aligned}
\int_A \int_B \alpha_k(x, y) q_k(x, dy)\pi(dx) \quad &= \quad \int_A \int_B \alpha_k(x, y) g_k(x, y) \xi_k(dx, dy) \\
&= \quad \int_A \int_B \alpha_k(y, x) g_k(y, x) \xi_k(dy, dx) \\
&= \quad \int_A \int_B \alpha_k(y, x) q_k(y, dx) \pi(dy).
\end{aligned}$$

Green defined a method that avoids having to explicitly define the appropriate symmetric measure. This method is called reversible jump (or trans-dimensional) MCMC and is based on the following dimension matching criterion. First consider a move from $\mathcal{M}_k$ to model $\mathcal{M}_j$. The way we usually generate such a move is by sampling a (possibly multivariate) variable $u \sim g(u)$ and proposing $h_1(\theta_k, u)$ the value of the deterministic function h_1 evaluated at the current θ_k and u. Similary, to move from model $\mathcal{M}_j$ to model $\mathcal{M}_k$ we sample $v \sim f(v)$ and propose $\theta_k = h_2(\theta_j, v)$ for some deterministic function h_2. The dimension matching criterion requires that

$$\dim(\theta_k, u) \quad = \quad \dim(\theta_j, v).$$

This means that there is a bijective function T such that

$$T(\theta_k, u) \quad = \quad (\theta_j, v).$$

Recall that our states are given by (i, θ_i) where i is the index of the model and θ_i are the corresponding model parameters. Suppose the probability of proposing a move from model k_1 to model k_2 is $p_{k_1 k_2}$. Green then worked out that the appropriate probability of accepting state (j, θ_j) where $\theta_j = h_1(\theta_k, u)$ given the chain is currently in state (k, θ_k) is given by

$$\alpha((k, \theta_k), (j, \theta_j)) \quad = \quad \min\left\{1, \frac{\pi(j, \theta_j | \underline{x})}{\pi(k, \theta_k | \underline{x})} \frac{p_{jk} f(v)}{p_{kj} g(u)} \left| \frac{\partial T(\theta_k, u)}{\partial(\theta_k, u)} \right| \right\}.$$

where the last factor is the absolute value of the Jacobian determinant of T and v is such that $\theta_k = h_2(\theta_j, v)$. This is often called the Green-Hastings acceptance probability.

Example 23 *Suppose we have two models: the first model $\mathcal{M}_1$ has parameter $\theta \in \mathbb{R}$ and the second model $\mathcal{M}_2$ has two parameters $\theta_1, \theta_2 \in \mathbb{R}$. When we would like to move from the second model to the first we may propose $\theta = (\theta_1 + \theta_2)/2$. Thus to go from the first model to the second model we can only use a univariate random variable u. So suppose we sample $u \sim g(u)$ and set $\theta_1 = \theta - u$ and $\theta_2 = \theta + u$. Then*

$$\dim(\theta, u) \quad = \quad \dim(\theta_1, \theta_2).$$

The Jacobian of $T(\theta, u) = (\theta - u, \theta + u)$ is

$$\left| \frac{\partial T(\theta, u)}{\partial(\theta, u)} \right| = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2$$

and so we accept the move from $(1, \theta)$ to $(2, \theta_1, \theta_2) = (2, \theta - u, \theta + u)$ with probability

$$\alpha((1, \theta), (2, \theta_1, \theta_2)) \quad = \quad \min\{1, \frac{\pi(2, \theta_1, \theta_2)}{\pi(1, \theta)} \frac{p_{21}}{p_{12}} \frac{2}{g((\theta_2 - \theta_1)/2)}\}.$$

As an exercise show that the probability of accepting a move from $(2, \theta_1, \theta_2)$ to $(1, \theta) = (1, (\theta_1 + \theta_2)/2)$ should be

$$\alpha((2, \theta_1, \theta_2), (1, \theta)) \quad = \quad \min\{1, \frac{\pi(1, \theta)}{\pi(2, \theta_1, \theta_2)} \frac{p_{12}}{p_{21}} \frac{g((\theta_2 - \theta_1)/2)}{2}\}.$$

Example 24 *Suppose you have an iid sample* $\underline{y} = (y_1, \ldots, y_n)$*. We believe that the data is either from an Exponential distribution with parameter* λ *or from a Gamma(*β*,*γ*) distribution. We choose the prior probabilities of each model equal to 0.5 and the following prior distributions for the model parameters:*

$$\begin{aligned}\lambda &\sim \Gamma(a, b)\\ \beta &\sim \Gamma(c, d)\\ \gamma &\sim \Gamma(e, f).\end{aligned}$$

First consider the posterior distribution conditional on the first model. We have

$$\pi((1, \lambda)|\underline{y}) \quad \propto \quad \lambda^{n+a-1}\exp(-\lambda(b + \sum_{i=1}^{n} y_i))$$

and so λ *has a Gamma(*$n + a - 1$*,* $b + \sum_{i=1}^{n} y_i$*) distribution. Now consider the second model and set* $z = \prod_{i=1}^{n} y_i$*. We have*

$$\begin{aligned}\pi(2, \beta, \gamma|\underline{y}) &\propto \beta^{c-1}\exp(-\beta d)\gamma^{e-1}\exp(-\gamma f)\frac{\gamma^{n\beta}}{\Gamma(\beta)^n}z^{\beta-1}\exp(-\gamma\sum_{i=1}^{n} y_i), \text{ and so}\\ \pi(\beta|2, \gamma, \underline{y}) &\propto \beta^{c-1}\exp(-\beta d)\frac{\gamma^{n\beta}}{\Gamma(\beta)^n}z^{\beta-1}\\ \pi(\gamma|2, \beta, \underline{y}) &\propto \gamma^{n\beta+e-1}\exp(-\gamma(f + \sum_{i=1}^{n} y_i))\end{aligned}$$

Thus in Model 2, the full conditional distribution for γ *is Gamma(*$n\beta + e, f + \sum_{i=1}^{n} y_i$*). The full conditional distribution for* β *is non-standard, but we can use a Metropolis-Hastings step to update* β *as follows. Choose some value* $\epsilon > 0$*, then we choose a new value* β' *uniformly from the interval* $[\beta - \epsilon, \beta + \epsilon]$*. This may produce negative values for* β' *but we then simply reject* β'*. The acceptance probability is thus given by*

$$\alpha((2, \beta, \gamma), (2, \beta', \gamma)) = \mathbf{1}_{[\beta'>0]}\min\{1, \frac{\pi(\beta'|2, \gamma, \underline{y})}{\pi(\beta|2, \gamma, \underline{y})}\}$$

This specifies the within model moves, but we also need to specify the between model moves. The easiest thing to do is to always propose the other model. Thus if we are in model 1 we propose model 2 and vice versa. Now we need to decide how to generate the parameters of the proposed model. A common approach is to choose the parameters such that the models have the same mean as this will facilitate good mixing properties between the model. This means if we have β *and* γ *then we propose* $\lambda = \gamma/\beta$*. This fully determines the dimensions and we need to use a univariate random variables to generate values for* β *and* γ *given* λ*. One way of doing this is to sample* $u \sim \Gamma(c, d)$ *and setting* $\beta = u$ *and* $\gamma = \lambda u$*. The corresponding Jacobian is given by*

$$\begin{vmatrix} 0 & 1 \\ u & \lambda \end{vmatrix}$$

which has a determinant equal to u*. This means we have acceptance probabilities*

$$\alpha((1,\lambda),(2,\beta,\gamma)) = \min\{1, \frac{\pi(2,\beta,\lambda|\underline{y})}{\pi(1,\lambda|\underline{y})} \frac{1}{\frac{d^c}{\Gamma(c)}\beta^{c-1}\exp(-d\beta)}\beta\}$$

and

$$\alpha((2,\beta,\gamma),(1,\lambda)) = \min\{1, \frac{\pi(1,\lambda|\underline{y})}{\pi(2,\beta,\gamma|\underline{y})} \frac{\frac{d^c}{\Gamma(c)}\beta^{c-1}\exp(-d\beta)}{\beta}\}$$

respectively.

The whole reversible jump algorithm now works as follows. Suppose the current state of the chain is $X_t = (1,\lambda)$ *then we first perform a within model update followed by a between model update:*

1. *Sample* $\lambda' \sim$ *Gamma(*$n+a-1$*,* $b + \sum_{i=1}^n y_i$*) and set* $X_{t+1} = (1,\lambda')$*.*

2. *Sample* $u \sim$ *Gamma(*c, d*) and set* $\beta = u$ *and* $\gamma = \lambda' u$*.*

3. *With probability*

   $$\alpha((1,\lambda'),(2,\beta,\gamma)) = \min\{1, \frac{\pi(2,\beta,\lambda)}{\pi(1,\lambda)} \frac{1}{\frac{d^c}{\Gamma(c)}\beta^{c-1}\exp(-d\beta)}\beta\}$$

   set $X_{t+2} = (2,\beta,\gamma)$*. Alternatively set* $X_{t+2} = (1,\lambda')$*.*

Similarly, suppose $X_t = (2,\beta,\gamma)$*, then perform the following steps.*

1. *Sample* $\gamma' \sim$ *Gamma(*$n\beta + e$*,*$f + \sum_{i=1}^n y_i$*).*

2. *Sample* β' *uniformly on* $[\beta-\epsilon, \beta+\epsilon]$*. With probability*

   $$\alpha((2,\beta,\gamma'),(2,\beta',\gamma')) = \mathbf{1}_{[\beta'>0]}\min\{1, \frac{\pi(2,\beta',\gamma')}{\pi(2,\beta,\gamma')}\}$$

   set $X_{t+1} = (2,\beta_{t+1},\gamma_{t+1}) = (2,\beta',\gamma')$*. Else set* $X_{t+1} = (2,\beta_{t+1},\gamma_{t+1}) = (2,\beta,\gamma')$*.*

3. *Set* $\lambda = \gamma_{t+1}/\beta_{t+1}$*. Then with probability*

   $$\alpha((2,\beta_{t+1},\gamma_{t+1}),(1,\lambda)) = \min\{1, \frac{\pi(1,\lambda)}{\pi(2,\beta_{t+1},\gamma_{t+1})} \frac{\frac{d^c}{\Gamma(c)}\beta_{t+1}^{c-1}\exp(-d\beta_{t+1})}{\beta_{t+1}}\}$$

   set $X_{t+2} = (1,\lambda)$*. Else set* $X_{t+2} = X_{t+1}$*.*

9 Simulated annealing, simulated tempering and parallel tempering

Consider a set of distributions $\pi_T(x) \propto \pi(x)^{\frac{1}{T}}$ where $T > 0$. Borrowing terminology from physics we call the scaling parameter T temperature. Now consider a

Metropolis-Hastings algorithm with stationary distribution π_T. For large T, that is at hot temperatures, π_T is relative flat and thus the MH-algorithm will find it easy to explore the support of π_T. As T decreases, i.e. becomes colder, modes will become more pronounced and it will be harder to explore the state space. As T decreases towards 0 the states visited by the chain will become concentrated around the local maxima of π.

9.1 Simulated tempering

Here we exploit the fact that for hotter distributions the corresponding Markov chains tend to mix better. We use a finite number of temperatures $T_0 = 1 < T_1 < \ldots < T_m$ which are used to define an augmented target distribution. We introduce an auxiliary variable that denotes the index of the current temperature and define the augmented target distribution as

$$\pi(x, j) \quad \propto \quad c_j \; \pi_{T_j}(x)$$

We then design a Markov chain whose stationary distribution is the above augmented target distribution. Given a run of the chain we then simply retain the states (x_t, i_t) for which $i_t = 1$ and this gives us a sample whose asymptotic distribution is π.

A chain whose invariant distribution is $\pi(x, i)$ can now be designed as follows. We alternate updates of the x component with updates of the temperature index i. Suppose that the proposal kernel for the x component is defined via $q(x, y)$. We define a proposal probability mass function $r(i, j)$ for the temperature index as follows:

$$\begin{aligned} r(i, i+1) &= r(i, i-1) &= \frac{1}{2} \qquad \text{for } i \in \{2, \ldots, m-1\} \\ r(m, m-1) &= r(1, 2) &= 1. \end{aligned}$$

The algorithm now works as follows. Suppose the current state of the chain is $X_t = (x_t, i_t)$.

1. Sample $I = j$ according to $r(i_t, j)$.
2. Accept $I = j$ and set $i_{t+1} = j$ with probability

   $$\alpha\Big(x_t, i_t, (x_t, j)\Big) = \min\{1, \frac{c_j \pi_{T_j}(x_t) r(j, i_t)}{c_{i_t} \pi_{T_{i_t}}(x_t) r(i_t, j)}\}$$

   Else set $i_{t+1} = i_t$.
3. Sample $Y = y$ from $q(x_t, y)$.
4. Accept $Y = y$ and thus set $x_{t+1} = y$ with probability

   $$\alpha\Big((x_t, i_{t+1}), (y, i_{t+1})\Big) = \min\{1, \frac{c_{i_{t+1}} \pi_{T_{i_{t+1}}}(y) q(y, x_t)}{c_{i_{t+1}} \pi_{T_{i_{t+1}}}(x_t) q(x_t, y)}\}.$$

   Else set $x_{t+1} = x_t$.
5. Set $X_{t+1} = (x_{t+1}, i_{t+1})$.

The idea of this approach is to exploit the fact that MCMC algorithms tend to mix better in ”flatter” distributions. Whenever the chain visits a hot distribution it will be easier for it to move from one part of the state space to another. Thus by the time it returns to the cold target distribution it will have moved to a different region of the space. For this to work efficiently it is generally recommended that the constants c_i are chosen such that the chain spends roughly the same amount of time in each of the m temperatures. If π_{T_i} is normalised then this will happen if all $c_i = 1$. However, if they are not normalised then c_i should be equal to normalising constant of the corresponding density/pmf. However, in this lecture course we have encountered numerous distributions where MCMC was necessary for the very reason that we did not know the normalising constant. The solution is to use approximations to the normalising constants often based on trial runs or alternative use an adaptive algorithm. (If you would like to find out more search for the keywords “logistic regression” or “Wang-Landau” in the MCMC literature).

Another implementational issue of this method is how to choose the temperature levels. There is a trade-off between having a sufficiently high acceptance probability for moving from one temperature to another and not having too many temperature levels. Again there is a vast amount of literature on this issue, notably in Physics.

9.2 Parallel tempering

Parallel tempering or Metropolis-coupled MCMC also uses different levels of temperatures. In contrast to simulated tempering it runs m chains in parallel such that the ith chain has stationary distribution π_{T_i}. After an update within each of the m chains we pick a pair of chains and propose to swap their states. This is followed by a Metropolis-Hastings accept/reject step to maintain the correct stationary distribution. Here is a summary of the algorithm. Suppose the current state of the jth chain is $X_t^{(j)} = x_t^{(j)}$ for $j \in \{1, \dots, m\}$. Then

1. For $j \in \{1, \dots, m\}$ sample $Y^{(j)} = y^{(j)}$ according to $q_j(x_t^{(j)}, y^{(j)})$ and set $x_{t+1}^{(j)} = y^{(j)}$ with probability

   $$\alpha(x_t^{(j)}, y^{(j)}) = \min\{1, \frac{\pi_{T_j}(y^{(j)}) q_j(y^{(j)}, x_t^{(j)})}{\pi_{T_j}(x_t^{(j)}) q_j(x_t^{(j)}, y^{(j)})}\}.$$

   Else set $x_{t+1}^{(j)} = x_t^{(j)}$.

2. Choose $i, j \in \{1, \dots, m\}$ at random subject to $i \neq j$.

3. With probability

   $$\alpha(x_{t+1}^{(i)}, x_{t+1}^{(j)}) = \min\{1, \frac{\pi_{T_i}(x_t^{(j)}) \pi_{T_j}(x_{t+1}^{(i)})}{\pi_{T_j}(x_{t+1}^{(j)}) \pi_{T_i}(x_{t+1}^{(i)})}\}.$$

   Set $X_{t+2}^{(i)} = x_{t+1}^{(j)}$, $X_{t+2}^{(j)} = x_{t+1}^{(i)}$ and $X_{t+2}^{(k)} = x_{t+1}^{(k)}$ for $k \neq i, j$.

The advantage of parallel tempering is that we do not have to estimate the normalizing constants c_i, but this comes at the cost of having m parallel chains. However, with todays parallel computer technology this is not really a problem.

9.3 Simulated annealing

The simulated annealing algorithm is an optimisation technique that is strongly related to the MCMC methods we have discussed so far. The aim of simulated annealing is to find the location of the maximum of a function $h(x)$. We can easily transform the problem into finding the mode of a distribution by setting $\pi(x) \propto \exp(h(x))$. We now run an MCMC algorithm but with a target distribution that changes with time. At iteration t the target distribution is $\pi_{T(t)}(x)$ where $T(t) \to 0$ as $t \to \infty$. Note that it is important that $T(t)$ does not decrease too quickly towards 0 to avoid getting stuck in a local mode. Temperature schemes for which convergence is assured can be found in the literature.

Example 25 *Suppose we would like to maximize*

$$h(x) \quad = \quad [\cos(50x) + \sin(20x)]^2 \qquad \textit{for } x \in [0,1].$$

Our MCMC algorithm proposes new values uniformly within distance r from the current value of the chain. Thus if $X_t = x_t$ is the current state of the chain we produce state X_{t+1} as follows:

1. *Sample $Y = y$ uniformly from the interval (a_t, b_t) where $a_t = \max\{x_t - r, 0\}$ and $b_t = \min\{x_t + r, 1\}$ and propose y as the new state.*
2. *Let $a = \max\{y - r, 0\}$ and $b = \min\{y + r, 1\}$ then we accept y with probability*

   $$\alpha_t(x_t, y) \quad = \quad \min\left\{1, \exp\left(\frac{1}{T(t)}(h(y) - h(x_t))\right)\frac{b_t - a_t}{b - a}\right\}.$$

   If accepted set $x_{t+1} = y$, else set $x_{t+1} = x_t$.
3. *Reduce the temperature to $T(t+1)$.*

A suitable cooling scheme in this context would be $T(t) = 1/\log(t)$.

A more complex example where the temperature is related to physical reality is the Ising model. Below is a pseudo-code description of the algorithm.

Example 26

```
Initialize $x$ and $T$
for $t = 1$ to $N$
    for $i \in \mathcal{S}$
        $d = \beta \sum_{j:i\sim j} x^{(i)} x^{(j)}$
        $U \sim$ Uniform(0,1)
        if $\log(U) < \min\{0, -2d/T\}$ then
            $x^{(i)} = -x^{(i)}$
    $T = T(t)$
```

10 The EM algorithm

10.1 Introduction

The aim of the EM algorithm is to maximise the likelihood function in data augmentation problems, that is when data is missing or when the introduction of a latent variable makes the likelihood function much simpler.

Similarly to what we discussed in the section on the data augmentation algorithm, we assume that $X = x$ is the observed data, $Y = y$ the missing or latent data, and $Z = (x, y)$ the complete data. The "incomplete data" likelihood is given by

$$L(\theta|x) \quad = \quad p(x|\theta) \quad = \quad \int p(x, y|\theta)dy.$$

Here, $p(x, y|\theta)$ is the joint probability density or mass function for the observed and missing data and $p(x|\theta)$ is the appropriate marginal density/pmf for the observed data. The integration step on the right hand side makes maximisation of the incomplete data likelihood computationally difficult.

The "complete data" likelihood is given by

$$L(\theta|x, y) \quad = \quad p(x, y|\theta) \quad = \quad p(x|\theta)p(y|x, \theta).$$

Note that

$$p(x|\theta) \quad = \quad \frac{p(x, y|\theta)}{p(y|x, \theta)},$$

so we can write the incomplete data log-likelihood as

$$l(\theta|x) \quad = \quad l(\theta|x, y) - \log\Big(p(y|x, \theta)\Big).$$

This is of course conditional on $Y = y$. Assume $\theta^{(i)}$ is an approximation on θ and take expectations with respect to the conditional distribution of Y given x and $\theta^{(i)}$. Then we can write

$$\begin{aligned} l(\theta|x) &= \mathbb{E}_{Y|x,\theta^{(i)}}\Big(l(\theta|x, y)\Big) - \mathbb{E}_{Y|x,\theta^{(i)}}\Big(\log(p(y|x, \theta))\Big) \\ &= Q(\theta, \theta^{(i)}) - H(\theta, \theta^{i}). \end{aligned} \tag{3}$$

Example 27 Censored Normal data:
Suppose $z = \{z_1, \ldots, z_n\}$ *is iid* $\mathcal{N}(\mu, 1)$*. We observe the censored data* $x = \{x_1, \ldots, x_n\} =$

$\{z_1\mathbf{1}_{[z_1>0]}, \dots, z_n\mathbf{1}_{[z_n>0]}\}$. Assume m is such that $x_k > 0$ for $1 \le k \le m$ and $x_k = 0$ for $m+1 \le k \le n$. The incomplete data likelihood is then given by

$$L(\mu|x) \quad = \quad \prod_{k=1}^{m} \phi(x_k - \mu) \prod_{k=m+1}^{n} \Phi(-\mu),$$

where ϕ is the pdf of a standard Normal distribution and Φ is the corresponding cdf. The complete data likelihood is given by

$$L(\mu|x, y) \quad = \quad \prod_{k=1}^{m} \phi(x_k - \mu) \prod_{k=m+1}^{n} \phi(y_k - \mu)\mathbf{1}_{[y_k \le 0]}.$$

Note that

$$\int L(\mu|x, y)dy \quad = \quad L(\mu|x).$$

Now assuming $y_{m+1}, \dots y_n$ are less or equal to zero we have

$$\begin{aligned}
l(\mu|x, y) \quad &= \quad -\frac{1}{2}\sum_{k=1}^{m}(x_k - \mu)^2 - \frac{1}{2}\sum_{k=m+1}^{n}(y_i - \mu)^2 \quad (+\log(c)) \\
\textit{and so} \quad & \\
Q(\mu, \mu^{(i)}) \quad &= \quad \mathbb{E}_{Y|x,\mu^{(i)}}\Big(l(\theta|x, y)\Big) \\
&= \quad -\frac{1}{2}\sum_{k=1}^{m}(x_k - \mu)^2 - \frac{1}{2}\sum_{k=m+1}^{n}\mathbb{E}_{Y|x,\mu^{(i)}}\Big((Y_k - \mu)^2 \;\Big|\; Y_k \sim \mathcal{N}(\mu^{(i)}, 1), Y_k \le 0\Big).
\end{aligned}$$

We will use the notation

$$\mathbb{E}_{\mu^{(i)},0}\Big(f(Y)\Big) \quad = \quad \mathbb{E}_{Y|x,\mu^{(i)}}\Big(f(Y) \;\Big|\; Y \sim \mathcal{N}_(\mu^{(i)}, 1), Y \le 0\Big).$$

Now, under appropriate regularity conditions, the argument θ that maximises the incomplete data log-likelihood is a root of the following derivative

$$\frac{\partial l(\theta|x)}{\partial\theta} \quad = \quad \frac{\partial Q(\theta, \theta^{(i)})}{\partial\theta} \quad - \quad \frac{\partial H(\theta, \theta^{(i)})}{\partial\theta}.$$

The partial derivatives of $H(\theta, \theta^{(i)})$ can be shown to be zero if $\theta = \theta^{(i)}$, so for now we concentrate on finding the roots of the partial derivatives of $Q(\cdot, \theta^{(i)})$. The EM-algorithm is an iterative procedure that proceeds as follows:

1. Set an initial value $\theta^{(0)}$ and then iterate the following two steps:
2. **E-step**:
   Compute the expectation

$$Q(\theta, \theta^{(i)}) \quad = \quad \mathbb{E}_{Y|x,\theta^{(i)}}\Big(l(\theta|x, y)\Big) \quad = \quad \int \log(p(x, y|\theta))p(y|x, \theta^{(i)})dy.$$

3. **M-step:**
   Maximise $Q(\theta, \theta^{(i)})$ and set

$$\theta^{(i+1)} \quad = \quad \mathrm{argmax}_\theta \; Q(\theta, \theta^{(i)}).$$

Example 28 *Censored Normal distribution continued:*
E-step:

$$Q(\mu, \mu^{(i)}) \quad = \quad -\frac{1}{2}\sum_{k=1}^{m}(x_i - \mu)^2 - \frac{1}{2}\sum_{k=m+1}^{n} \mathbb{E}_{\mu^{(i)},0}\Big((Y_i - \mu)^2\Big).$$

M-step:

$$\begin{aligned}\frac{\partial Q(\mu, \mu^{(i)})}{\partial \mu} &= \sum_{k=1}^{m}(x_i - \mu) - \frac{1}{2}\sum_{k=m+1}^{n} \mathbb{E}_{\mu^{(i)},0}\Big(\frac{\partial (Y_k - \mu)^2}{\partial \mu}\Big) \\ &= \Big(\sum_{k=1}^{m} x_i\Big) - m\mu + \sum_{k=m+1}^{n} \mathbb{E}_{\mu^{(i)},0}\Big(Y_k - \mu\Big) \\ &= \Big(\sum_{k=1}^{m} x_i\Big) - m\mu + (n-m)\Big(\mathbb{E}_{\mu^{(i)},0}\Big(Y_n\Big) - \mu\Big).\end{aligned}$$

The latter expression is zero if

$$\mu \quad = \quad \frac{1}{n}\Big(\sum_{k=1}^{m} x_i + (n-m)\mathbb{E}_{\mu^{(i)},0}\Big(Y_n\Big)\Big),$$

so we set

$$\mu^{(i+1)} \quad = \quad \frac{1}{n}\Big(\sum_{k=1}^{m} x_i + (n-m)\Big(\mu^{(i)} - \frac{\phi(-\mu^{(i)})}{\Phi(-\mu^{(i)})}\Big)\Big)$$

as

$$\mathbb{E}_{\mu^{(i)},0}\Big(Y_n\Big) \quad = \quad \mu^{(i)} - \frac{\phi(-\mu^{(i)})}{\Phi(-\mu^{(i)})}.$$

Lemma 9 *In each iteration the EM-algorithm increases the likelihood $L(\theta|x)$, that is $L(\theta^{(i+1)}|x) \geq L(\theta^{(i)}|x)$.*

It follows that, under mild regularity conditions, the EM-algorithm converges to a local maximum of the likelihood function. To prove the above lemma we recall Jensen's inequality which implies the following result:

Lemma 10

$$\mathbb{E}_g\Big(\log\Big(\frac{f(X)}{g(X)}\Big)\Big) \quad \leq \quad 0.$$

Proof of Lemma 9:
As $\theta^{(i+1)} = \text{argmax}\, Q(\theta, \theta^{(i)})$ it follows that

$$Q(\theta^{(i+1)}, \theta^{(i)}) \quad \geq \quad Q(\theta^{(i)}, \theta^{(i)})$$

Moreover,

$$H(\theta, \theta^{(i)}) \quad = \quad \mathbb{E}_{Y|x,\theta^{(i)}}\Big(\log(p(y|x,\theta))\Big) \quad = \quad \int \log(p(y|x,\theta))p(y|x,\theta^{(i)})dy.$$

and so by the above lemma

$$H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)}) = \int \log\Big(\frac{p(y|x,\theta^{(i+1)})}{p(y|x,\theta^{(i)})}\Big)p(y|x,\theta^{(i)})dy \quad \leq \quad 0.$$

Hence

$$\begin{aligned} l(\theta^{(i+1)}|x) - l(\theta^{(i)}|x) \quad &= \quad Q(\theta^{(i+1)}, \theta^{(i)}) - Q(\theta^{(i)}, \theta^{(i)}) \\ &- \quad \Big(H(\theta^{(i+1)}, \theta^{(i)}) - H(\theta^{(i)}, \theta^{(i)})\Big) \quad \geq \quad 0. \end{aligned}$$

Example 29 Normal mixture distribution:
Consider data $x = \{x_1, \dots, x_n\}$ which is assumed to be independent and identically distributed according to the mixture density

$$p(\cdot|\theta) \quad = \quad \sum_{j=1}^{m} \alpha_j \; p_j(\cdot|\mu_j)$$

where $p_j(\cdot|\mu_j)$ is a Normal density with mean μ_j and variance 1, $\sum_{j=1}^m \alpha_j = 1$ and $0 < \alpha_j < 1$ for all $j \in \{1, \dots, m\}$. Let $\theta = (\alpha_1, \dots, \alpha_m, \gamma_1, \dots \gamma_m)$. Then the incomplete data log-likelihood is given by

$$\log(L(\theta|x) \quad = \quad \sum_{k=1}^{n} \log\Big(\sum_{j=1}^{m} \alpha_j p_j(x_k|\mu_j)\Big).$$

We introduce auxiliary (or latent variables) y such that for $k \in \{1, \dots, n\}$ and $j \in \{1, \dots M\}$

$$y_{kj} \quad = \quad \begin{cases} 1 & \text{if } x_k \sim p_j(\cdot|\mu_j) \\ 0 & \text{otherwise} \end{cases}$$

The complete data log-likelihood is now given by

$$l(\theta|x,y) \quad = \quad l(\theta|x,y) \quad = \quad \sum_{k=1}^{n}\sum_{j=1}^{m} y_{kj} \log\Big(\alpha_j p_j(x_k|\mu_j)\Big).$$

For the E-step set

$$t_{kj}^{(i)} \quad = \quad \mathbb{P}\Big(Y_{kj} = 1|x, \theta^{(i)}\Big) \quad = \quad \frac{\alpha_j^{(i)} p_j(x_k|\mu_j^{(i)})}{\sum_{l=1}^m \alpha_l^{(i)} p_l(x_k|\mu_l^{(i)})}$$

and note that

$$\sum_{j=1}^{m} t_{kj}^{(i)} \quad = \quad 1.$$

Hence,

$$\begin{aligned} Q(\theta,\theta^{(i)}) &= \sum_{k=1}^{n}\sum_{j=1}^{m}\mathbb{E}_{Y|x,\theta^{(i)}}(Y_{kj})\log\Big(\alpha_j p_j(x_k|\mu_j)\Big) \\ &= \sum_{k=1}^{n}\sum_{j=1}^{m} t_{kj}^{(i)}\log(\alpha_j) \quad + \quad \sum_{k=1}^{n}\sum_{j=1}^{m} t_{kj}^{(i)} p_j(x_k|\mu_j) \\ &= A + B \end{aligned}$$

Now consider the M-step in which we determine $\theta^{(i+1)} = \mathrm{argmax}_\theta\, Q(\theta,\theta^{(i)})$*. In the above equation we can maximise* A *and* B *seperately. To maximise* A *use a Lagrange multiplier* λ *to impose the constraint* $\sum_{j=1}^{m}\alpha_j = 1$*.We need to determine the roots of*

$$\frac{\partial}{\partial\alpha_j}\Big[\sum_{j=1}^{m}\sum_{k=1}^{n} t_{kj}^{(i)}\log(\alpha_j) - \lambda\Big(\sum_{j=1}^{m}\alpha_j - 1\Big)\Big].$$

We deduce that

$$\alpha_j^{(i+1)} \quad = \quad \frac{1}{n}\sum_{k=1}^{n} t_{kj}^{(i)}.$$

Next we need to maximise

$$B \quad = \quad \sum_{k=1}^{n}\sum_{j=1}^{m} t_{kj}^{(i)} p_j(x_k|\mu_j)$$

with respect to $\mu_1,\ldots,\mu_m$*. We have*

$$\frac{\partial}{\partial\mu_j}\Big(\sum_{k=1}^{n}\sum_{j=1}^{m} t_{kj}^{(i)}\log(p_j(x_k|\mu_j))\Big) \quad = \quad \sum_{k=1}^{n} t_{kj}^{(i)}(x_k-\mu_j)$$

and so

$$\mu_j^{(i+1)} \quad = \quad \frac{1}{n}\sum_{k=1}^{n} t_{kj}^{(i)} x_k.$$

10.2 EM Standard Errors

From statistical theory we know that under mild regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is asymptotically Normal with mean θ and variance-covariance matrix $I^{-1}(\theta)$ where $I(\theta)$ is the Fisher information about θ defined as

$$I(\theta) \quad = \quad \mathbb{E}_X\Big(-\frac{\partial^2 l(\theta|x)}{\partial\theta^2}\Big) \quad = \quad \mathbb{E}_X\Big(\Big[-\frac{\partial^2 l(\theta|x)}{\partial\theta}\Big]^2\Big).$$

In a missing data set-up we can exploit the following equality:

$$\begin{aligned} l(\theta|x) &= Q(\theta,\theta^{(i)}) - H(\theta,\theta^{(i)}) \quad \text{and so} \\ \frac{\partial^2 l(\theta|x)}{\partial\theta^2} &= \frac{\partial^2 Q(\theta,\theta^{(i)})}{\partial\theta^2}_{|_{\theta^{(i)}=\theta}} - \frac{\partial^2 H(\theta,\theta^{(i)})}{\partial\theta^2}_{|_{\theta^{(i)}=\theta}} \end{aligned}$$

Now note that

$$\begin{aligned} \frac{\partial H(\theta,\theta^{(i)})}{\partial\theta}_{|_{\theta^{(i)}=\theta}} &= \frac{\partial}{\partial\theta}\mathbb{E}_{Y|x,\theta^{(i)}}\Big(\log(p(y|x,\theta))\Big)_{|_{\theta^{(i)}=\theta}} \\ &= 0 \quad \text{and so} \\ \frac{\partial^2 H(\theta,\theta^{(i)})}{\partial\theta^2}_{|_{\theta^{(i)}=\theta}} &= \frac{\partial^2}{\partial\theta^2}\mathbb{E}_{Y|x,\theta^{(i)}}\Big(\log(p(y|x,\theta))\Big)_{|_{\theta^{(i)}=\theta}} \\ &= -\mathbb{E}_{Y|x,\theta^{(i)}}\Big(\Big[\frac{\partial}{\partial\theta}\log(p(y|x,\theta))\Big]^2\Big)_{|_{\theta^{(i)}=\theta}} \\ &= -\mathrm{Var}_{Y|x,\theta^{(i)}}\Big(\frac{\partial}{\partial\theta}\log(p(y|x,\theta))\Big)_{|_{\theta^{(i)}=\theta}} \end{aligned}$$

It follows that

$$\frac{\partial^2 l(\theta|x)}{\partial\theta^2} = \frac{\partial^2 Q(\theta,\theta^{(i)})}{\partial\theta^2}_{|_{\theta^{(i)}=\theta}} + \mathrm{Var}_{Y|x,\theta^{(i)}}\Big(\frac{\partial}{\partial\theta}\log(p(y|x,\theta))\Big)_{|_{\theta^{(i)}=\theta}}$$

Example 30 Censored Normal Distribution:

$$\begin{aligned} Q(\mu,\mu^{(k)}) &= -\frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{1}{2}\sum_{i=m+1}^{n}\mathbb{E}_{\mu^{(k)},0}\Big((Y_i-\mu)^2\Big) \\ \frac{\partial^2 Q(\mu,\mu^{(k)})}{\partial\mu^2} &= -m-(n-m) = -n. \end{aligned}$$

Moreover,

$$\begin{aligned} -\frac{\partial^2 H(\mu,\mu^{(k)})}{\partial\mu^2} &= \mathrm{Var}_{Y|x,\mu^{(k)}}\Big(\frac{\partial\log(p(Y|x,\mu))}{\partial\mu}\Big)_{|_{\mu^{(k)}=\mu}} \\ \log(p(y|x,\mu)) &= \log\Big(\prod_{i=m+1}^{n}\phi(y_i-\mu)\Big) - (n-m)\log(\Phi(-\mu)) \\ &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\Big(\sum_{i=m+1}^{n}(y_i-\mu)^2\Big) - (n-m)\log(\Phi(-\mu)), \end{aligned}$$

and so

$$\begin{aligned}
\frac{\partial \log(p(y|x,\mu))}{\partial \mu} &= \sum_{i=m+1}^{n} (y_i - \mu) - (n-m)\frac{\partial \log(\Phi(-\mu))}{\partial \mu}, \qquad \textit{hence} \\
\mathrm{Var}_{Y|x,\mu^{(k)}}\Big(\frac{\partial \log(p(Y|x,\mu))}{\partial \mu}\Big) &= \sum_{i=m+1}^{n} \mathrm{Var}_{Y|x,\mu^{(k)}}(Y) \\
&= (n-m)\Big[1 - \frac{\phi(-\mu)}{\Phi(-\mu)}\Big(\frac{\phi(-\mu)}{\Phi(-\mu)} + \mu\Big)\Big].
\end{aligned}$$

It follows that the variance of the MLE $\hat{\theta}$ is approximately equal to

$$\frac{1}{n-(n-m)[1-\frac{\phi(-\mu)}{\Phi(-\mu)}(\frac{\phi(-\mu)}{\Phi(-\mu)}+\mu)]}.$$

10.3 Extensions of the EM algorithm

1. Monte Carlo EM:
   This may be used if the direct evaluation of Q and thus the E-step is computationally difficult. We use Monte Carlo estimation to approximate Q. At iteration i the E-step is thus replaced by sampling $y_1, \ldots, y_n$ iid from $p(y|x, \theta^{(i-1)})$ and setting

   $$Q(\theta, \theta^{(i-1)}) = \frac{1}{m}\sum_{j=1}^{n} \log(L(\theta|y_j, x)).$$

2. Generalized EM:
   Here instead of maximising Q we simply find a $\theta^{(i+1)}$ such that

   $$Q(\theta^{(i+1)}, \theta^{(i)}) > Q(\theta^{(i)}, \theta^{(i)}).$$