AALBORG UNIVERSITY

A short diversion into the theory of Markov chains, with a view to Markov chain Monte Carlo methods

by

Kasper K. Berthelsen and Jesper Møller

June 2004

2004-01

DEPARTMENT OF MATHEMATICAL SCIENCES
AALBORG UNIVERSITY
Fredrik Bajers Vej 7 G ▪ DK - 9220 Aalborg Øst ▪ Denmark
Phone: +45 96 35 80 80 ▪ Telefax: +45 98 15 81 29
URL: http://www.math.aau.dk

A short diversion into the theory of Markov chains, with a view to Markov chain Monte Carlo methods

1 Introduction

These lecture notes was written for a PhD-course, “Stochastic simulation and example of applications” held at Aalborg University, May-June 2004. It concerns Markov chain Monte Carlo (MCMC) methods, particularly the Metropolis-Hastings algorithm and special cases of this (Metropolis algorithm, Gibbs sampling, Metropolis within Gibbs). Various exercises and examples of applications will be presented throughout the text. Particularly, examples of Bayesian MCMC analysis will be discussed.

We start with a short diversion into the theory of Markov chains. So far we have for simplicity considered Markov chains defined on a finite state space, but for many statistical problems we need a more complicated state space Ω. For the purpose of simulation from a given target density π, we clearly need to have that $\Omega \supseteq \{x : \pi(x) > 0\}$ (often $\Omega = \{x : \pi(x) > 0\}$). For specificity we consider in the following the case where $\Omega \subseteq \mathbb{R}^d$ is the state space for a d-dimensional continuous random vector (the reader may easily modify things to the discrete case by replacing integrals by sums). As we shall see concepts like e.g. irreducibility and convergence of Markov chains become slightly more technical, but we shall try to keep technicalities at a minimum.

The exposition will be much directed by what is needed of Markov chain theory (Sections 2–5) in order to understand MCMC methods (Sections 6–11). We skip most proofs in this tutorial (proofs can be found in the references given in Section 12).

2 Examples and basic definitions of Markov chains

2.1 Introductory example and exercise

Example 1 In this example we consider the constructions of two Markov chains with state space $\Omega = \mathbb{R}$. For $x, y \in \mathbb{R}$, we define

$$q(x,y) = \exp\left(-(y-x)^2/2\right)/\sqrt{2\pi}.$$

Then $y \mapsto q(x,y)$ is the density of $\mathrm{N}(x,1)$.

In the first construction we obtain a Markov chain $(X_0, X_1, \ldots)$ by setting

$$X_0 = R_0, \qquad X_{n+1} = X_n + R_{n+1}, \qquad n = 0, 1, \ldots,$$

where $R_0, R_1, \ldots$ are i.i.d. standard normal distributed. This Markov chain is called a *random walk*. To see that it is indeed a Markov chain (a formal definition of the concept “Markov chain” is given in Definition 1 below), we observe that for any $x_0, \ldots, x_n \in \mathbb{R}$ and any $A \subseteq \mathbb{R}$,

$$\begin{aligned}
&P(X_{n+1} \in A | X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) \\
=&P(x_n + R_{n+1} \in A | R_0 = x_0, R_1 = x_1 - x_0, \ldots, R_n = x_n - x_{n-1}) \\
=&P(x_n + R_{n+1} \in A) && \text{(since } R_0, \ldots, R_{n+1} \text{ are independent)} \\
=&\int \mathbf{1}[x_n + r \in A] \frac{1}{\sqrt{2\pi}} \exp\left(-r^2/2\right) dr && \text{(because } R_{n+1} \text{ is N}(0,1)\text{-distributed)} \\
=&\int_A \frac{1}{\sqrt{2\pi}} \exp\left(-(y - x_n)^2/2\right) dy \\
=&\int_A q(x_n, y) dy
\end{aligned}$$

which only depends on x_n. Instead of these calculations we could simply argue that if $X_n = x_n$, then $X_{n+1} = x_n + R_{n+1}$ is the same no matter the values of $X_0, \ldots, X_{n-1}$[1]. Consequently, neither which argument we use, the conditional distribution of X_{n+1} given $X_0, \ldots, X_n$ is seen to be the same as the conditional distribution of X_{n+1} given X_n. Thus it is a Markov chain. We have also shown that the conditional distribution of X_{n+1} given $X_n = x$ is a $\mathrm{N}(x, 1)$-distribution.

In the second construction we let $\pi(x)$ denote an arbitrary density function which is strictly positive on $\mathbb{R}$, and define

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

Furthermore, we let $U_1, U_2, \ldots$ be i.i.d. unif(0,1)-distributed (and independent of $R_0, R_1, \ldots$ as defined above). Then we obtain a new Markov chain $(X_0, X_1, \ldots)$ by $X_0 = R_0$ and

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{if } U_{n+1} \le a(X_n, Y_{n+1}) \\ X_n & \text{otherwise} \end{cases}$$

for $n = 0, 1, \ldots$, where Y_{n+1} is defined by $Y_{n+1} = X_n + R_{n+1}$. We can think of Y_{n+1} as a proposal for X_{n+1} at the $(n + 1)$th update. By construction we accept this proposal with probability $a(X_n, Y_{n+1})$. Since $U_{n+1} \le 1$, we can replace the condition for accepting the proposal by $U_{n+1} \le \pi(Y_{n+1})/\pi(X_n)$. Note that we always accept the proposal Y_{n+1} if it is

[1] When we make a computer code for generating a Markov chain it is formally given as $X_{n+1} = \varphi(X_n, V_{n+1})$. Here φ is a deterministic function called the updating function, and $V_1, V_2, \ldots$ are i.i.d. random variables (or vectors). For the random walk we have $\varphi(x, y) = x + y$ and $V_n = R_n$. Conversely, it is easily seen that any stochastic recursive sequence given by $X_{n+1} = \varphi(X_n, V_{n+1})$ is a Markov chain.

more likely than X_n (i.e. $\pi(Y_{n+1}) \geq \pi(X_n)$ implies $X_{n+1} = Y_{n+1}$), while if it is less likely there is still a chance of accepting the proposal: the chance is given by $\pi(Y_{n+1})/\pi(X_n)$.

The latter construction is an example of the so-called *Metropolis random walk algorithm* (Section 7). Note that if we redefine $a(x, y) = 1$, then the first case is formally a particular case of the latter case (we write "formally" because there exists no density $\pi(x)$ so that both (1) and $a(x, y) = 1$ hold). In the sequel we therefore refer to the first case as the case $a(x, y) = 1$.

Exercise 1

1. Argue in detail why the second construction in Example 1 specifies a Markov chain.
2. For simplicity let π be the density of $\mathrm{N}(0, 1)$. In order to simulate $X_1, \ldots, X_n$ for the two cases considered in Example 1 when $X_0 = x$ is the initial state, write R-functions `random.walk(n,x)` and `Metropolis.random.walk(n,x)`.
3. Let $X_0 = 0$ and simulate some sample paths (i.e. realisations of $X_1, \ldots, X_n$ in the two cases) and compare the results.

2.2 General setting

The setting considered below is motivated by how simulation algorithms (so-called Metropolis-Hastings algorithms) are constructed later on. So far the reader is recommended to relate the following to the second case in Example 1.

For any $x \in \Omega$, suppose that $y \mapsto q(x, y)$ is a density function on Ω and define a probability measure by

$$Q(x, A) = \int_A q(x, y) dy, \qquad A \subseteq \Omega.$$ [2]

We call $q(x, y)$ a *proposal density* and $Q(x, A)$ a *proposal distribution*. Furthermore, suppose that $a(x, y)$ is a number between 0 and 1 (Example 1 shows examples of such functions $a(x, y)$, and we shall later consider many other examples). Given x we think of $a(x, y)$ as a probability for accepting a proposal y drawn from the density $q(x, y)$. In other words, we think of $a(x, y)$ as a *probability of accepting* a *proposal* $Y = y$ drawn from the distribution $Q(x, A)$. We do not accept the proposal with probability

$$r(x) = 1 - \int a(x, y) q(x, y) dy;$$

[2] Readers familiar with measure theory will more carefully read this as "for all measurable $A \subseteq \Omega$". However, practically all subsets of $\mathbb{R}^d$ are measurable (this is the class of Borel sets), so we do not worry about such details in this course.

in that case we will retain x. Thus, for fixed x,

$$P(x, A) = r(x)\mathbf{1}[x \in A] + \int q(x, y)a(x, y)\mathbf{1}[y \in A]dy, \qquad A \subseteq \Omega, \tag{2}$$

(where $\mathbf{1}[x \in A]$ denotes the indicator function which is 1 if $x \in A$ and 0 if $x \notin A$) is the probability measure describing what happens when we either accept the proposal y or retain x. Informally we can also write this as

$$P(x, A) = P(Y \text{ is not accepted}|x)\mathbf{1}[x \in A] + \int_A q(x, y)P(Y = y \text{ is accepted}|x)dy.$$

Finally, we call $P(Y \text{ is not accepted}|x) = P(x, \{x\}) = r(x)$ the *rejection probability*.

Why this becomes a very useful strategy for making simulations is probably so far a mystery for most readers. Before we can understand this we need to establish some theory as outlined in the sequel.

Example 2 In the first case in Example 1 where $a(x, y) = 1$, we have that $X_n = R_0 + \cdots + R_n$ where the R_i are i.i.d. and N(0,1)-distributed. Hence the conditional distribution of X_n given $X_0 = x$ is N(x, n). So the variance tends to infinity as $n \to \infty$ (did you observe that when solving Exercise 1?). However, what is the (conditional or marginal) distribution of X_n in the second case of Example 1? The answer will be given later in Theorem 2 (Section 4).

2.3 Some basic definitions

Definition 1 A random process $(X_0, X_1, \ldots)$ with state space Ω is said to be a *(homogeneous) Markov chain* with *transition kernel* P if for all integers $n \geq 0$, all $A \subseteq \Omega$, and all $x_0, \ldots, x_n \in \Omega$ we have

$$P(X_{n+1} \in A|X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_{n+1} \in A|X_n = x_n) = P(x_n, A).$$

In other words, the conditional distribution of X_{n+1} given $X_0, \ldots, X_n$ is identical to the conditional distribution of X_{n+1} given X_n.

Definition 2 The *initial distribution* of a Markov chain $(X_0, X_1, \ldots)$ is the distribution of X_0. Furthermore,

$$P^n(x, A) = P(X_n \in A|X_0 = x), \qquad A \subseteq \Omega,$$

denotes the conditional distribution of X_n given $X_0 = x$; this is called the *n-step transition kernel*.

It can be shown that the n-step transition kernel can be expressed in terms of the (1 step) transition kernel, but in applications the expression is often complicated (in the random walk example it was simple enough in the first case, but it is not in the case (1) unless we let the initial distribution be given by π as noticed in Exercise 2 below). One important point here is that *the distribution of a Markov chain can be shown to be completely specified by its initial distribution and its transition kernel.*

3 Invariant distributions and reversibility

In the sequel we consider a Markov chain $(X_0, X_1, \ldots)$ with transition kernel P and unless otherwise stated an arbitrary initial distribution.

For the purpose of simulation from a given target density π defined on Ω, we want to construct the chain such that π becomes an invariant density (see Definition 3 below). As we shall see later, such a chain can be constructed in many ways, but it needs at least to be irreducible (see Definition 4 and Theorem 1 below). We let

$$\Pi(A) = \int_A \pi(x)dx, \qquad A \subseteq \Omega,$$

denote the *target distribution*.

Definition 3 A Markov chain with transition kernel P has π as its *invariant density*[3] if for all $A \subseteq \Omega$,

$$\int \pi(x)P(x, A)dx = \Pi(A).$$

Moreover, the chain is *reversible* if (X_0, X_1) and (X_1, X_0) are identically distributed when $X_0 \sim \pi$.

We say that the chain satisfies the *detailed balance condition (DBC)* if for all different states $x, y \in \Omega$,

$$\pi(x)p(x, y) = \pi(y)p(y, x) \tag{3}$$

where we define

$$p(x, y) = a(x, y)q(x, y).$$

Note that by (2),

$$P(x, A) = \int_A p(x, y)dy \qquad \text{if } x \notin A.$$

[3] Also called a *stationary density* or an *equilibrium density.*

It can be shown that the DBC implies both that π is an invariant density, the chain is reversible, and $(X_0, \ldots, X_n)$ and $(X_n, \ldots, X_0)$ are identically distributed when $X_0 \sim \pi$ (see Exercise 2).

Exercise 2

1. Verify that if X_n is distributed according to an invariant distribution Π, then X_m has distribution Π for $m = n + 1, n + 2, \ldots$.
   Hint: *Argue that by induction it suffices to consider the case* $m = n + 1$.

2. For simplicity consider the discrete case and show that (3) implies that
   a) π is an invariant density for the chain,
   b) the chain is reversible,
   c) and for any integer $n \geq 1$, $(X_0, \ldots, X_n)$ and $(X_n, \ldots, X_0)$ are identically distributed when X_0 is distributed in accordance with the invariant density.
   Hint: *a) Argue that it suffices to verify that* $\pi(x) = \sum_y \pi(y)p(y, x)$.

3. The DBC will be satisfied for many Markov chains used for simulation, including the Metropolis-Hastings in Section 6. However, consider the random walk example (Example 1), and show that
   a) it does not satisfy the DBC in the first case $a(x, y) = 1$ [4]
   b) but it does satisfy the DBC in the second case (1).
   Hint: *a) If* $\pi(x)$ *is constant for all* $x \in \mathbb{R}$, *then* π *is not a well-defined density (why?).*

4 Irreducibility and asymptotic results

Definition 4 Suppose a Markov chain has Π as its invariant distribution. The chain is then *irreducible* if for all $x \in \Omega$ and $A \subseteq \Omega$ with $\Pi(A) > 0$, there exists an n such that $P^n(x, A) > 0$; in other words the chain can get to any region A with $\Pi(A) > 0$. To stress the role of Π (or π) we also say that the chain is Π-*irreducible* (or π-irreducible). Furthermore, the chain is *Harris recurrent*[5] if

$$P(X_n \in A \text{ for infinite many } n \,|\, X_0 = x) = 1.$$

It becomes useful to notice that in the continuous case as considered in this tutorial, $\Pi(A) >$

[4] In fact even an invariant density does not exist in this case (you are not asked to verify that right now, but you are welcome to discuss why it cannot be the case).

[5] In the discrete state space case, Harris recurrence is identical to irreducibility.

0 is equivalent to

$$\int \mathbf{1}[x \in A,\, \pi(x) > 0]dx > 0. \tag{4}$$

It can be shown that irreducibility implies uniqueness of the invariant distribution[6]. The concept of Harris recurrent is stronger than irreducibility, and it is used in the following result.

Theorem 1 (The strong law of large numbers for Markov chains) Let $(X_0, X_1, \ldots)$ be a π-irreducible Markov chain where π is an invariant density, and let $h : \Omega \to \mathbb{R}$ be a function such that the mean $\theta = \int h(x)\pi(x)dx$ exists. For an arbitrary integer $m \geq 0$, define the *ergodic average*[7] by

$$\hat{\theta}_n = \frac{1}{n+1} \sum_{i=m}^{m+n} h(X_i).$$

Then there exists a set $C \subseteq \Omega$ such that $\Pi(C) = 1$ and for all $x \in C$,

$$P(\hat{\theta}_n \to \theta \text{ as } n \to \infty \,|\, X_0 = x) = 1.$$

Furthermore, if the chain is Harris recurrent, we can take $C = \Omega$.

Thus irreducibility implies consistency of the estimator $\hat{\theta}_n$ for all initial states $x \in C$; Harris recurrence ensures consistency for all initial states $x \in \Omega$, and so we do not need to worry about what happens if $x \notin C$. In this course we shall mainly focus on verifying irreducibility, since establishing Harris recurrence can be rather technical.

A natural question is if the choice of m in Theorem 1 is of any relevance in practice. We investigate this in the following Exercise 3 and comment further on this after the exercise.

Exercise 3 Consider an irreducible Markov chain with invariant distribution Π.

1. Show that $\hat{\theta}_n$ is an unbiased estimator if X_m is drawn from Π.
2. a) Is it unbiased if it is not drawn from π? b) Does it matter?
   Hint: *b) Use* `Metropolis.random.walk(n,x)` *from Exercise 1 when e.g.* $h(x) = x$ *is the identity mapping, and try with different values of* `(n,x)`.
3. Show that in the case (1) of the random walk Metropolis algorithm in Example 1, the chain is π-irreducible.
   Hint: *Recall (4).*

[6] More precisely, we mean uniqueness of the invariant distribution up to so-called null sets, but again we do not worry about such details, which have no practical importance.

[7] Also called the *empirical average*.

In practice we choose m as the so-called *burn-in*, that is a time at which the chain is considered to be *effectively in equilibrium* (i.e. m is large enough so that the distribution of X_m is in very close accordance with π; we shall later return to this issue in greater detail). We therefore now turn to the question when a Markov chain has a limiting distribution, and what it is. First we need the following definition.

Definition 5 A Π-irreducible Markov chain is *periodic* if there is a partition of the state space $\Omega = A_0 \cup A_1 \cup \cdots \cup A_{n-1} \cup A_n$ into $n+1 \geq 3$ disjoint sets $A_0, \ldots, A_n$ such that $\Pi(A_n) = 0$ and

$$x \in A_0 \Rightarrow P(x, A_1) = 1,\ x \in A_1 \Rightarrow P(x, A_2) = 1,\ \ldots,$$
$$x \in A_{n-2} \Rightarrow P(x, A_{n-1}) = 1,\ x \in A_{n-1} \Rightarrow P(x, A_0) = 1.$$

Otherwise the chain is said to be *aperiodic*.

Note that aperiodicity is not needed for the strong law of large numbers for Markov chains (Theorem 1). However, aperiodicity is needed in the following theorem.

Theorem 2 (The Markov chain convergence theorem) For a Π-irreducible and aperiodic Markov chain, where Π is the invariant distribution, there exists a set $C \subseteq \Omega$ such that $\Pi(C) = 1$ and for all $x \in C$ and $A \subseteq \Omega$,

$$P(X_n \in A \,|\, X_0 = x) \rightarrow \Pi(A) \qquad \text{as } n \rightarrow \infty.^{8}$$

If it is also Harris recurrent[9], then we can take $C = \Omega$.

Exercise 4 Consider again Example 1.

1. Does there exist a limiting distribution for the chain in the case $a(x, y) = 1$?
   Hint: *Recall Example 2.*

2. Argue why the chain in the case (1) is aperiodic. What is the limiting distribution?
   Hint: *Recall 3.b) in Exercise 2.*

Finally, we notice that a *central limit theorem* also hold for Markov chains satisfying certain conditions (including irreducibility and aperiodicity). For example, the Metropolis random walk algorithm satisfies these conditions, but since the conditions are a bit technical, we shall not discuss these further here.

[8] In fact we then have convergence with respect to the total variation norm.

[9] Harris recurrence and aperiodicity are not only sufficient but also necessary conditions for the Markov chain convergence theorem to hold for all initial states $x \in \Omega$.

Exercise 5 Consider again Example 1, fix the initial state to $X_0 = 0$, and let

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

be the ergodic average based on a sample of length $n \geq 2$.

1. Show that if $a(x, y) = 1$ then $\bar{X}_n \sim \mathrm{N}(0, n(n+1)(2n+1)/(6n^2))$. What does this mean as $n \to \infty$?
   Hint: *First show that*

   $$\bar{X}_n = \frac{1}{n}(nR_1 + (n-1)R_2 + (n-2)R_3 + \cdots + R_n)$$

   and next use that $1^2 + 2^2 + \cdots + n^2 = n(n+1)(2n+1)/6$.

2. Consider the case (1) with $\Pi = \mathrm{N}(0, 1)$.
   a) In each of the cases $n = 1000, 2000, \ldots, 10000$ simulate 10 i.i.d. realisations of $\bar{X}_n$ and estimate then the variance of $\bar{X}_n$.
   b) Does the variance seem to be a decreasing function of order $1/n$?
   c) Simulate 100 i.i.d. realisations of $\bar{X}_n$ when $n = 1000$, and discuss if $\sqrt{n}\bar{X}_n$ is (approximately) $\mathrm{N}(0, 1)$-distributed.
   Hint: *b) Plot* n *times the estimated variance versus* $n/1000$.
   c) Consider histograms and q-q plots.

5 Importance sampling based on Markov chains

We have earlier introduced importance sampling in the simple setting of an i.i.d. sample. This easily extends to Markov chains as follows.

Let still π denote our target density and suppose we want to estimate the mean $\theta = \int h(x)\pi(x)dx$. Assume that g is another density such that

$$\pi(x) > 0 \;\Rightarrow g(x) > 0$$

and we have constructed a g-irreducible and aperiodic Markov chain $Y_0, Y_1, \ldots$ with invariant density g. Let $m \geq 0$ denote the burn-in of the chain. By Theorem 1, if $Y_0 \in C$ where $\int_C g(x)dx = 1$, then

$$\tilde{\theta}_n = \frac{1}{n+1}\sum_{i=m}^{m+n} h(Y_i)\frac{\pi(Y_i)}{g(Y_i)}$$

is a consistent estimator[10] of θ.

As in the i.i.d. case, we call g the *instrumental or importance sampling density*, and

$$w(Y_i) = \frac{\pi(Y_i)}{g(Y_i)}, \qquad i = m, \ldots, m+n,$$

the *importance weights*. As before, the variation of the importance weights should not be too large.

Exercise 6 As an illustrative example, let us consider the situation where the importance sampling distribution is $\mathrm{N}(0,1)$ and the target distribution is $\mathrm{N}(\theta,1)$, where $\theta \geq 0$ is a parameter.

1. Show that the importance weight based on a realisation y is given by

$$w(y) = \exp\left(\theta y - \theta^2/2\right).$$

2. Simulate what happens with the importance weights as $\theta = 0, 1, 2, 3$ increases when we use

   (a) 1000 i.i.d. samples from $\mathrm{N}(0,1)$,

   (b) 1000 samples from the Metropolis random walk chain when this is started in equilibrium, i.e. when the burn-in is $m = 0$ and Y_0 is drawn from $\mathrm{N}(0,1)$.

6 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm (due to Metropolis et al., 1953, and Hastings, 1970) provides very general constructions of Markov chains with an arbitrary equilibrium density. The Metropolis random walk algorithm in Example 1 is just one example of a specific Metropolis-Hastings algorithm.

Consider the setting at the beginning of Section 2: given a target density π and a proposal density $q(x,y)$ how do we specify the acceptance probability $a(x,y)$? Recall the DBC given by (3):

$$\pi(x)q(x,y)a(x,y) = \pi(y)q(y,x)a(y,x). \tag{5}$$

If $\pi(x)q(x,y) > 0$, we can rewrite this as

$$a(x,y) = H(x,y)a(y,x).$$

[10] Strictly speaking aperiodicity is not needed here. Using an appropriate burn-in is usually a good idea as it may drastically reduce the variance of the estimator $\tilde{\theta}_n$.

where

$$H(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$$

is the so-called *Hastings ratio*. It turns out that it is not so important what $H(x,y)$ is if $\pi(x)q(x,y) = 0$, but for specificity let us set

$$H(x,y) = \infty \qquad \text{whenever } \pi(x)q(x,y) = 0.$$

If we set

$$a(x,y) = \min\{1, H(x,y)\} \tag{6}$$

then (5) is satisfied (Exercise 7). It can be shown that if (5) should be satisfied, then the highest acceptance probabilities are obtained by the choice (6). The Metropolis-Hastings algorithm uses this choice and can be described as follows.

Metropolis-Hastings algorithm Let the initial state $X_0 = x$ be such that $\pi(x) > 0$[11]. For $n = 0, 1, \ldots$, given X_n we

- generate $U_{n+1} \sim \text{unif}(0,1)$ and Y_{n+1} from the density $y \mapsto q(X_n, y)$ (where U_{n+1} and Y_{n+1} are independent given X_n),
- and then set

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{if } U_{n+1} \le H(X_n, Y_{n+1}) \\ X_n & \text{otherwise} \end{cases}$$

An important observation: Note that the Metropolis-Hastings algorithm only depends on π through the ratio $\pi(Y_{n+1})/\pi(X_n)$ from the Hastings ratio, so *in the Metropolis-Hastings algorithm we need only to know π up to proportionality* (because any constant factor will cancel in the ratio $\pi(Y_{n+1})/\pi(X_n)$). Thus, when we later consider examples of posterior densities as our target density π, *we need only to specify the posterior density up to proportionality*.

Theorem 3 By construction the Metropolis-Hastings algorithm is reversible with invariant density π.

Other properties (irreducibility etc.) need to established under further assumptions. For instance, if

$$q(x,y) > 0 \qquad \text{for all } x, y \in \Omega$$

[11] In most situations it is natural to require that $\pi(X_0) > 0$, because then (with probability one) $\pi(X_n) > 0$ for all $n \ge 0$. However, in some rare occasions it is advantageous to modify the Metropolis-Hastings algorithm so that $\pi(X_0) = 0$ is allowed: Then by definition, if $X_0 = x$ and if $Y_1 = y$ is proposed, $a(x,y) = 1$, and so $X_1 = Y_1$ is accepted. If further $q(x,y) > 0$ implies $\pi(y) > 0$, then we are certain that $\pi(X_1) > 0$, and so $\pi(X_n) > 0$ for all $n \ge 1$.

then the Metropolis-Hastings chain is π-irreducible. Furthermore, we have Harris recurrence because of the following result.

Theorem 4 If the Metropolis-Hastings chain is π-irreducible, it is Harris recurrent.

Finally, we observe that any π-irreducible Markov chain is aperiodic if the event $\{X_{n+1} = X_n\}$ is possible when $X_n \sim \pi$, i.e.

$$\int P(x, \{x\})\pi(x)dx > 0.$$

For the Metropolis-Hastings algorithm this means that

$$\int\int \mathbf{1}[\pi(y)q(y,x) < \pi(x)q(x,y)]q(x,y)\pi(x)dydx > 0.$$

Exercise 7 Fill in the details, showing that the Metropolis-Hastings algorithm satisfies the DBC.

7 Metropolis algorithm

This is the special case of the Metropolis-Hastings algorithm with a symmetric proposal density:

$$q(x,y) = q(y,x).$$

Then the Hastings ratio reduces to the *Metropolis ratio*

$$H(x,y) = \pi(y)/\pi(x)$$

whenever $q(x,y) > 0$.

For a *Metropolis random walk algorithm,*

$$q(x,y) = f(y-x)$$

where f is a symmetric density function, so $q(x,y) = q(y,x)$. Example 1 provides an example of such an algorithm. Theoretical results show that the Metropolis random walk algorithm "works best" if chosen so that the acceptance probability in average is between 0.2 and 0.4.

Exercise 8 Recall Exercise 5 in the text "Basic methods for simulation of random variables: 1. Inversion" regarding estimating the probability θ for a female birth given a special condition called placenta previa. There we performed a Bayesian analysis letting the data distribution be binomial (see the Example in “A brief introduction to (simulation based) Bayesian inference”) and assuming a non-standard prior density for θ. Posterior inference was done by sampling the posterior distribution of Θ, i.e. the conditional distribution of Θ given the number of female births x of the total number of births n. Specifically this was done by discretising the posterior density and then sampling from this discrete distribution using inversion.

An alternative to the discretisation approach is to apply a Metropolis algorithm for sampling the posterior distribution of θ. In this case the posterior density $\pi(\theta|x)$ of θ plays the role of π in the theory above and θ itself plays the role of x. Furthermore, we let θ' denote the proposal (corresponding to y in the theory above). Assume that if θ is the current state then the proposal θ' is generated from a normal distribution with mean θ and standard deviation $\sigma > 0$, i.e. $q(\theta, \theta') = \exp(-(\theta' - \theta)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$.

1. Implement an R-function for this Metropolis algorithm.
2. Use the R function to produce a sequence $\Theta_0, \Theta_1, \ldots, \Theta_n$ when $\sigma = 0.05$ and $n = 1000$. Summarise the simulation by a histogram and 2.5%, 50% and 97.5% quantiles. Furthermore, compare the histogram to a plot of the posterior density.
   Hint: *The R code used for Exercise 5 mentioned above may be useful.*

8 Gibbs sampler

Suppose that the state space $\Omega \subseteq \mathbb{R}^d$ is a product space

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_k \tag{7}$$

where $\Omega_1 \subseteq \mathbb{R}^{d_1}, \Omega_2 \subseteq \mathbb{R}^{d_2}, \ldots, \Omega_k \subseteq \mathbb{R}^{d_k}$, and $d_1 + d_2 + \cdots + d_k = d$. If $\mathbf{X}$ is a random vector following the target density π, we can write

$$\mathbf{X} = (X_1, X_2, \ldots, X_k) \tag{8}$$

where X_i is the projection of $\mathbf{X}$ on Ω_i, $i = 1, \ldots, k$; we refer to X_i as the ith component. We use bold face for $\mathbf{X}$ in (8) to emphasize that it is a vector and later to avoid confusing an indexed element of the type (8), say $\mathbf{X}_n$, with the component X_n of $\mathbf{X}$. Let

$$X^i = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k)$$

denote $\mathbf{X}$ minus X_i, i.e. X^i has state space $\Omega^i = \Omega_1 \times \cdots \times \Omega_{i-1} \times \Omega_{i+1} \times \cdots \times \Omega_k$. Gibbs sampling consists in simulating from the conditional distributions of X_i given X^i,

$i = 1, \ldots, k$; below we consider first the case of a cyclic updating scheme, i.e. when updating the components in cycles given by the first, the second, ..., the last component; later on we comment on other updating schemes.

For simplicity, let us assume that

$$\pi(x) > 0 \qquad \text{for all } x \in \Omega. \tag{9}$$

The density of X^i is then given by

$$\pi^i(x^i) = \int_{\Omega_i} \pi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k) dy_i, \qquad x^i \in \Omega^i,$$

and the conditional density of X_i given $X^i = x^i$ (for $x^i \in \Omega^i$) is given by

$$\pi_i(x_i|x^i) = \pi(x)/\pi^i(x^i), \qquad x_i \in \Omega_i,$$

where x is specified in accordance with x_i and x^i. So if $P_i(\cdot|x^i)$ denotes the conditional distribution of X_i given $X^i = x^i$, then

$$P_i(A|x^i) = P(X_i \in A|X^i = x^i) = \int_A \pi_i(x_i|x^i) dx_i, \qquad A \subseteq \Omega_i.$$

The densities $\pi_i(x_i|x^i)$ are called the *full conditionals*, and a particular feature of the Gibbs sampler is that these are the only densities used for simulation. Thus, even in high-dimensional problems, all of the simulations may be of low dimension (i.e. all d_i are small), which is usually an advantage.

Example 3 Before giving the details of the Gibbs sampler, let us consider a simple example where $k = d = 2$ and π is the density of a 2-dimensional normal distribution so that

$$EX_1 = EX_2 = 0,\ VarX_1 = VarX_2 = 1,\ Cov(X_1, X_2) = \rho$$

where $-1 < \rho < 1$ is a parameter (the correlation). Then

$$\pi(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x_1^2 + x_2^2 - 2\rho x_1 x_2)\right). \tag{10}$$

It can be shown that $X_1 \sim \mathrm{N}(0,1)$, $X_2 \sim \mathrm{N}(0,1)$, and

$$X_1|X_2 \sim \mathrm{N}(\rho X_2, 1-\rho^2), \qquad X_2|X_1 \sim \mathrm{N}(\rho X_1, 1-\rho^2). \tag{11}$$

For example, if we fix x_2, then

$$\pi(x_1, x_2) \propto \exp\left(-\frac{1}{2(1-\rho^2)}(x_1^2 - 2\rho x_1 x_2)\right) \propto \exp\left(-\frac{1}{2(1-\rho^2)}(x_1 - \rho x_2)^2)\right)$$

which is proportional to the density of $\mathrm{N}(\rho x_2, 1-\rho^2)$, whereby the last result in (11) is verified. Now, the cyclic Gibbs sampler generates

$$\mathbf{X}_n = (X_{1,n}, X_{2,n}), \qquad n = 0, 1, \ldots,$$

by the following:

$$X_{1,n+1} \text{ is drawn from } \mathrm{N}(\rho X_{2,n}, 1-\rho^2),$$
$$X_{2,n+1} \text{ is drawn from } \mathrm{N}(\rho X_{1,n+1}, 1-\rho^2).$$

Note that given $X_{2,n}$ we have that $X_{1,n+1}$ is independent of the "past" $(X_{1,0}, X_{2,0}, \ldots, X_{1,n})$; and given $X_{1,n+1}$ we have that $X_{2,n+1}$ is independent of the "past" $(X_{1,0}, X_{2,0}, \ldots, X_{1,n}, X_{2,n})$. Furthermore, the conditional distributions in (11) are of the same type (meaning that for any real number a, the full conditionals $X_1|X_2 = a$ and $X_2|X_1 = a$ are the same), so both $(X_{1,0}, X_{2,0}, X_{1,1}, X_{2,1}, \ldots)$ and $(\mathbf{X}_0, \mathbf{X}_1, \ldots)$ are (homogeneous) Markov chains. In many other examples of a cyclic Gibbs sampler, the conditional distributions of $X_1|X_2$ and $X_2|X_1$ are not of the same type (e.g. we could just modify the example above by letting $EX_2 = 1$), and so $(X_{1,0}, X_{2,0}, X_{1,1}, X_{2,1}, \ldots)$ becomes an inhomogeneous Markov chain, while $(\mathbf{X}_0, \mathbf{X}_1, \ldots)$ is still a homogeneous Markov chain.

Returning to the general setting, let

$$\mathbf{X}_n = (X_{1,n}, X_{2,n}, \ldots, X_{k,n}), \qquad n = 0, 1, \ldots, \tag{12}$$

denote the Markov chain in Gibbs sampling. Then for *Gibbs sampling using a cyclic updating scheme*, given $\mathbf{X}_n$ we generate $\mathbf{X}_{n+1}$ by updating first $X_{1,n+1}$, second $X_{2,n+1}$, ..., and finally $X_{k,n+1}$ in accordance to the following:

$$\begin{aligned}
&X_{1,n+1} \text{ is drawn from } P_1(\cdot|X_{2,n}, \ldots, X_{k,n})\\
&X_{2,n+1} \text{ is drawn from } P_2(\cdot|X_{1,n+1}, X_{3,n}, \ldots, X_{k,n})\\
&\vdots\\
&X_{k,n+1} \text{ is drawn from } P_k(\cdot|X_{1,n+1}, \ldots, X_{k-1,n+1}).
\end{aligned}$$

Since

$$X_{i,n+1}|(X_{1,0}, X_{2,0}, \ldots, X_{i-1,n+1}) \sim P_i(\cdot|X_{1,n+1}, \ldots, X_{i-1,n+1}, X_{i+1,n}, \ldots, X_{k,n})$$

(with obvious modifications if $i = 1$), we have that $X_{i,n+1}$ and $(X_{1,0}, X_{2,0}, \ldots, X_{i-1,n})$ are independent given $(X_{i+1,n}, \ldots, X_{k,n}, X_{1,n+1}, \ldots, X_{i-1,n+1})$. So we say that

$$(X_{1,0}, X_{2,0}, \ldots, X_{k,0}, X_{1,1}, X_{2,1}, \ldots, X_{k,1}, \ldots)$$

is a Markov chain of order $k-1$ (it might be inhomogeneous, cf. Example 3). It follows that $(\mathbf{X}_0, \mathbf{X}_1, \ldots)$ is a (homogeneous) Markov chain.

The cyclic Gibbs sampler is not reversible. However, each update is in a way reversible, since for any $x \in \Omega$ and $y_i \in \Omega_i$, letting $y = (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k)$, we have detailed balance:

$$\pi(x)\pi_i(y_i|x^i) = \pi(x)\frac{\pi(y)}{\pi^i(x^i)} = \pi(y)\pi_i(x_i|x^i). \tag{13}$$

Because of this it can be verified that the chain (12) has π as an invariant density; this is also verified in Section 9. Furthermore, it can be shown that (9) implies that the cyclic Gibbs sampler is Harris recurrent (and hence irreducible) and aperiodic.

Sometimes other updating schemes are used. A *systematic updating scheme* which in fact leads to a *reversible Gibbs sampler* is given by a forward cycle followed by a backward cycle: Let $(X_{1,n}, X_{2,n}, \ldots, X_{k,n})$ denote the result after the nth forward cycle and

$$\mathbf{X}_n = (X_{k,n}, X_{k+1,n}, \ldots, X_{2k-1,n})$$

the result after the nth backward cycle, then

$$\begin{aligned}
&X_{1,n+1} \text{ is drawn from } P_1(\cdot|X_{2k-2,n}, \ldots, X_{k,n})\\
&X_{2,n+1} \text{ is drawn from } P_2(\cdot|X_{1,n+1}, X_{2k-3,n}, \ldots, X_{k,n})\\
&\vdots\\
&X_{k,n+1} \text{ is drawn from } P_k(\cdot|X_{1,n+1}, \ldots, X_{k-1,n+1})\\
&X_{k+1,n+1} \text{ is drawn from } P_{k-1}(\cdot|X_{1,n+1}, \ldots, X_{k-2,n+1}, X_{k,n+1})\\
&\vdots\\
&X_{2k-1,n+1} \text{ is drawn from } P_1(\cdot|X_{2k-2,n+1}, \ldots, X_{k+1,n+1}, X_{k,n+1}).
\end{aligned}$$

Alternatively, a *random updating scheme can be used to ensure reversibility*: Then given $\mathbf{X}_n$ we update $\mathbf{X}_{n+1}$ by first generating a random variable I_{n+1} from the uniform distribution on $\{1, 2, \ldots, k\}$. Suppose that $I_{n+1} = i$. Then we next generate $X_{i,n+1}$ from $P_i(\cdot|X_n^i)$, and we keep the rest, i.e. $X_{j,n+1} = X_{j,n}$ for $j \neq i$. This is also a reversible Gibbs sampler.

Exercise 9: Pump failure data *This exercise mainly consists in understanding and discussing the following statistical analysis and in implementing a Gibbs sampler for the problem. We give many details as this might be the first time the reader is seeing a Bayesian analysis for a higher-dimensional posterior distribution. These details may later appear quite trivial and will often be omitted in Bayesian textbooks and papers. In particular we try to be careful in relating the notation to that presented above on Gibbs sampling in a general context.*

At each of 10 pump stations, the number of failures y_i over a time span of length $t_i > 0$ was observed, where $i = 1, \ldots, 10$. We assume that the time span vector $t = (t_1, \ldots, t_{10})$ is known and fixed and that $y = (y_1, \ldots, y_{10})$ is our data, see Table 1. We consider y

as a realisation of a stochastic vector $\mathbf{Y} = (Y_1, \ldots, Y_{10})$ where each Y_i has state space $\{0, 1, 2, \ldots\}$.

Pump i	Failures y_i	Time span t_i
1	5	94.320
2	1	15.720
3	5	62.880
4	14	125.760
5	3	5.240
6	19	31.440
7	1	1.048
8	1	1.048
9	4	2.096
10	22	10.480

Table 1: Pump failure data

Below we consider a Bayesian MCMC analysis of these data. We assume the following model structure illustrated in Figure 1:

1. The distribution of each Y_i depends on the realisation of a positive random variable Λ_i, which we interpret as an unobserved failure rate. Specifically, the conditional distribution of Y_i given $\Lambda_i = \lambda_i$ is assumed to be a Poisson distribution with mean $t_i\lambda_i$ (can you argue why this may be a reasonable model assumption?).

2. Conditional on $\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_{10})$, we assume that $Y_1, \ldots, Y_{10}$ are independent.

3. The distribution of $\mathbf{\Lambda}$ depends on a positive random variable B, which we interpret as the mean level of the unobserved failure rates. Specifically, conditional on $B = \beta$, we assume that $\Lambda_1, \ldots, \Lambda_{10}$ are independent, and each Λ_i is exponentially distributed with parameter β.

4. We assume that B is exponentially distributed with parameter 40 (briefly, this value was obtained by some experimentation).

5. Finally, we assume that the conditional distribution of $\mathbf{Y}$ given $(\mathbf{\Lambda}, B)$ does not depend on B but only on $\mathbf{\Lambda}$ (in fact this is something one can immediately obtain from the graphical representation used in Figure 1, as there is no direct arrow from B to $Y_1, \ldots, Y_{10}$).

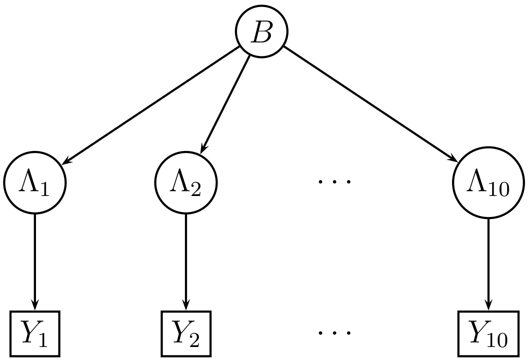

Figure 1: Hierarchical structure for pump failure data.

A model with this kind of structure is known as a *hierarchical model*. Such models are widely used within Bayesian statistics.

Specification of data distribution: This is given by the model assumptions 1., 2. and 5. above: By 1. and 2., given that $\Lambda_1 = \lambda_1, \ldots, \Lambda_{10} = \lambda_{10}$, then $\mathbf{Y}$ has conditional density

$$\pi(y|\lambda) = \prod_{i=1}^{10} \pi(y_i|\lambda_i) = \prod_{i=1}^{10} \frac{(t_i\lambda_i)^{y_i}}{y_i!} e^{-t_i\lambda_i}, \qquad y_1, \ldots, y_{10} \in \{0, 1, 2, \ldots\} \tag{14}$$

where $y = (y_1, \ldots, y_{10})$ and $\lambda = (\lambda_1, \ldots, \lambda_{10})$. Furthermore, by 5., the conditional density of $\mathbf{Y}$ given $(\Lambda, B) = (\lambda, \beta)$ does not depend on β:

$$\pi(y|\lambda, \beta) = \pi(y|\lambda). \tag{15}$$

Specification of prior distribution: This is given by the model assumptions 3. and 4. above[12]: B has density

$$\pi(\beta) = 40e^{-40\beta}, \qquad \beta > 0,^{13} \tag{16}$$

and conditional on $B = \beta$ we have that $\mathbf{\Lambda}$ has density

$$\pi(\lambda|\beta) = \prod_{i=1}^{10} \pi(\lambda_i|\beta) = \prod_{i=1}^{10} \beta \exp(-\beta\lambda_i) = \beta^{10} \exp(-\beta\lambda.) \tag{17}$$

for $\lambda = (\lambda_1, \ldots, \lambda_{10}) \in (0, \infty)^{10}$, where $\lambda. = \lambda_1 + \ldots + \lambda_{10}$. The prior density of $(\mathbf{\Lambda}, B)$ is simply given by $\pi(\lambda, \beta) = \pi(\lambda|\beta)\pi(\beta)$.

[12] These rather simple prior assumptions are mainly made for illustrative purposes. More "realistic" prior assumptions can be imposed but leads to a more complicated analysis.

[13] In Bayesian statistics B is called a hyper parameter and $\pi(\beta)$ a hyper prior.

Specification of posterior distribution: This is specified by the conditional density of $(\mathbf{\Lambda}, B)$ given the data $\mathbf{Y} = y$:

$$\pi(\lambda, \beta|y) = \pi(y, \lambda, \beta)/\pi(y) \propto \pi(y, \lambda, \beta)$$

where the proportionality follows from the fact that $\pi(y)$ is a constant when the data $Y = y$ is given. The joint density $\pi(y, \lambda, \beta)$ is determined by (14)–(17):

$$\pi(y, \lambda, \beta) = \pi(y|\lambda, \beta)\pi(\lambda, \beta) = \pi(y|\lambda)\pi(\lambda|\beta)\pi(\beta),$$

and so the posterior distribution of (Λ, B) has unnormalised density

$$\pi(\lambda, \beta|y) \propto \beta^{10} e^{-40\beta} \prod_{i=1}^{10} \lambda_i^{y_i} e^{-(t_i+\beta)\lambda_i}, \qquad (\lambda, \beta) \in (0, \infty)^{11}, \tag{18}$$

where again we have omitted any factors not depending on (λ, β).

Specification of full conditionals: Consider the random vector $\mathbf{X} = (\Lambda_1, \dots, \Lambda_{10}, B)$ consisting of $k = 11$ one-dimensional components. For the Gibbs sampler below we need to find the 11 full conditionals obtained from the posterior density (18). In this connection we make the following points: Recall that a gamma distribution with parameters $a > 0$ and $b > 0$ has density

$$f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \qquad x > 0.$$

Furthermore, if we briefly return to the general setting of a Gibbs sampler, recall that a full conditional is given by $\pi_i(x_i|x^i) = \pi(x)/\pi^i(x^i)$. Consider x^i to be fixed. Then ignoring both the denumerator $\pi^i(x^i)$ and any other constant factors only depending on x^i it is usually easy from

$$\pi_i(x_i|x^i) \propto \pi(x)$$

to determine if $\pi_i(x_i|x^i)$ is proportional to the density of a known distribution. So e.g. if we obtain that $\pi_i(x_i|x^i) \propto x_i^{a-1} e^{-bx_i}$, we know that this is an unnormalised gamma density with parameters a and b (remember to check if this density is well-defined, i.e. if $a > 0$ and $b > 0$).

We show first that the full conditional $\pi_i(\lambda_i|\lambda^i, \beta, y) = \pi_i(\lambda_i|\beta, y)$ does not depend on $\lambda^i = (\lambda_1, \dots, \lambda_{i-1}, \lambda_{i+1}, \dots, \lambda_{10})$. For fixed β, (18) is a product of 10 terms depending on $\lambda_1, \dots, \lambda_{10}$, respectively (this means that given $(B, \mathbf{Y})$, we have that $\Lambda_1, \dots, \Lambda_{10}$ are independent). Consequently, $\pi_i(\lambda_i|\lambda^i, \beta, y) = \pi_i(\lambda_i|\beta, y)$. Moreover, it follows from (18) that Λ_i given $B = \beta$ and $\mathbf{Y} = y$ has conditional density

$$\pi_i(\lambda_i|\beta, y) \propto \lambda_i^{y_i} e^{-(t_i+\beta)\lambda_i}.$$

This is recognised as an unnormalised gamma density with parameters $y_i + 1$ and $t_i + \beta$.

It follows also from (18) that B given $\Lambda = \lambda$ and $\mathbf{Y} = y$ has conditional density

$$\pi_{11}(\beta|\lambda, y) \propto \beta^{10} e^{-(40+\lambda.)\beta}$$

which is recognised as an unnormalised gamma density with parameters 11 and $40 + \lambda.$ (in accordance with model assumption 5., it follows from this that given $\mathbf{\Lambda}$, we have that B and $\mathbf{Y}$ are independent).

Gibbs sampling: We want to sample from the posterior distribution using Gibbs sampling (with a cyclic updating scheme) based on the 11 full conditionals specified above. This leads to the following Gibbs sampler for producing a Markov chain $\mathbf{X}_n = (\mathbf{\Lambda}_n, B_n)$, $n = 0, 1, 2, \ldots$, where $\mathbf{\Lambda}_n = (\Lambda_{1,n}, \ldots, \Lambda_{10,n})$:

```
I  Choose initial values of $\mathbf{\Lambda}_0 = (\Lambda_{1,0}, \ldots, \Lambda_{10,0})$ and $B_0$.
II For $n = 0, 1, 2 \ldots$ do
    (a) for $i = 1, \ldots, 10$, generate $\Lambda_{i,n+1}$ from a gamma distribution with parameters
        $y_i + 1$ and $t_i + B_n$,
    (b) generate $B_{n+1}$ from a gamma distribution with parameters 11 and $\Lambda_{1,n+1} + \cdots +$
        $\Lambda_{10,n+1} + 40$.
```

Problems:

1. Implement this Gibbs sampler as a function in R.
2. Use the function to simulate realisations of the chain $\mathbf{X}_0, \ldots, \mathbf{X}_{1000}$. For $i = 1, \ldots, 10$, summarise the simulation by a histogram based on $\Lambda_{i,0}, \ldots, \Lambda_{i,1000}$, and compare with the naive estimate y_i/t_i.

9 Metropolis within Gibbs or hybrid Metropolis-Hastings algorithms

Consider again the situation in (7) and (8). Recall that the Gibbs sampler (with a cyclic or another given updating scheme) consists in updating from the full conditionals, and each such update is reversible in the sense of (13). Suppose it is not convenient to generate simulations from the full conditional $\pi_i(x_i|x^i)$. Then instead a Metropolis-Hastings update might be used: Let $q_i(y_i|x)$ be a proposal density, where y_i plays the role of the proposal and $x = (x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k)$ is the current value before the update of x_i. Note that the proposal density is allowed not only to depend on x^i but also on x_i. For instance, $q_i(y_i|x) = f_i(y_i - x_i)$ could specify a random walk type proposal, where f_i is a symmetric density. Define the Hastings ratio by

$$H_i(x_i, y_i|x^i) = \frac{\pi_i(y_i|x^i) q_i(x_i|(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k))}{\pi_i(x_i|x^i) q_i(y_i|(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k))}$$

and the acceptance probability by

$$a_i(x, y_i) = \min\{1, H_i(x_i, y_i|x^i)\}.$$

In the random walk case $q_i(y_i|x) = f_i(y_i - x_i)$ we obtain

$$H_i(x_i, y_i|x^i) = \frac{\pi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k)}{\pi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k)}$$

whenever $f_i(y_i - x_i) > 0$. The Gibbs sampler is the special case where all $q_i(y_i|x) = \pi_i(y_i|x^i)$, whereby $a_i(x, y_i) = 1$.

MCMC algorithms based on such combinations of Gibbs updates and Metropolis-Hastings updates are called *Metropolis within Gibbs* or *hybrid Metropolis-Hastings algorithms*. Exercise 10 below shows a specific example of such an algorithm. As in the Gibbs sampler, a Metropolis within Gibbs algorithms has π as its invariant density: consider an update of the ith component when x^i is fixed; it follows from similar arguments as in Section 6 that the update of the ith component satisfies detailed balance and hence that the ith full conditional $\pi_i(\cdot|x^i)$ is invariant; consequently, π is invariant. Properties such as irreducubility, aperiodicity, etc. have to be established for the particular construction considered. Note that

$$H_i(x_i, y_i|x^i) = \frac{\pi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k) q_i(y_i|(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k))}{\pi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k) q_i(y_i|(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_k))}$$

so we actually do not need to determine the full conditional $\pi_i(x_i|x^i)$. Moreover, as in the original Metropolis-Hastings algorithm, *we need only to know the density π up to proportionality*.

Exercise 10 : Rat tumor data *This exercise concerns a statistical analysis similar to the one considered in Exercise 9. The main difference is that a Metropolis within Gibbs algorithm is needed for the simulation of the posterior. For these reasons this exercise contains less details compared to Exercise 9.*

In the following we consider the results of a clinical study of a specific type of tumor among rats. The study consisted of 71 experiments, where the ith experiment consisted

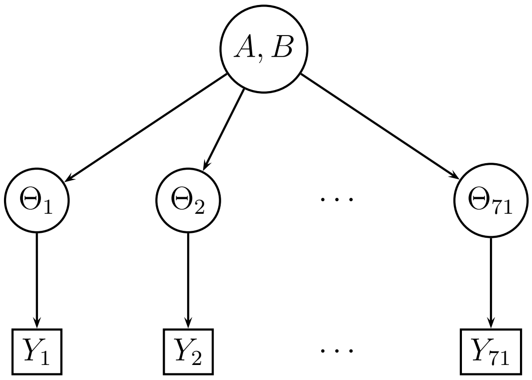

Figure 2: Hierarchical structure for rat tumor data.

in counting the number of tumor cases y_i among the n_i rats in that experiment. Since all rats in the data set were in a control group, the rates were not exposed to any special treatment. We assume that the vector of group sizes $n = (n_1, \ldots, n_{71})$ is known and fixed and that $y = (y_1, \ldots, y_{71})$ is our data. We consider y as a realisation of a stochastic vector $\mathbf{Y} = (Y_1, \ldots, Y_{71})$ where each Y_i has state space $\{0, 1, 2, \ldots, n_i\}$.

As in Exercise 9 we assume a hierarchical model structure, which is illustrated in Figure 2 and specified as follows.

1. The distribution of each Y_i depends on a realisation of a random variable $\Theta_i \in (0, 1)$ which we interpret as the death rate in the ith group of rats. Specifically we assume that the conditional distribution of Y_i given $\Theta_i = \theta_i$ is binomial with parameters θ_i and n_i.

2. Conditional on $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_{71})$ we assume that $Y_1, \ldots, Y_{71}$ are independent.

3. The distribution of $\boldsymbol{\Theta}$ depends on a realisation of two positive random variables A and B: given $A = \alpha$ and $B = \beta$, we assume that $\Theta_1, \ldots, \Theta_{71}$ are independent and Θ_i is beta distributed with parameters α and β.

4. We assume that the joint distribution of A and B has density [14] $\pi(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$.

5. Finally, analog to Exercise 9, we assume that the conditional distribution of $\mathbf{Y}$ given $(\boldsymbol{\Theta}, A, B)$ does not depend on (A, B).

[14] See Gelman et al. (2004) for a discussion of why this is an appropriate choice.

Specification of data distribution: Model assumptions 1. and 2. imply that the conditional density of $\mathbf{Y}$ given $\Theta_1 = \theta_1, \ldots, \Theta_{71} = \theta_{71}$ is

$$\pi(y|\theta) = \prod_{i=1}^{71} \pi(y_i|\theta_i) = \prod_{i=1}^{71} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta)^{n_i-y_i}, \qquad y_i \in \{0, 1, \ldots, n_i\}, i = 1, \ldots, 71,$$

where $y = (y_1, \ldots, y_{71})$ and $\theta = (\theta_1, \ldots, \theta_{71})$. By 5., $\pi(y|\theta, \alpha, \beta) = \pi(y|\theta)$ does not depend on (α, β).

Specification of prior distribution: By 3., conditional on $A = \alpha$ and $B = \beta$, the conditional density of $\mathbf{\Theta}$ is

$$\pi(\theta|\alpha, \beta) = \prod_{i=1}^{71} \pi(\theta_i|\alpha, \beta) = \prod_{i=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1},$$

$\theta = (\theta_1, \ldots, \theta_{71}) \in (0,1)^{71}$. The prior density of $(\mathbf{\Theta}, A, B)$ is then given by $\pi(\theta, \alpha, \beta) = \pi(\theta|\alpha, \beta)\pi(\alpha, \beta)$, where $\pi(\alpha, \beta)$ is given in 4.

Specification of posterior distribution: By similar arguments as in Exercise 9 we obtain that the posterior distribution of $(\mathbf{\Theta}, A, B)$ given data $\mathbf{Y} = y$ has density

$$\begin{aligned}
\pi(\theta, \alpha, \beta|y) &\propto \pi(\alpha, \beta)\pi(\theta|\alpha, \beta)\pi(y|\theta) \\
&= \pi(\alpha, \beta) \prod_{i=1}^{71} \pi(\theta_i|\alpha, \beta) \prod_{i=1}^{71} \pi(y_i|\theta_i) \\
&\propto (\alpha+\beta)^{-5/2} \left\{ \prod_{i=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \right\} \\
&\quad \left\{ \prod_{i=1}^{71} \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \right\} \qquad (19) \\
&\propto (\alpha+\beta)^{-5/2} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \right)^{71} \prod_{i=1}^{71} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1} \qquad (20)
\end{aligned}$$

for $\theta \in (0,1)^{71}$, $\alpha > 0$ and $\beta > 0$.

Specification of full conditionals: Using argument like in Exercise 9 verify the following points A.-C.

A. Given $(A, B) = (\alpha, \beta)$ and $\mathbf{Y} = y$ the conditional distribution of $\mathbf{\Theta}$ has density

$$\pi(\theta|\alpha, \beta, y) \propto \prod_{i=1}^{71} \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1}, \qquad \theta \in (0,1)^{71}. \qquad (21)$$

B. The conditional distribution (21) implies that, conditional on $(A, B) = (\alpha, \beta)$ and $\mathbf{Y} = y$, we have that $\Theta_1, \dots, \Theta_{71}$ are independent and Θ_i is beta distributed with parameters $\alpha + y_i$ and $\beta + n_i - y_i$.
Hint: *A beta distributed random variable with parameters $a > 0$ and $b > 0$ has unnormalised density*

$$f(x) \propto x^{a-1}(1-x)^{b-1} \quad \text{for } 0 < x < 1.$$

C. The joint conditional distribution of (A, B) given $\boldsymbol{\Theta} = \theta$ and $\mathbf{Y} = y$ has unnormalised density

$$\pi(\alpha, \beta|\theta, y) \propto (\alpha+\beta)^{-5/2} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{71} \prod_{i=1}^{71} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}, \qquad (\alpha, \beta) \in (0, \infty)^2. \tag{22}$$

Notice that this is not the (unnormalised) density of any standard distribution.

Metropolis within Gibbs sampling: As it is not immediate how to sample (A, B) directly from (22) we propose to use a Metropolis within Gibbs algorithm for sampling from the posterior density (20). The Metropolis within Gibbs algorithm has 72 components: $\Theta_1, \dots, \Theta_{71}$ and (A, B). By A. above we can easily (by a Gibbs step) sample from the conditional distribution of Θ_i given $(A, B) = (\alpha, \beta)$ (this conditional distribution does not depend on $\Theta_1, \dots, \Theta_{i-1}, \Theta_{i+1}, \dots, \Theta_{71}$). For (A, B) we propose to use a Metropolis random walk update where the proposal comes from a bivariate normal distribution as specified in II(b) below.

The Metropolis within Gibbs sampler generates a Markov chain $\mathbf{X}_n = (\boldsymbol{\Theta}_n, A_n, B_n)$, $n = 0, 1, 2, \dots$, as follows, where $\boldsymbol{\Theta}_n = (\Theta_{1,n}, \dots, \Theta_{71,n})$:

I Choose initial values of $\boldsymbol{\Theta}_0 = (\Theta_{1,0}, \dots, \Theta_{71,0})$, A_0 and B_0.

II For $n = 0, 1, 2, \dots$ do

(a) For $i = 1, \dots, 71$ sample $\Theta_{i,n+1}$ given A_n and B_n from a beta distribution with parameters $A_n + y_i$ and $B_n + n_i - y_i$.

(b) Generate independent proposals $A'_{n+1} \sim \mathrm{N}(A_n, \sigma_\alpha^2)$ and $B'_{n+1} \sim \mathrm{N}(B_n, \sigma_\beta^2)$ (where $\sigma_\alpha^2 > 0$ and $\sigma_\beta^2 > 0$ are user specified parameters).

(c) Generate $U_{n+1} \sim \texttt{unif}(0, 1)$.

(d) If $U_{n+1} < \pi(A'_{n+1}, B'_{n+1}|\boldsymbol{\Theta}_{n+1}, y)/\pi(A_n, B_n|\boldsymbol{\Theta}_{n+1}, y)$ then $(A_{n+1}, B_{n+1}) = (A'_{n+1}, B'_{n+1})$ otherwise $(A_{n+1}, B_{n+1}) = (A_n, B_n)$.

Implementing the Metropolis within Gibbs algorithm:

D. Implement the above Metropolis within Gibbs algorithm in R. You should implement the algorithm so that the mean acceptance probability for the Metropolis update of (A, B) can be calculated. The data can be loaded into R using

```
data <- read.table("/user/kkb/rats.dat",header=T)
```

The data consists of two variables y and n (use `names(data)` to verify this).

E. As the initial state choose $\Theta_{i,0} = y_i/n_i$, $i = 1, \dots, 71$, $A_0 = 1.6$ and $B_0 = 10$. Further, initially let $\sigma_\alpha^2 = 0.5^2$, $\sigma_\beta^2 = 2.5^2$ and let the sample length be 500. You may want to experiment with the values of σ_α^2, σ_β^2 and the sample length. Does different choices of σ_α^2 and σ_β^2 affect the mean acceptance probability?

F. Consider how to summarise the results. One possibility is to derive the 2.5%, 50% and 97.5% quantiles for each Θ_i, $i = 1, \dots, 71$ — in this case consider how you could display all results regarding $\Theta_1, \dots, \Theta_{71}$ in a single plot. Further, you should compare your results to the naive estimate y_i/n_i of Θ_i, $i = 1, \dots, 71$. You may also want to examine a plot of B against A.
Hint: *You should evaluate* (22) *on the log scale to avoid numerical problems in R.*

10 Output analysis

When MCMC samples are used instead of samples of independent simulations at least three problems emerge: assessment of convergence of Markov chains (burn-in, cf. Section 4); computation of auto-correlations and the asymptotic variance of a Monte Carlo estimate; and subsampling of a Markov chain. We consider each of these issues in the sequel.

10.1 Assessment of convergence

The *burn-in* is the time $j \geq 0$ at which the marginal distribution of a Markov chain state X_j is sufficiently close to its limit distribution Π for all practical purposes (provided Theorem 2 in Section 4 applies). The states in the initial part of the chain may be far from the limiting distribution due to the choice of the value of X_0, so to reduce the bias of Monte Carlo estimates, it is sensible to use only $X_m,\ m \geq j$. Below we consider a simple graphical method for determining the burn-in.

Visual inspection of *trace plots* $k(X_m),\ m = 0, 1, \dots,$ for various real functions k is a commonly used method to assess if the chain has or has not reached equilibrium. Figures 3 and 4 are examples of trace plots.

Suppose trace plots are obtained from two or more chains of length n with different initial values. Often one considers different extreme starting values. If the chains behave differently according to the trace plots for large values of $m \leq n$, the burn-in for at least one of the chains will be greater than n.

Example 4 In Exercise 10 we performed a Bayesian analysis of the number of tumor cases among rats. There we constructed a Metropolis within Gibbs algorithm for sampling the posterior distribution. One question is how we should choose the sample length n of the Metropolis within Gibbs algorithm so that the resulting approximation of the posterior distribution is "good enough". By "good enough" we mean that the Markov chain is effectively a sample from the target distribution (in this case the posterior distribution) and that the Markov chain is long enough that it is possible to estimate posterior quantities of interest satisfactory.

Figure 3 shows trace plots of Θ_{35}, A and B resulting from a application of the Metropolis within Gibbs sampler in Exercise 10 when $n = 500$. The trace plot of Θ_{35} indicates that convergence has effectively been reached after just a few updates; the trace plots for the other Θ_i lead to similar conclusions. However, it is unclear from the trace plots of A and B if the chain has (effectively) converged. So can we still trust results for Θ_{35}? We reapplied the Metropolis within Gibbs sampler, but this time with $n = 5000$, whereby Figure 4 was obtained. The trace plots look then more satisfactory, though the high picks after about 3700 iterations may indicate some concern. Possibly one would like to run the chain for even longer. . . .

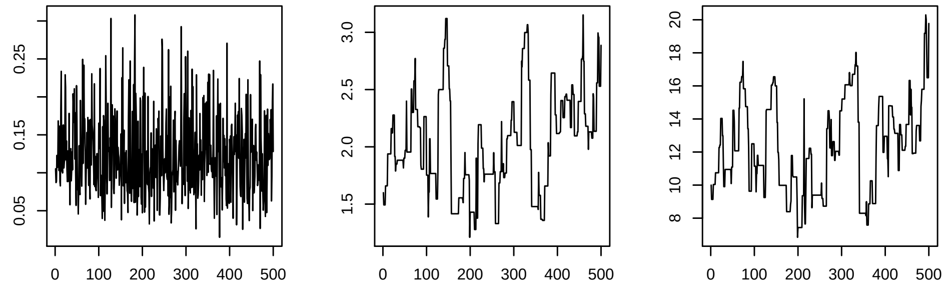

Figure 3: From left to right: trace plots of Θ_{35}, A and B when $n = 500$.

In Figure 5 we compare the posterior distributions of $\boldsymbol{\Theta}$ estimated from the Markov chain of length 500 in Figure 3 with the corresponding estimates based on the last 4500 elements of the Markov chain of length 5000 in Figure 4. A visual inspection of the two plots in Figure 5 show only minor differences suggesting that $n = 500$ and no burn-in is "good enough" for the posterior analysis of $\boldsymbol{\Theta}$ (which is the parameter of main interest in this

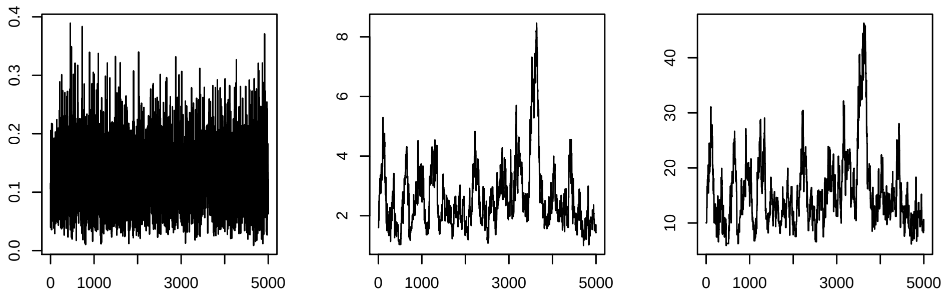

Figure 4: From left to right: trace plots of Θ_{35}, A and B when $n = 5000$.

application).

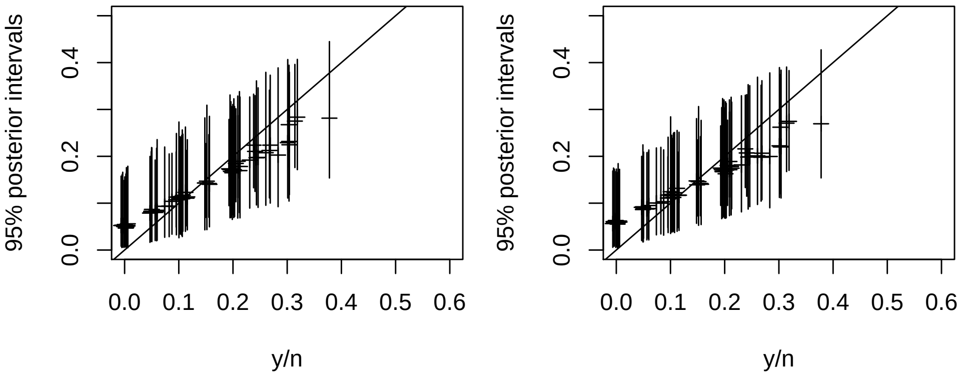

Figure 5: Left plot: summary of posterior distribution estimated from the Markov chain in Figure 3. Right plot: summary of posterior distribution estimated from the last 4500 iterations of the Markov chain in Figure 4. Both plots: each vertical line represents the 95% central posterior interval for a Θ_i, and the small horizontal line represent the corresponding median. Small random jitter has been added in the horisontal direction to help distinguishing the vertical lines.

Example 5 Assume that we want to sample (X_1, X_2) from a bivariate normal distribution with density $\pi(x_1, x_2)$ given by (10) when $\rho = 0.5$. We use the following Metropolis random walk algorithm:

I Choose initial value $\mathbf{X}_0 = (X_{1,0}, X_{2,0})$.

II For $i = 0, \ldots, n-1$ do

(a) Generate independent proposals $Y_{1,i+1} \sim \mathrm{N}(X_{1,i}, \sigma^2)$ and $Y_{2,i+1} \sim \mathrm{N}(X_{2,i}, \sigma^2)$ (where σ^2 is a user specified parameter).

(b) Generate $U_{i+1} \sim \texttt{unif}(0, 1)$.

(c) If $U_{i+1} \leq \pi(Y_{1,i+1}, Y_{2,i+1})/\pi(X_{1,i}, X_{2,i})$ then $(X_{1,i+1}, X_{2,i+1}) = (Y_{1,i+1}, Y_{2,i+1})$ otherwise $(X_{1,i+1}, X_{2,i+1}) = (X_{1,i}, X_{2,i})$.

We have applied this Metropolis random walk algorithm using four different starting values $(X_{1,0}, X_{2,0}) = (-5, -5), (-5, 5), (5, -5), (5, 5)$ and $n = 200$. Figure 6 shows the trace plots of the initial 10, 20, 50 and 200 states of the four Markov chains. Looking at these plots it is clear that the chains are far from the target distribution after both 10 and 25 updates. After 200 updates we seem to have effectively converged and even so after 100 updates. This suggests a burn-in of length 100.

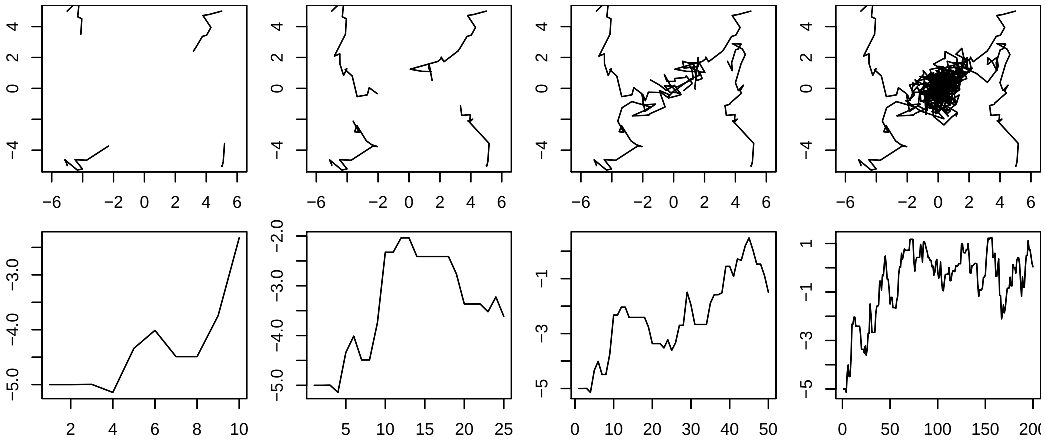

Figure 6: Top row: four trace plots of the first 10, 25, 50 and 200 states of a 2-dimensional Metropolis random walk chain. Bottom row: similar as in the top row but showing only the trace plots for the first component.

10.2 Estimation of correlations and asymptotic variances

Plots of estimated auto-correlations and cross-correlations for different statistics often provide good indications of the chain's *mixing behaviour*. Assume that $X_j \sim \Pi$ for some time $j \geq 0$. For a given real function k with finite variance

$$\sigma^2 = Var(k(X_j)),$$

define the lag m *auto-correlation* by

$$\rho_m = Corr(k(X_j), k(X_{j+m})), \quad m = 0, 1, \ldots.$$

Under fairly weak conditions, $\rho_m \to 0$ as $m \to \infty$. Similarly, given two functions $k^{(1)}$ and $k^{(2)}$ with finite variances, the lag m *cross-correlations* are defined by

$$\rho_m^{i_1,i_2} = Corr(k^{(i_1)}(X_j), k^{(i_2)}(X_{j+m})), \quad m = 0, 1, \ldots,\ i_1, i_2 \in \{1, 2\}.$$

In the reversible case, $\rho_m^{1,2} = \rho_m^{2,1}$. Under fairly weak conditions, $\rho_m^{i_1,i_2} \to 0$ as $m \to \infty$. The chain is slowly respectively rapidly mixing if the correlations are slowly respectively rapidly decaying to 0.

For the estimation of ρ_m and $\rho_m^{i_1,i_2}$, let us for ease of presentation assume that $j = 0$ (in practice a burn-in $j \geq 0$ may have been used, however, the following estimates are also consistent without a burn-in) and that we have generated n states $X_0, \ldots, X_{n-1}$ of the Markov chain. The lag m *auto-covariance* $\gamma_m = \sigma^2 \rho_m$ is estimated by the empirical auto-covariance

$$\hat{\gamma}_m = \frac{1}{n} \sum_{i=0}^{n-1-m} (k(X_i) - \bar{k}_n)(k(X_{i+m}) - \bar{k}_n)$$

for $m = 0, \ldots, n-1$ (there are good arguments for using the divisor n rather than $n - m$). Here

$$\bar{k}_n = \frac{1}{n} \sum_{i=0}^{n-1} k(X_i)$$

is the Monte Carlo estimate of $Ek(X_0)$. From this we obtain natural estimates

$$\hat{\sigma}^2 = \hat{\gamma}_0, \qquad \hat{\rho}_m = \hat{\gamma}_m / \hat{\sigma}^2, \qquad m = 0, 1, \ldots.$$

Similar methods apply for estimation of cross-correlations.

The *Monte Carlo error* of $\bar{k}_n$ can be expressed by the *Monte Carlo variance*

$$Var(\bar{k}_n) = \frac{\sigma^2}{n} \left[1 + 2 \sum_{m=1}^{n-1} \left(1 - \frac{m}{n} \right) \rho_m \right].$$

Under fairly general conditions a central limit theorem applies (see Section 4): then the *asymptotic variance* is well defined and finite and given by

$$\bar{\sigma}^2 = \lim_{n \to \infty} n Var\big(\bar{k}_n\big) = \sigma^2 \tau$$

where

$$\tau = 1 + 2 \sum_{m=1}^{\infty} \rho_m \tag{23}$$

is called the *integrated auto-correlation time*; and as $n \to \infty$, $\sqrt{n}(\bar{k}_n - Ek(X_0))$ is asymptotically normally distributed with mean 0 and variance $\bar{\sigma}^2$. The asymptotic variance determines the *efficiency of a Monte Carlo estimate*. In the special i.i.d. case, $\tau = 1$. Note that finiteness of τ implies by (23) that $\rho_m \to 0$ as $m \to \infty$.

Although $\hat{\gamma}_m$ is a consistent estimate of γ_m, it is well known that the obvious estimate $1 + 2\sum_{m=1}^{n-1} \hat{\rho}_m$ of τ is not consistent as $n \to \infty$.

One method for estimation of the asymptotic variance is the method of *batch means*: Suppose that $n = n_1 n_2$ where n_1 and n_2 are integers and n_2 is so large that we can treat the n_1 batch mean estimates

$$\bar{k}^{(i)}_{n_1,n_2} = \frac{1}{n_2} \sum_{m=(i-1)n_2}^{in_2-1} k(X_m), \quad i = 1, \dots, n_1,$$

as being (approximately) uncorrelated. Note that $\bar{k}_n = \sum_{i=1}^{n_1} \bar{k}^{(i)}_{n_1,n_2}/n_1$. When n_2 is sufficiently large, the batch mean estimates are approximately normally distributed. This suggests to estimate $Var(\bar{k}_n)$ by

$$\sum_{i=1}^{n_1} (\bar{k}^{(i)}_{n_1,n_2} - \bar{k}_n)^2/(n_1(n_1-1)).$$

Another method for estimating the asymptotic variance can be used in the reversible case: It can be shown that for an irreducible and reversible Markov chain, $\Gamma_m = \gamma_{2m} + \gamma_{2m+1}$ is a strictly positive, strictly decreasing, and strictly convex function of $m = 0, 1, \ldots$. Here strict convexity means that the sequence $\Gamma_0 - \Gamma_1, \Gamma_1 - \Gamma_2, \Gamma_2 - \Gamma_3, \ldots$ is strictly decreasing. Let $l_s \leq (n-2)/2$, $s = \text{pos}, \text{mon}, \text{conv}$, be the largest integers so that $\hat{\Gamma}_m = \hat{\gamma}_{2m} + \hat{\gamma}_{2m+1}$, $m = 0, \ldots, l_s$, is respectively strictly positive, strictly decreasing, or strictly convex. Then it can verified that the *initial sequence estimates*

$$\hat{\tau}_s = 1 + 2 \sum_{m=1}^{2L_s+1} \hat{\rho}_m, \tag{24}$$

where

$$L_{\text{pos}} = l_{\text{pos}}, \; L_{\text{mon}} = \min\{L_{\text{pos}}, l_{\text{mon}}\}, \; L_{\text{conv}} = \min\{L_{\text{mon}}, l_{\text{conv}}\},$$

provide consistent conservative estimates of τ, i.e.

$$\lim \inf_{n \to \infty} (\hat{\sigma}^2 \hat{\tau}_s) \geq \bar{\sigma}^2 \qquad \text{for } s = \text{pos}, \text{mon}, \text{conv}$$

where $\hat{\tau}_{\text{pos}} \geq \hat{\tau}_{\text{mon}} \geq \hat{\tau}_{\text{conv}}$.

Example 6 In Figure 7 we have shown results based on simulating three Markov chains using the Metropolis random walk algorithm in Exercise 8 when $n = 1000$ and $\sigma = 0.005, 0.05, 0.5$, respectively. The corresponding mean acceptance probabilities are $\bar{a} =$ 0.892, 0.331 and 0.028; the case $\sigma = 0.05$ is the only one where $\bar{a}$ is between 0.2 and 0.4.

Comparing the three trace plots in Figure 7, the case $\sigma = 0.05$ seems fastest converging and best mixing. The chain is less stable when $\sigma = 0.005$, and for $\sigma = 0.5$ the chain gets stuck for long periods of time.

Comparing the plot of lag m auto-correlations, again the case $\sigma = 0.05$ looks best with the auto-correlation quickly getting close to zero. For the case $\sigma = 0.005$ the auto-correlation is reaching zero slower because the proposals are close to current value. For the case $\sigma = 0.5$ the auto-correlation reaches zero slowly because of the "stickiness" of the Markov chain in this case.

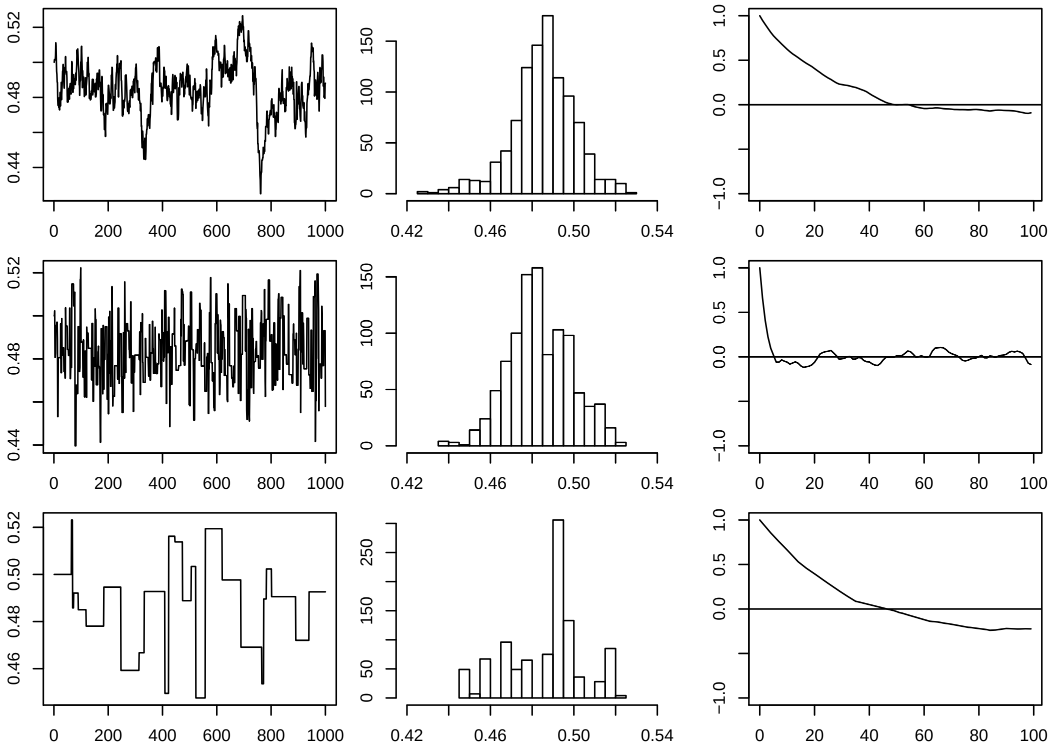

Figure 7: Left coloumn: trace plots for Markov chains obtained using the Metropolis random walk algorithm in Exercise 8 when $n = 1000$. Centre column: histograms based on the sampled chains. Right column: lag m auto-correlations for $m = 1, \ldots, 100$ estimated from the sampled chains. Top to bottom: $\sigma = 0.005, 0.05, 0.5$, respectively.

To demonstrate the methods of batch means and the initial sequence estimate, we have simulated three Markov chains by the Metropolis random walk algorithm in Exercise 8 when $n = 50000$ and $\sigma = 0.005, 0.05, 0.5$, respectively. Figure 8 shows $\hat{\Gamma}_m$ estimated from the three chains. The corresponding values of the initial sequence estimate $\hat{\tau}_{\text{conv}}$ are 33.8, 4.62 and 38.63, and the estimates $\hat{\sigma}^2\hat{\tau}_{\text{conv}}$ of the upper bound on the asymptotic variance $\bar{\sigma}^2$ are 7.09×10^{-3}, 9.97×10^{-4} and 7.89×10^{-3}. Using batch means with $n_1 = 100$ and $n_2 = 500$ we obtain the following estimates of $Var(\bar{k}_n)$: 2.37×10^{-7}, 1.56×10^{-8} and 2.05×10^{-7}. Multiplying these numbers by n we obtain estimates of the asymptotic variance $\bar{\sigma}^2$: 9.77×10^{-3}, 9.11×10^{-4} and 7.11×10^{-3}. These figures are rather similar to those obtained using the initial sequence estimate.

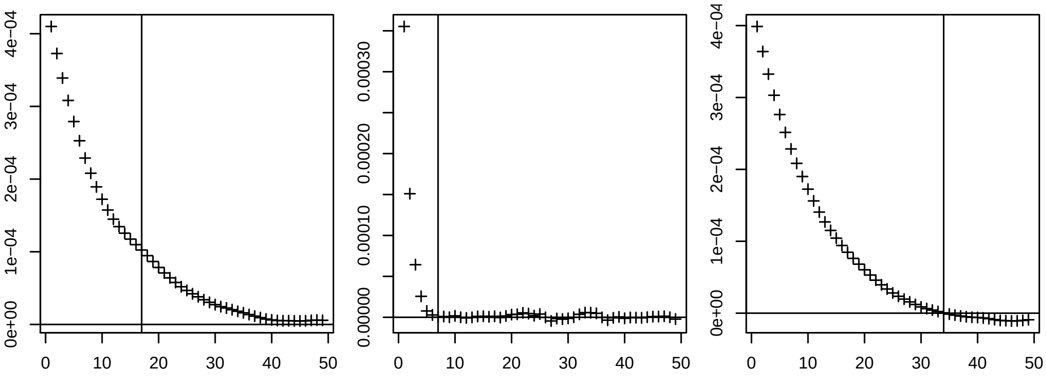

Figure 8: Plots of $\hat{\Gamma}_m$ estimated from three Markov chains generated using the random walk algorithm in Exercise 8 with $n = 50000$ and (from left to right) $\sigma = 0.005, 0.05, 0.5$. The vertical lines in each plot is located at L_{conv}.

10.3 Subsampling

Sometimes *subsampling* with a *spacing* $s \geq 2$ is used, i.e. we use only the subchain $X_j, X_{j+s}, X_{j+2s}, \ldots$ for some given $j \in \mathbb{N}_0$. There may be various reasons for using a subsample: storage problems may be reduced if it is required to store a sample e.g. for plotting; trace and auto-correlation plots may be more informative; and more efficient Monte Carlo estimates may be obtained. It is not always optimal to use subsampling, however, if the samples are highly auto-correlated and the evaluation of $k(X_m)$ is expensive, then a large spacing s may be desirable.

For the Markov chain $X_j, X_{j+s}, X_{j+2s}, \ldots$, we proceed as above by substituting the original chain by the subsampled chain. For example, the asymptotic variance of the Monte Carlo average

$$\bar{k}_n^s = \sum_{m=0}^{n-1} k(X_{j+ms})/n$$

based on the subsampled chain is given by $\sigma^2\tau^s$, where

$$\tau^s = 1 + 2\sum_{m=1}^{\infty} \rho_{ms}$$

is the integrated auto-correlation time for the subsampled chain.

11 A final application example and exercise

This section consists of two main parts. In the first part we consider an example of a so-called *Gibbs distribution*, namely an Ising model. This distribution has been widely use in the analysis of digital images, and the Gibbs sampler, when it was named so by Stuart and Donald Geman in 1984, was applied on this model.

The second part concerns a Bayesian analysis of a data set related to archaeology. In this analysis we use the Ising model as the prior, and we discuss how to handle missing values in the data.

11.1 The Ising model

Consider the following situation: A rectangular region S is divided into J smaller disjoint rectangular (sub)regions $S_i \subseteq S$, so that $S = \cup_{i=1}^J S_i$. To each region S_i associate a binary variable $X_i \in \{0, 1\}$. We can think of S_i as a pixel in a binary digital image S and X_i as the “colour” of that pixel, in the sense that $X_i = 0$ corresponds to S_i being a black pixel and $X_i = 1$ corresponds to X_i being a white pixel.

For many applications it is natural to assume that neighbouring pixels are more likely to have the same colour than pixels far apart. For this reason we let the binary vector $\mathbf{X} = (X_1, \ldots, X_J)$ be distributed according to the *Ising density*

$$\pi(x) = \frac{1}{c(\beta)} \prod_{i=1}^{J} \exp(\beta u_i(x_i)), \qquad x = (x_1, \ldots, x_J) \in \{0,1\}^J, \tag{25}$$

where $c(\beta)$ is the normalising constant and

$$u_i(k) = \sum_{\text{8-neighbours of } S_i} \mathbf{1}[x_j = k]$$

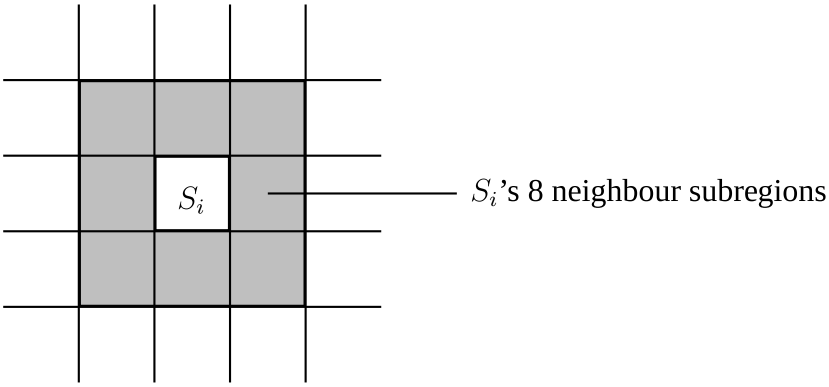

Figure 9: Shaded regions correspond to the neighbour regions of S_i.

is the number of 8-neighbours of S_i with the colour $k \in \{0,1\}$. More precisely, by " 8-neighbours of S_i" we mean the (up to) 4 nearest neighbours and (up to) 4 second nearest neighbours to "pixel" S_i, see Figure 9. Notice that pixels at the border of S have less than 8 neighbours. Further, notice that the notation $u_i(x_i)$ is a bit misleading, since $u_i(x_i)$ depends not only on x_i but the values x_j with S_j a nearest or second nearest neighbour to S_i. Furthermore, β is a real parameter and $c(\beta)$ is an unknown normalising constant in the sense that it is infeasible to calculate $c(\beta)$ (except for $\beta = 0$) even for moderate values of J. In fact $c(\beta) = \sum_x \prod_{i=1}^J \exp(\beta u_i(x_i))$ where the sum is over the 2^J different possible configurations of x. For example, if $J = 16^2$ (as in the examples below) then $2^J \approx 1.16 \times 10^{77}$.

To get a better understanding of (25) we consider the significance of β. For $\beta = 0$ we have $\pi(x) \propto 1$, so $\mathbf{X}$ is then uniformly distributed over the 2^J possible configurations of x. For $\beta > 0$ configurations of x containing many pairs of neighbouring pixels with the same colour have higher probability than configuration with few pairs of neighbouring pixels with the same colour. To see this notice that

$$\prod_{i=1}^J \exp(\beta u_i(x_i)) = \exp(\beta \sum_{i=1}^J u_i(x_i)),$$

where $\frac{1}{2}\sum_{i=1}^J u_i(x_i)$ is the number of pairs of neighbour pixels in x with the same colour. So for increasing β there is an increasing tendency towards preferring configurations of $\mathbf{X}$ where many pairs of neighbour pixels have the same colour, i.e. resulting in clumps of black and white pixels.

To get an even better understanding of (25) and the significance of β we want to produce realisations of $\mathbf{X}$ for different values of β. Simulation of $\mathbf{X}$ can be done using a Gibbs sampler with a cyclic updating scheme where each pixel value X_i corresponds to a single component. Thus we need the full conditionals:

A. Verify that the full conditional for X_i, $i = 1, \dots, J$, is given by

$$\pi(x_i|x^i) = \frac{\exp(\beta u_i(x_i))}{\sum_{k=0}^{1} \exp(\beta u_i(k))}, \qquad x_i \in \{0, 1\}.$$

B. Implement a Gibbs sampler in R for producing approximate samples of (25) in the case where S is divided into $J = 16 \times 16$ subregions.

   (a) It is convenient to consider $\mathbf{X}$ as a 16×16 matrix and implement the sampler as a function `gibbs(X,beta,n)` where the input `(X,beta,n)` corresponds to the initial state $\mathbf{X}_0$, β and the number of Gibbs updates n. As output the function should return the final state $\mathbf{X}_n$ of the chain (we can use `image(x)` to view the state of a matrix `x`).

   (b) Experiment with different choices of β, say $\beta = 0, 0.2, 0.4, 0.6$, and number of Gibbs updates, say $n = 1, 2, 5, 10, 25$. How does the final state $\mathbf{X}_n$ change with different choices of β? Does the choice of n have any influence on the final state? You may also consider different choices of $\mathbf{X}_0$, e.g. the zero-matrix

   ```
   X <- matrix(0,16,16)
   ```

   or a random matrix, e.g.

   ```
   X <- matrix(rbinom(256,1,0.5),nrow=16).
   ```

Hint: (a) *The number of neighbours of* `x[i,j]` *with value* 1 *can be counted by*

```
sum(x[max(i-1,1):min(i+1,16),
  max(j-1,1):min(j+1,16)]) - x[i,j]
```

The total number of neighbours of `x[i,j]` *is given by*

```
(min(i+1,16)-max(i-1,1)+1)*(min(j+1,16)-max(j-1,1)+1)-1
```

11.2 Analysis of archaeological data

In the following we consider a data set consisting of a grid of $J = 16 \times 16$ soil measurements $y_1, \dots, y_{256}$ of the log phosphate concentrations taken at 10 m intervals. Enhanced soil phosphate content, the result of decomposition of organic material, is often found at sites of known archaeological activity. Thus, measurements of phosphate concentration over a study region can provide a useful aid in locating sites that are already known to exist. We consider our data $y = (y_1, \dots, y_{256})$ to be a realisation of a stochastic vector $\mathbf{Y} = (Y_1, \dots, Y_{256})$ where $Y_i \in \mathbb{R}$. In the following we consider a Bayesian analysis of these data assuming the following model structure:

1. The distribution of each Y_i depends on the realisation of a binary variable $X_i \in \{0, 1\}$ with the following interpretation. The site S_i is said to be archaeologically active if $X_i = 1$, and archaeologically inactive otherwise. Specifically, conditional on $X_i = x_i$ we assume that Y_i is normal distributed with mean $\mu + \Delta x_i$ and variance κ^2. For these data it is reported that $\mu = 1$ and $\Delta = 1$. The variance κ^2 is so far treated as a known parameter (at the end we discuss how to estimate κ^2).

2. Conditional on $\mathbf{X} = (X_1, \dots, X_{256})$ we assume that $Y_1, \dots, Y_{256}$ are independent.

3. A priori we assume that $\mathbf{X}$ is distributed according to the Ising model (25), where we so far treat $\beta > 0$ as a known parameter (at the end we also discuss how to estimate β).

Notice that with the interpretation of X_i in 1. the prior assumption in 3. seems natural as nearby sites would be expected to be more likely to have the same level of archaeological activity than sites far apart.

Specification of the data distribution: The model assumptions 1. and 2. imply that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = (x_1, \dots, x_{256})$ has density

$$\pi(y|x) = \prod_{i=1}^{256} \frac{1}{\sqrt{2\pi\kappa^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \mu - \Delta x_i)^2}{\kappa^2} \right), \qquad y = (y_1, \dots, y_{256}) \in \mathbb{R}^{256}. \quad (26)$$

Specification of posterior distribution: As the prior distribution of $\mathbf{X}$ has density given by (25), the conditional distribution of $\mathbf{X}$ given $\mathbf{Y} = y$ has density

$$\begin{aligned} \pi(x|y) &\propto \pi(x)\pi(y|x) \\ &\propto \left\{ \prod_{i=1}^{256} \exp(\beta u_i(x_i)) \right\} \left\{ \prod_{i=1}^{256} \exp(-(y_i - \mu - \Delta x_i)^2/(2\kappa^2)) \right\} \\ &= \prod_{i=1}^{256} \exp(\beta u_i(x_i) - (y_i - \mu - \Delta x_i)^2/(2\kappa^2)). \end{aligned} \quad (27)$$

Specification of full conditionals:

A. Verify that the full conditional for X_i, $i = 1, \dots, 256$, is

$$\pi(x_i|x^i, y) = \frac{\exp(\beta u_i(x_i) - (y_i - \mu - \Delta x_i)^2/(2\kappa^2))}{\sum_{k=0}^{1} \exp(\beta u_i(k) - (y_i - \mu - \Delta k)^2/(2\kappa^2))}. \quad (28)$$

Gibbs sampling: Given the full conditionals (28) it is in straightforward to implement a Gibbs sampler for sampling from the posterior density (27): for simulation from the full conditionals we simply use inversion.

Missing data: The data under consideration is not complete. In fact the phosphate concentration has not been measured at all sites, so some y_i are missing. Sampling (27) using a Gibbs sampler then becomes a problem as the full conditionals in (28) depend on y.

We now consider briefly how to deal with missing data — a recurring situation in applied statistics. Let Y_{obs} denote the observed part and Y_{mis} the missing part of Y. Then the observed data is distributed according to

$$\pi(y_{\text{obs}}|x) = \int \pi(y_{\text{obs}}, y_{\text{mis}}|\theta)dy_{\text{mis}},$$

where the integral is replaced by a sum in the case of a discrete state space. Posterior inference is then done as usual but with the density $\pi(y|x)$ replaced by the data density $\pi(y_{\text{obs}}|x)$, so the conditional distribution of X given $Y_{\text{obs}} = y_{\text{obs}}$ has density

$$\pi(x|y_{\text{obs}}) \propto \pi(x)\pi(y_{\text{obs}}|x).$$

Specification of data distribution: In the present setting recall that given $\mathbf{X} = x$ the complete data is distributed according to

$$\pi(y|x) \propto \prod_{i=1}^{256} \exp(-(y_i - \mu - \Delta x_i)^2/(2\kappa^2)).$$

Using the approach above for dealing with missing data

$$\begin{aligned} \pi(y_{\text{obs}}|x) &\propto \int \prod_{i=1}^{256} \exp(-(y_i - \mu - \Delta x_i)^2/(2\kappa^2))dy_{\text{mis}} && (29) \\ &\propto \prod_{y_i \text{ is observed}} \exp(-(y_i - \mu - \Delta x_i)^2/(2\kappa^2)). && (30) \end{aligned}$$

This simple form is due to the model assumption 2.

Specification of the posterior distribution: The conditional distribution of $\mathbf{X}$ given $Y_{\text{obs}} = y_{\text{obs}}$ has density

$$\pi(x|y_{\text{obs}}) \propto \prod_{i=1}^{256} \exp(\beta u_i(x_i)) \prod_{y_i \text{ is observed}} \exp(-(y_i - \mu - \Delta x_i)^2/(2\kappa^2)). \qquad (31)$$

Specification of full conditionals:

B. Verify that

   (a) if Y_i is missing, the full conditional for X_i does not depend on y_{obs}, since

$$\pi(x_i|x^i, y_{\text{obs}}) = \frac{\exp(\beta u_i(x_i))}{\sum_{k=0}^{1} \exp(\beta u_i(k))};$$

   (b) if Y_i is observed, the full conditional for X_i is the same as in the case with no missing data, that is,

$$\pi(x_i|x^i, y_{\text{obs}}) = \frac{\exp(\beta u_i(x_i) - (y_i - \mu - \Delta x_i)^2/(2\kappa^2))}{\sum_{k=0}^{1} \exp(\beta u_i(k) - (y_i - \mu - \Delta k)^2/(2\kappa^2))}.$$

   These simple forms are again due to the model assumption 2.

Gibbs sampling: In a Bayesian analysis of the archaeological data we want to estimate for each site the posterior probability of that site being archaeologically active. We do this by sampling the posterior distribution in accordance with (31) using a cyclic Gibbs sampler where, as above, each activity level x_i corresponds to a single component in the Gibbs sampler.

Performing a Bayesian analysis of the archaeological data: The data is loaded into a 16×16 matrix `y` where missing values are given the value `NA` (this is the standard R notation for missing value) using

```
y <- read.table("/user/kkb/julian.dat")
y <- as.matrix(y)
y[y==1.5] <- NA
```

The command `is.na(x[i,j])` will determine if `x[i,j]` has value `NA`. We need to use an extra option `na.rm=T` for some commands in order that they ignore any occurrence of `NA`.

C. Implement a Gibbs sampler for sampling (31) in R as a function

```
gibbs.2(X,beta,kappa,mu,Delta,n)
```

   where the input `(X,beta,kappa,mu,Delta,n)` corresponds to the initial state $\mathbf{X}_0$, the parameters β, κ, μ and Δ, and the number of cyclic Gibbs updates n. As output the function should return both the final state $\mathbf{X}_n$ as a matrix and the mean $\mathbf{M} = (M_1, \dots, M_{256}) = \frac{1}{n+1}\sum_{i=0}^{n} \mathbf{X}_i$, also as a matrix. The quantity $M_i \in (0, 1)$ is the estimated posterior probability that the site S_i is archaeologically active.

Predicting unobserved variables: The conditional density for (Y_{mis}) given $X = x$ and the data $Y_{\text{obs}} = y_{\text{obs}}$ is

$$\pi(y_{\text{mis}}|x, y_{\text{obs}}) \propto \pi(y_{\text{mis}}|x) = \prod_{y_i \text{ is missing}} \frac{1}{\sqrt{2\pi\kappa^2}} \exp(-(y_i - \mu - \Delta x_i)^2/(2\kappa^2))$$

because of the model assumptions 1. and 2. In other words, given $X = x$ and the data $Y_{\text{obs}} = y_{\text{obs}}$, the unobserved Y_i are independent and Y_i is normally distributed with mean $\mu + \Delta x_i$ and variance κ^2. Thus we can easily predict the unobserved Y_i by simulating from these independent normal distributions.

Choosing $\mathbf{X}_0$*:* Consider the following thresholding: for $i = 1, \dots, 256$ set $\hat{x}_i = 1$ if $y_i > \mu + \Delta/2$ and $\hat{x}_i = 0$ otherwise. An appropriate initial state is then $\mathbf{X}_0 = \hat{x} = (\hat{x}_1, \dots, \hat{x}_{256})$. In R this is done by

```
X <- matrix(0,16,16)
X[y>(mu + Delta*0.5)] <- 1
```

Choosing κ *and* β*:* The binary vector $\hat{x}$ obtained via the thresholding above can be seen as an naive estimate of $\mathbf{X}$. Given this estimate we can obtain an estimate of κ^2 by

$$\hat{\kappa}^2 = \frac{1}{n_{\text{obs}}} \sum_{y_i \text{ is observed}} (y_i - \mu - \Delta \hat{x}_i)^2$$

where n_{obs} denotes the number of observed phosphate concentrations.

One way to choose β is so that it maximises (25) when $x = \hat{x}$ — this is a so-called maximum likelihood estimate. As $c(\beta)$ in (25) is unknown, maximum likelihood estimation is far from straightforward. Instead initially assume that $\beta = 0.8$. As we have little knowledge about how to choose β you should experiment with different values of β to see how it affects the results of the analysis.

Dealing with unknown normalising constants: In the previous examples we have assumed that a parameter like β is a realisation of a random variable B, say. The reader may wonder why we have not done it here as we have little knowledge about what β should be. Assume that β is in fact a realisation of a random variable B which is distributed according to a hyper prior density $\pi(\beta)$. Assume further that the above Gibbs sampler is replaced by a Metropolis within Gibbs sampler where X_i is updated as before (using a Gibbs update) and B is updated in accordance to the posterior density of (X, B), using e.g. a Metropolis random walk update. If β is the current value of B and β' is the proposal, then the acceptance probability would be

$$\min\left\{1, \frac{\pi(x|y,\beta')\pi(\beta')}{\pi(x|y,\beta)\pi(\beta)}\right\}, \tag{32}$$

where $\pi(x|y,\beta)$ is given by (27) (here we just have made the dependence on β explicit). Evaluating (32) then involves evaluating a ratio $c(\beta)/c(\beta')$ of unknown normalising constants making it impossible to apply the Metropolis algorithm as presented here. There exists a number of solution to this problem.

One solution is to estimate the ratio of unknown normalising constants. One way of doing this is based on importance sampling as follows. Assume that $\pi(x|\theta) = c(\theta)^{-1} f(x|\theta)$

is a normalised density with unknown normalising constant $c(\theta) = \int f(x|\theta)dx$. Let $X_0, X_1, \ldots, X_n$ be a Markov chain with $\pi(x|\theta)$ as its equilibrium density, e.g. obtain using a Metropolis-Hastings algorithm. Then

$$\frac{1}{n+1}\sum_{i=0}^{n}\frac{f(x_i|\beta')}{f(x_i|\beta)}$$

is an estimate of $c(\theta')/c(\theta)$. See Gelman & Meng (1998) for more details on this and other more sophisticated ways of estimating ratios of unknown normalising constants. See also Møller et al. (2004).

12 C in R

As you may have noticed R can be very slow when used for heavy computations. An effective solution is to implement parts of the R code in C. Here we just consider a small example of how to implement C in R. For more details see the help pages in R.

The following example implements a Metropolis random walk algorithm in R by calling the C-function below. The Metropolis random walk algorithm has a standard normal density as its target density. Proposals are uniformly distributed on an interval of length 2 centred at the current value of the Markov chain.
First we make a file named `MRW.c` containing the following C code for the Metropolis random walk algorithm:

```
#include <R.h>
/************************************/
/* Metropolis random walk algorithm */
/************************************/
void MetropolisRandomWalk(int *n, double *x, double *mc){
  int i;
  double u, proposal;
  *mc = *x;
  for(i=1; i<=*n; i++){
     u = drand48();
     proposal = *(mc+i-1)+2*drand48()-1;
     if(u<= exp(0.5*(-pow(proposal,2)+pow(*(mc+i-1),2))))
      *(mc+i) = proposal;
     else *(mc+i) = *(mc+i-1);
  }

}
```

Before we can use the C code in R it should be compiled. In UNIX this is done by typing the following command at the command prompt in an x-term window:

```
 R CMD SHLIB -o MRW.so MRW.c
```

Note that this command applies to UNIX. For Windows another procedure applies.

To use the C code in R you then need the following piece of code

```
  dyn.load("MRW.so")
  Metropolis.random.walk <- function(n,x){
    .C("MetropolisRandomWalk",as.integer(n),as.double(x),
      mc=double(length=n))$mc
  }

  #* Demonstration

  x <- Metropolis.random.walk(1000,0)
  par(mfrow=c(3,1))
  plot(x,type="l")
  hist(x,freq=F)
  z <- seq(min(x),max(x),length=1000)
  lines(z,dnorm(z))
  qqnorm(x)
  abline(0,1)
```

13 Literature

An elementary exposition of Markov chains on finite state spaces are given in

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press.

A thorough but technical discussion of convergence properties of Markov chain can be found in

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer.

Metropolis-Hastings algorithms are studied in many recent books, for example,

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer.

An introduction to Bayesian inference, including many examples of applications can be found in

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC.

A review of some methods for estimating (ratios of) unknown normalising constants can be found in

Gelmam, A., Meng. X.-L. (1998). Simulating normalizing constants constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

A further development of such methods appears in

J. Møller, A.N. Pettitt, K.K. Berthelsen and R.W. Reeves. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. Research Report R-2004-02, Department of Mathematical Sciences, Aalborg University. *Submitted for publication*.