,

STK 2130: Modellering av stokastiske prosesser
Vår 2010

Course Notes and Exercises
by Nils Lid Hjort

– This version: as of 1 May 2010 –

1. The Chapman–Kolmogorov equations for Markov chains

Let $X_0, X_1, X_2, \ldots$ be a time stationary Markov chain over some finite state space, say $\{1, \ldots, k\}$. Its probability mechanism is then governed by the one-step transition probabilities

$$P_{i,j} = \Pr(X_{n+1} = j \,|\, X_n = i) \quad \text{for all } i, j = 1, \ldots, k,$$

and it is famously convenient to collect these into one matrix

$$P = \begin{pmatrix} P_{1,1} & \cdots & P_{1,k} \\ \vdots & & \vdots \\ P_{k,1} & \cdots & P_{k,k} \end{pmatrix},$$

termed the transition probability matrix. (A brief typographical comment: One finds both '$P_{i,j}$' and P_{ij}' in standard literature, i.e. with or without the comma to separate start state i and end state j, and Ross's book prefers not using the comma.)

(a) Let $\mathbf{1} = (1, \ldots, 1)^{\mathrm{t}}$ be the column vector of 1s. Show that

$$P\mathbf{1} = \mathbf{1}.$$

In particular, every transition probability matrix has 1 as an eigenvalue.

(b) Show that the two-step transition probabilities can be computed via the one-step transition probabilities, as follows:

$$P^2_{i,j} = \Pr(X_{n+2} = j \,|\, X_n = i) = \sum_{\ell} P_{i,\ell} P_{\ell,j}.$$

Note that these are identical to the elements of the matrix P^2. (In Markov chain literature one often sees n-step probabilities given as $P^{(n)}_{i,j}$ rather than as in Ross's book, where they are given simply as $P^n_{i,j}$. This is fine as long as one knows 'what is what'; do not confuse $P^{10}_{i,j}$, which I would prefer to denote as $P^{(10)}_{i,j}$, with $(P_{i,j})^{10}$.)

(c) Show more generally that the $(m+n)$-step probabilities may be computed via the m-step and n-step probabilities:

$$P^{m+n}_{i,j} = \Pr(X_{m+n} = j \,|\, X_0 = i) = \sum_{\ell} P^m_{i,\ell} P^n_{\ell,j}.$$

These are the *Chapman–Kolmogorov equations* for Markov chains (derived independently by Sydney Chapman and Андрей Колмогоров, c. 1930). In particular, we

have the important and practical result that the n-step transition probabilities are found simply by computing the n-power of the one-step probability matrix:

$$P^{(n)} = \big(P(X_n = j \,|\, X_0 = i)\big)_{i,j=1,\ldots,k} = P^n.$$

In yet other words, the laws of probability for Markov chains fit like the proverbial hand in glove with matrix power multiplication.

2. Matrix power computations

This exercise demonstrates how one may perform simple matrix power computations in `R` (calculations in other software packages may be handled similarly). Consider a Markov chain over the states 1, 2, 3, 4 and with transition probability matrix

$$P = \begin{pmatrix} 0.28 & 0.37 & 0.24 & 0.11 \\ 0.20 & 0.24 & 0.26 & 0.30 \\ 0.47 & 0.13 & 0.16 & 0.24 \\ 0.03 & 0.52 & 0.17 & 0.28 \end{pmatrix}.$$

(a) Compute P^2, P^4, P^8, P^{16}, by repeated multiplications. Note that P^{16} has 'almost converged' to a limit. (Note that matrix multiplication AB in `R` is accomplished using `A %*% B`. Asking for `A * B` gives something different, namely coordinate-wise multiplication.)

(b) Using the `array` option and a simple programming loop one may easily enough compute all P^n powers up to a given limit, say `nmax=100`. Write and edit a little programme in your computer, as follows. First you need to read in the P matrix, e.g. using

```
P <- rbind(
c(0.28, 0.37, 0.24, 0.11),
c(0.20, 0.24, 0.26, 0.30),
c(0.47, 0.13, 0.16, 0.24),
c(0.03, 0.52, 0.17, 0.28))
```

Then you create an 'array' of dimension `(k,k,nmax)`, where `k=ncol(P)` (in this case, equal to 4), and where `P[ , ,n]` is P^n:

```
Ppower <- array(data=0, dim=c(kk, kk, nmax))
Ppower[ , , 1] <- P
for (n in 2:nmax)
{ Ppower[ , , n] <- Ppower[ , , n-1] %*% P }
```

Note that there is relatively quick convergence of P^n to a limit matrix where each row is the same, say $(\pi_1, \pi_2, \pi_3, \pi_4)$. Convergence to limits may be monitored e.g. via

```
plot(Ppower[1,3, ])
```

Experiment a bit with other transition matrices of different sizes, e.g. with a 100-state Markov chain, where you generate a suitable distribution for each of the 100×100 matrix by normalising a 100-sample from some distribution. Is there 'quick convergence' of P^n to its limit?

3. Simulating Markov chains

One interesting and perhaps mildly provocative definition of 'I understand a certain stochastic process' is 'I am able to write down a programme that produces realisations from that stochastic process in my computer'. This exercise is about being able to simulate chains from a given Markov chain mechanism. A key step is the ability to draw a random sample from a given finite probability distribution, which in `R` is accomplished using

```
x <- sample(list, m, prob=problist)
```

The outcome is say $x_1, \ldots, x_m$, a total of m independent draws from the sample space `list` of possible outcomes, and with probabilities governed by `problist`. Along with a suitable `for` loop, this makes it easy to simulate a short or long Markov chain.

(a) Going back to the 4×4 matrix P of Exercise 2, let us generate say $X_1, \ldots, X_{1000}$ from that chain mechanism, with initial state say $X_0 = 3$. Write and edit a suitable programme in your computer, along these lines:

```
X0 <- 3
sim <- 1000
Xsim <- 0*(1:sim)
Xsim[1] <- sample(1:kk, 1, prob=P[X0, ])
for (n in 2:sim)
{ Xsim[n] <- sample(1:kk, 1, prob=P[Xsim[n-1], ]) }
```

(b) Check that the relative frequencies with which the chain has visited the different states 1, 2, 3, 4 reasonably well match the limits $\pi_1, \pi_2, \pi_3, \pi_4$ found on the rows of P^n for large n. Formulate a precise mathematical statement about this.

(c) Use this little machinery to experiment with other Markov chains.

4. The equilibrium distribution for Markov chains

The fundamental limit theorem for Markov chains is that there under mild regularity conditions will be an equilibrium distribution (also called stationary distribution or limiting distribution):

$$P^n_{i,i} \to \pi_i \quad \text{for all } i.$$

Ross states his Theorem 4.1 without proof. This and the following exercise reflect a slightly higher ambition level by actually providing proofs of this and some related important results. This takes a bit of linear algebra and matrix decomposition theory.

The framework worked with now is that of a Markov chain on a finite state space, say $\{1, \ldots, k\}$, with only one class (we say that the chain is irreducible), and where the chain is also aperiodic. This latter assumption is that *the period* $d(i)$ is equal to 1, where this quantity is formally defined via

$$d(i) = \text{lcd}\{n \geq 1 \colon P^n_{i,i} > 0\},$$

the least common denominator ('minste felles multiplum') of all time spans for which transition from i back to i is possible. Thus the period is 1 if e.g. $P^n_{i,i}$ is positive for $n = 4, 5$ (and many more n), since 4 and 5 have no common divisor greater than 1.

From linear algebra and matrix theory, recall that λ is an eigenvalue for P if $Pu = \lambda u$ for some non-zero u, which is then called an eigenvector. Thus $\lambda = 1$ is an eigenvalue, with $\mathbf{1}$ and eigenvector, by Exercise 1(a). All eigenvalues are roots of the equation $Q_n(\lambda) = |P - \lambda I_k|$, with $I_k = \mathrm{diag}(1, \ldots, 1)$ the identity matrix of size $k \times k$; this is a polynomial of degree k, and its roots may be complex. In `R` one computes the eigenvalues λ_j and eigenvectors u_j, via

```
lambda <- eigen(P)$values
A <- eigen(P)$vectors
```

Then $Pu_j = \lambda_j u_j$ for $j = 1, \ldots, k$, where u_j is the jth column of A. This is typically organised such that $|\lambda_1| \geq \cdots \geq |\lambda_k|$. The crucial point, in the present connection, is that we have a *spectral decomposition*,

$$A^{-1}PA = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_k).$$

(a) Show that $P^n u_j = \lambda_j^n u_j$ for each n. Use this to show that $|\lambda_j| \leq 1$ – if a λ_j has a size bigger than 1, then λ_j^n explodes, which contradicts the property that each row in P^n is a probability distribution. Attempt furthermore to show that under the given assumptions (in particular, $d(i) = 1$), each λ_j has a size strictly smaller than 1, apart from the first: $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_k|$.

(b) Show that $P = ADA^{-1}$ and that $P^n = AD^nA^{-1}$. But

$$D^n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \lambda_2^n & \cdots & 0 \\ \vdots & \vdots & \vdots & \\ 0 & 0 & \ldots & \lambda_k^n \end{pmatrix},$$

which converges to the simple diagonal matrix $E_1 = \mathrm{diag}(1, 0, \ldots, 0)$. Hence conclude that

$$P^n \to AE_1A^{-1}.$$

(c) The formula above may be used to read off the π_j limits in practice, but it is not yet clear from the mathematics that each row in the limit matrix is the same. Show however that this is necessarily the case: the limit matrix $L = AE_1A^{-1}$ has (i, j)-element

$$(AE_1A^{-1})_{i,j} = a_{i,1}a^{1,j} \quad \text{for } i, j = 1, \ldots, k,$$

where $a_{i,1}$ are the elements in the very first eigenvector, associated with $\lambda_1 = 1$, and hence this first column vector of A is proportional to $\mathbf{1}$, with all elements equal. Also, $a^{1,j}$ are the elements of the first row of A^{-1}. We have now proved that

$$P^n_{i,j} \to \pi_j \quad \text{for all } i, j,$$

with limit independent of starting point.

(d) Flowing from the above is a simple numerical recipe for finding the equilibrium distribution in practice (from a given P matrix), as

$$\pi_j = a_{1,1}a^{1,j} \quad \text{for } j = 1, \ldots, k.$$

Carry out these calculations for the case of the four-state Markov chain worked with in Exercise 2. Compare with the other easy recipe which is to compute P^n numerically for moderately large n.

(e) Show from the analysis above that there is an explicit formula for P^n, in terms of its eigenvalues and eigenvectors:

$$P^n = AE_1A^{-1} + \lambda_2^n AE_2A^{-1} + \cdots + \lambda_k^n AE_kA^{-1},$$

where E_j is the matrix which has zeroes everywhere apart from a 1 at position (j, j). We also learn from this that the convergence is *exponentially fast* (each λ_j^n goes much faster to zero than e.g. $1/n$ or even $1/n^{1000}$). Show more precisely that there is a constant c and a positive number $\rho < 1$ such that

$$|P^n_{i,i} - \pi_i| \le c\rho^n \quad \text{for all } n,$$

with neither c nor ρ depending on i; we may in fact take ρ equal to the size of the second biggest eigenvalue. This finding has consequences for how quickly a Markov chain forgets where it comes from.

5. The 'Master Theorem' for Markov chains

Via matrix decomposition methods we were able to prove the essence of Ross's Theorem 4.1 (which in his book is stated without proof). We now give a more complete 'Master Theorem', spelling out also some other consequences that are partly not well highlighted in Ross's book. The framework is that of Exercise 4, with a Markov chain on a finite state space, aperiodic and irreducible (one class, all states having period equal to 1).

(1) There is an equilibrium distribution $\pi_1, \ldots, \pi_k$ to which there is convergence

$$P^n_{i,i} \to \pi_i \quad \text{for each } i.$$

(2) There is also convergence

$$P^n_{j,i} \to \pi_i \quad \text{for all } j, i,$$

i.e. the chain 'forgets where it started'.

(3) Let $B_n(i) = \sum_{j=1}^n I(X_j = i)$ be the number of visits in state i, in the course of the first n time points. Then the relative frequencies converge,

$$A_n(i) = B_n(i)/n \to_p \pi_i \quad \text{for each } i.$$

The convergence here is 'in probability', and also takes place in the sense that

$$\mathrm{E}\, A_n(i) \to \pi_i \quad \text{and} \quad \mathrm{Var}\, A_n(i) \to 0.$$

(4) State i is recurrent and will hence be visited, re-visited, etc., with probability 1. Let $V_{i,1}, V_{i,2}, \ldots$ be the waiting times in state i; the visiting times are thus $S_{i,1} = V_{i,1}$, $S_{i,2} = V_{i,1} + V_{i,2}$, $S_{i,3} = V_{i,1} + V_{i,2} + V_{i,3}$, etc. Then the V_j have expected value

$$\xi_i = \mathrm{E}\, V_{i,j} = \mathrm{E}(S_{i,1} \,|\, X_0 = i) = 1/\pi_i,$$

i.e. the mean return time is inversely proportional to the equilibrium probability.

(5) The limiting probabilities π_i are positive and satisfy the equations

$$\pi_i = \sum_j \pi_j P_{j,i} \quad \text{for all } i,$$

along with $\sum_j \pi_j = 1$. The solution to these equations is unique.

We have already seen that (1) and (2) hold true, cf. Exercise 4, and now set out to show (3), (4), (5).

(a) Show that with any convergent sequence, also the sequence of averages converges, to the same limit:

$$c_n \to c \quad \text{implies} \quad \bar{c}_n = n^{-1} \sum_{i=1}^{n} c_i \to c.$$

An 'ε and n_0' argument is required.

(b) Regarding statement (3), this holds regardless of initial state X_0, but let us for simplicity of presentation assume that the chain starts in the same state i, i.e. $X_0 = i$. Show that

$$\mathrm{E}\, A_n(i) = n^{-1} \sum_{j=1}^{n} P_{i,i}^j$$

and then also that this sequence tends to π_i.

(c) We then need to show that

$$\mathrm{Var}\, A_n(i) = (1/n^2)\mathrm{Var}(D_1 + \cdots + D_n) \to 0,$$

where $D_j = I(X_j = i)$. The crucial step here is to bound the covariances, as

$$\mathrm{Var}(D_1 + \cdots + D_n) = \sum_{j=1}^{n} \mathrm{Var}\, D_j + 2 \sum_{j<\ell} \mathrm{cov}(D_j, D_\ell),$$

with the second sum containing $n(n-1)/2$ terms. Show in fact that

$$\mathrm{cov}(D_j, D_\ell) = P_{ii}^j (P_{ii}^{\ell - j} - P_{ii}^\ell)$$

and that

$$|\mathrm{cov}(D_j, D_\ell)| \le c\rho^{\ell - j} + c\rho^\ell,$$

with $\rho < 1$, as per Exercise 4(e). Use this to prove that $\mathrm{Var}\, A_n(i)$ tends to zero, as required.

(d) To prove (4), note first that the waiting times $V_{i,1}, V_{i,2}, \ldots$ are independent with the same distribution (i.i.d.), by the nature of the Markov chain; it returns an infinite number of times to i, and each times it returns, the process re-starts, with no further memory of the past. Thus

$$\frac{S_{i,n}}{n} = n^{-1} \sum_{j=1}^{n} V_{i,j} \to \xi_i = \mathrm{E}\, V_{i,j},$$

by the law of large numbers. Show now that

$$B_n(i) = m \quad \text{is equivalent to} \quad S_{i,m} \le n < S_{i,m+1}.$$

Deduce that

$$\frac{S_{i,B_n(i)}}{B_n(i)} \le \frac{n}{B_n(i)} < \frac{S_{i,B_n(i)+1}}{B_n(i)}.$$

Use a sandwich argument to conclude that ξ_i must be equal to $1/\pi_i$, as claimed.

(e) Use the Chapman–Kolmogorov equations

$$P^n = PP^{n-1} = P^{n-1}P$$

to deduce the two identities

$$P^n_{i,i} = \sum_j P_{i,j} P^{n-1}_{j,i} = \sum_j P^{n-1}_{i,j} P_{j,i}.$$

Then take the limit as $n \to \infty$:

$$\pi_i = \sum_j P_{i,j} \pi_i = \sum_j \pi_j P_{j,i}.$$

The first result is absolutely true and absolutely uninteresting (in the same class as '0 = 0', a theorem every mathematician regularly rediscovers). The second result is however interesting and crucial.

(f) Show that the equations of (5) can be neatly summarised in

$$\pi = \pi\, P,$$

where $\pi = (\pi_1, \ldots, \pi_k)$ is seen as a row vector. Show also that $\pi\, P^m = \pi$ for every m:

$$\pi_i = \sum_j \pi_j P^m_{j,i} \quad \text{for all } i.$$

6. The Gambler's Ruin

The so-called Gambler's Ruin model comes in various variants, but the simplest version is that laid out in Ross's Section 4.5: If $X_n = i$, with $1 \le i \le N-1$, then

$$X_{n+1} = \begin{cases} i+1 & \text{with probability } p, \\ i-1 & \text{with probability } q, \end{cases}$$

with $q = 1-p$. States $1, \ldots, N-1$ are transient while states 0 and N are absorbing. You may think of X_n as your own position in a game (perhaps the number of points, or Kroner, on some scale); if X_n ends in 0 you're dead, and if X_n ends in N you have won. Below we go through some of the results reached in Ross's book, but also put in a bit more.

(a) Consider

$$u_i = P(X_n \text{ ends in } N \,|\, X_0 = i),$$

so that $1 - u_i$ correspondingly is the probability that the game sooner or later ends in zero. Show that

$$u_i = qu_{i-1} + pu_{i+1} \quad \text{for } 1 \le i \le N-1,$$

and that $u_0 = 0$, $u_N = 1$.

(b) To solve these equations it turns out to be practical to work with *the differences*, say

$$\Delta_i = u_{i+1} - u_i \quad \text{for } i = 0, 1, \ldots, N-1.$$

Show that $\Delta_i = \alpha\Delta_{i-1}$, with $\alpha = q/p$, and use this to infer that

$$\Delta_i = \alpha^i \Delta_0 \quad \text{for } i = 0, 1, \ldots, N-1.$$

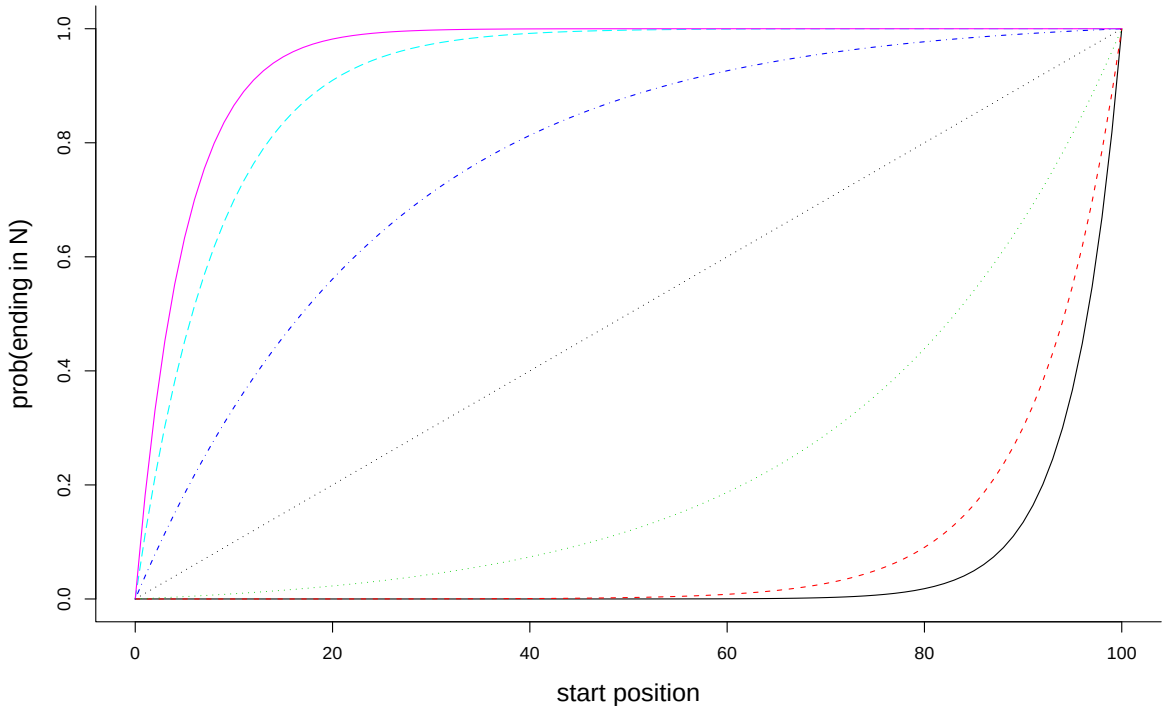

Figure 1: For $N = 100$, curves giving $u_i = P(X_n \text{ ends in } N \,|\, X_0 = i)$ are displayed as a function of start position $X_0 = i$, for p equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55.

(c) Sum all differences to find

$$u_i - u_0 = \Delta_0 + \cdots + \Delta_{i-1} = \Delta_0(1 + \cdots + \alpha^{i-1}) = \Delta_0 \frac{1-\alpha^i}{1-\alpha}.$$

Use the side conditions $u_0 = 0$ and $u_N = 1$ to find Δ_0. Conclude that

$$u_i = P(\text{you win in the end} \,|\, X_0 = i) = \frac{1 - \alpha^i}{1 - \alpha^N} \quad \text{for } i = 0, 1, \ldots, N.$$

This is valid for the case of $p \neq \frac{1}{2}$. For the symmetric case of $p = q = \frac{1}{2}$ one cannot use the same formula for $1 + \alpha + \cdots + \alpha^{i-1}$, but the arguments may be used and lead to the simpler formula $u_i = i/N$.

(d) Write a computer programme to copy Figure 1, which presents u_i as a function of start position $X_0 = i$, for the case of $N = 100$ and for values of p equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55. Note that small differences in p lead to rather drastic differences in u_i.

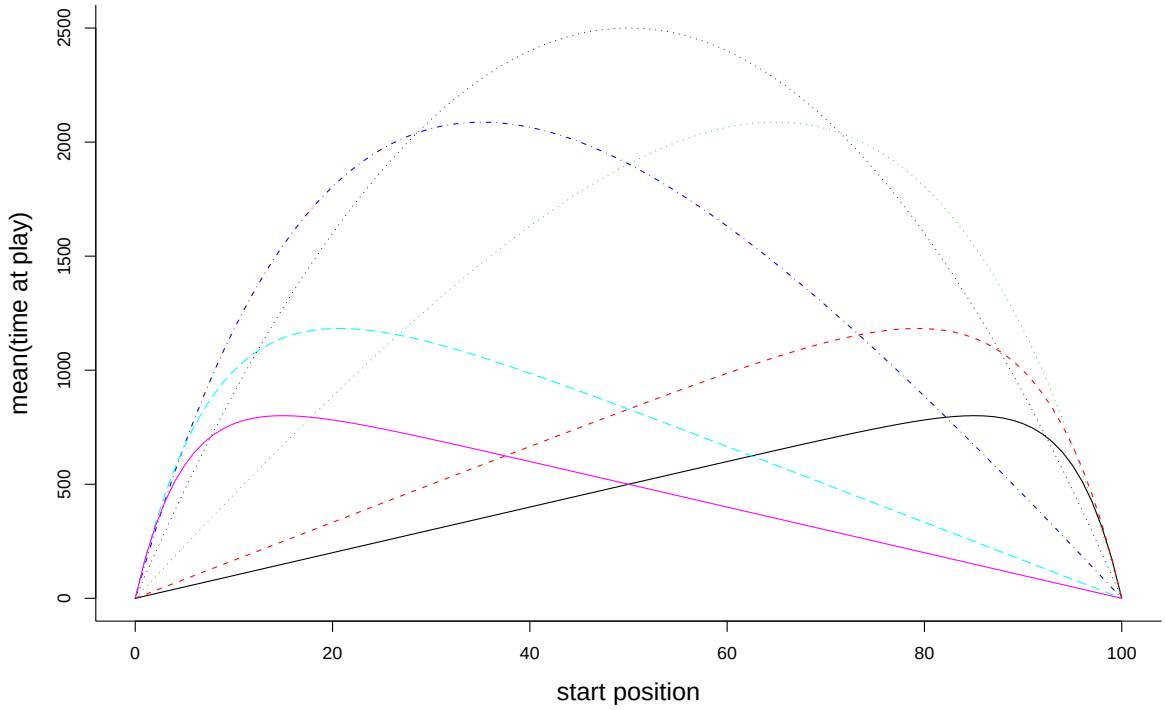

Figure 2: For $N = 100$, curves giving $v_i = \mathrm{E}(T \,|\, X_0 = i)$ are displayed as a function of start position $X_0 = i$, for p equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55. Here T is the first time X_n settles at 0 or N.

(e) We next set out to find

$$v_i = \mathrm{E}(T \,|\, X_0 = i) \quad \text{for } i = 0, 1, \ldots, N,$$

where T denotes the time the process first settles down at one of the endpoints, i.e. the first n where X_n is either 0 or N. Show that

$$v_i = 1 + q v_{i-1} + p v_{i-1} \quad \text{for } i = 1, \ldots, N - 1,$$

with side conditions $v_0 = v_N = 0$.

(f) As above it is mathematically convenient to transform the problem into equations involving the differences $\delta_i = v_{i+1} - V_i$. Show that

$$q\delta_{i-1} = 1 + p\delta_i \quad \text{for } i = 1, \ldots, N-1,$$

giving $\delta_i = \alpha\delta_{i-1} - d$, with $\alpha = q/p$ and $d = 1/p$. Show that this leads to

$$\delta_i = \alpha^i \delta_0 - d\frac{1-\alpha^i}{1-\alpha} \quad \text{for } i = 0, 1, \ldots, N-1.$$

(g) Then sum all differences to arrive at

$$v_i - v_0 = \delta_0 + \cdots + \delta_{i-1} = \delta_0 \frac{1-\alpha^i}{1-\alpha} - \frac{d}{1-\alpha}\Big(i - \frac{1-\alpha^i}{1-\alpha}\Big).$$

Using the side conditions, show that

$$\delta_0 = d\Big(\frac{N}{1-\alpha^N} - \frac{1}{1-\alpha}\Big),$$

which gives a formula for v_i, the expected time spent until the game ends.

(h) The above assumes $p \neq \frac{1}{2}$, since we were using the geometric sum formula for $1 + \alpha + \cdots + \alpha^{i-1}$. For the symmetric case $p = q = \frac{1}{2}$, use similar but actually simpler arguments to show that $v_i = i(N-i)$.

(i) Finally write a computer programme to copy Figure 2, which presents mean game length time v_i as a function of start position $X_0 = i$, for the case of $N = 100$ and for values of p equal to 0.45, 0.47, 0.49, 0.50, 0.51, 0.53, 0.55. Note again that small differences in p lead to rather drastic differences in v_i.

7. MCMC: Markov Chain Monte Carlo

[Exercises 7–8 serve as 'alternative curriculum' to Ross's Sections 4.8–4.9]

The typical modus of thinking and working regarding Markov chains, both generally and inside this course, is that one constructs a Markov chain model to emulate a phenomenon of interest (perhaps with parameters estimated from data etc.), after which one typically finds the equilibrium distribution, perhaps computes some probabilities of interest, etc. A historical footnote of just sufficient relevance to cause my mind to decide to tell you about it in this sentence without waiting for another chance is that Markov chains were invented in 1906 and that Markov himself did the world's first ever modelling of something real via Markov chains in 1913, with the data being the sequence of the first 20,000 letters of Pushkin's fantastic Евгений Онегин 1833 poem, unpoetically viewed on this occasion as a string of vowels and consonants. He did just what I mentioned above: (i) modelling something 'real' as a Markov chain, with transition probabilities $P_{i,j}$ etc.; (ii) computing the equilibrium distribution π_i etc. (along with other aspects).

In this exercise, however, we turn things the other way around – we start with the π_i distribution and then invent a Markov chain $P_{i,j}$ that happens to have the π_i as its equilibrium distribution! Got it?

This is perhaps a simple point, but turns out to have grandiose consequences, opening up doors to a space of applications where complicated probability distributions are assessed and worked with and simulated from not directly (because it may be too difficult or impossible), but via Markov chains that converge to the desired distribution: the chain $X_0, X_1, X_2, \ldots$ will after a burn-in period behave as samples from the π_i distribution. This is the MCMC (Markov Chain Monte Carlo) idea.

(a) Let $\{\pi_i : i \in I\}$ be some probability distribution on a set I (finite or countably infinite), assumed for simplicity of presentation to be positive everywhere. Assume that P is a Markov matrix of $P_{i,j}$ (that we think of as constructed or chosen after knowledge of the π_i) with the property that

$$\pi_i P_{i,j} = \pi_j P_{j,i} \quad \text{for all } i, j. \tag{B}$$

We also assume that the P chain is aperiodic and recurrent. Show that the π_i must be the equilibrium distribution of the P chain.

(b) Just to pause for a minute for you to make sure you know what this means – show that

$$P^n \to \begin{pmatrix} \pi_1 & \pi_2 & \cdots \\ \pi_1 & \pi_2 & \cdots \\ \vdots & \vdots & \cdots \end{pmatrix}$$

and that

$$P^n_{i_0,i} = P(X_n = i \,|\, X_0 = i_0) \to \pi_i \quad \text{for all } i,$$

regardless of starting point i_0. Yet another way to look at this is that if $X_0, X_1, X_2, \ldots$ is a chain drawn from P, then the distribution of X_n tends to the $\{\pi_i : i \in I\}$.

(c) This opens the door to this idea: to simulate realisations from the π_i distribution, you may set up a Markov Chain obeying condition (B) and then simulate $X_0, X_1, \ldots$ from this chain.

(d) Now let us construct such a Markov chain, with the desired property (B). Suppose there is a 'proposal distribution'

$$Q_{i,j} = P(X^{\text{cand}}_{n+1} = j \,|\, X_n = i)$$

that draws a candidate value j after the current i, and that this mechanism is symmetric (i.e. $Q_{i,j} = Q_{j,i}$ for all i, j). Assume next that we actually accept the proposed candidate j with acceptance probability

$$\text{pracc}(i,j) = \min\Big(1, \frac{\pi_j}{\pi_i}\Big).$$

In other words, such a chain has

$$X_{n+1} = \begin{cases} X_{n+1}^{\text{cand}} & \text{if this proposed value is accepted,} \\ X_n & \text{if else.} \end{cases}$$

Show that this corresponds to the Markov chain mechanism with transition probabilities

$$P_{i,j} = Q_{i,j}\text{pracc}(i,j) = Q_{i,j}\min(1, \pi_j/\pi_i) \quad \text{for } j \neq i$$

and

$$P_{i,i} = Q_{i,i} + \sum_{j \neq i} Q_{i,j}\{1 - \text{pracc}(i,j)\} = 1 - \sum_{j \neq i} Q_{i,j}\min(1, \pi_j/\pi_i).$$

(e) Show that this Markov chain indeed satisfies condition (B). (There are many other constructions and solutions; this is one type that is easy to implement and often works well.)

Markov, A.A. (1906). Распространение закона больших чисел на величины, зависящие друг от друга. [Extending the law of large numbers for variables that are dependent of each other.] Известия Физико-математического общества при Казанском университете **15** (2-я серия), 124–156.

Markov, A.A. (1913). Пример статистического исследования над текстом "Евгения Онегина", иллюстрирующий связь испытаний в цепь. [Example of a statistical investigation illustrating the transitions in the chain for the 'Evgenii Onegin' text.] Известия Академии Наук, Санкт-Петербург **7** (6-я серия), 153–162.

8. MCMC: an example

To give an illustration of the MCMC idea, consider the following moderately strange probability distribution

$$\pi_i = \frac{1}{K} f(i) = \frac{1}{K} \exp(-\beta|i)|i - \tfrac{1}{2}|^\alpha \exp\{\gamma \sin(\pi i)\} \quad \text{for } i = 0, \pm 1, \pm 2, \pm 3, \ldots,$$

for certain parameters α, β, γ, where $K = \sum_{\text{all } i} f(i)$. I invented it just now just to have an example of something we haven't seen before, and for which there is no free simulation recipe. The task is to produce a sequence of realisations from this distribution.

(a) Show that the π_i here really define a proper probability distribution, as long as $\beta > 1$. What happens if $\beta \leq 1$?

(b) Follow the MCMC recipe of the previous exercise, with symmetric proposal distribution $(\frac{1}{2}, \frac{1}{2})$ for $X_{n+1}^{\text{cand}} = j = i \pm 1$ when $X_n = i$, and accept with

$$\text{pracc} = \min(1, \pi_j/\pi_i) = \min\Big(1, \frac{f(X_{n+1}^{\text{cand}})}{f(i)}\Big).$$

Explain that X_n converges in distribution to the π_i distribution. Note that we do not need the numerical value of K to set this in motion.

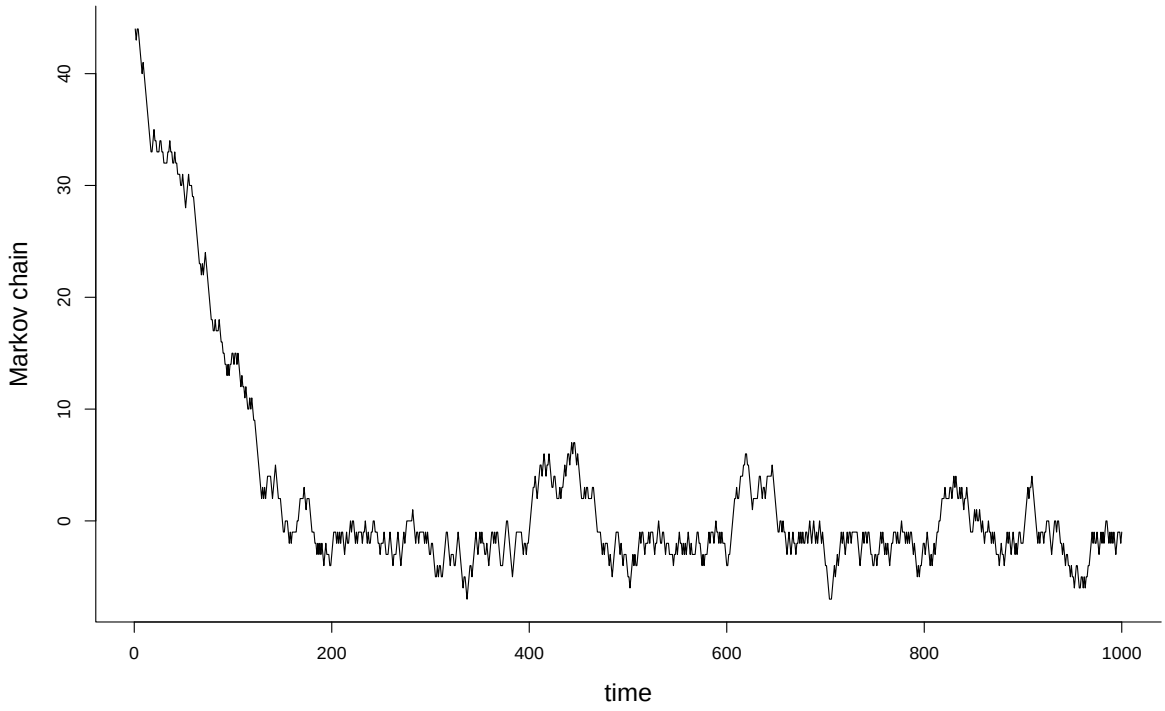

Figure 3: These are the first 1,000 steps of the MCMC, started (a bit idiosyncratically) at position $X_1 = 44$. After some 300-400 steps the chain has reached its equilibrium distribution.

(c) Implement this MCMC scheme (don't {don't do this at home}) – check the `R` programme `com3d` that I have uploaded to the course website. My illustration, see Figures 3 and 4, use $\alpha = 2.222$, $\beta = 1.111$, $\gamma = 7.777$. Check the acceptance rate for this proposal distribution (I had about 75% in my implementation). Check that the chain actually converges even if it started far away from mainstream (but requiring a longer burn-in phase; with $X_1 = 1000$ I need around 1,500 steps before equilibrium is reached).

(d) Display the histogram of the MCMC, after properly trimming away the burn-in period (taken in my illustration to be 1:1000). Compute mean, standard deviation and skewness for the π_i distribution based on this.

(e) Use the above type of MCMC to simulate from the Poisson distribution with parameter $\lambda = 10$, say. Again use the symmetric $X_{n+1}^{\text{cand}} = i \pm 1$ proposal from $X_n = i$, as long as $i \geq 1$, and with the adjustment $Q_{0,0} = \frac{1}{2} = Q_{0,1}$ for the case of $i = 0$. Afterwards, when you are sure your algorithm works, carry out simulations and some distribution assessment for the 'extended Poisson model' with

$$\pi_i = \frac{1}{K(\lambda, \gamma)} \frac{\lambda^i}{(i!)^\gamma} \quad \text{for } i = 0, 1, 2, \ldots,$$

where $\gamma = 1$ corresponds to the special Poisson case.

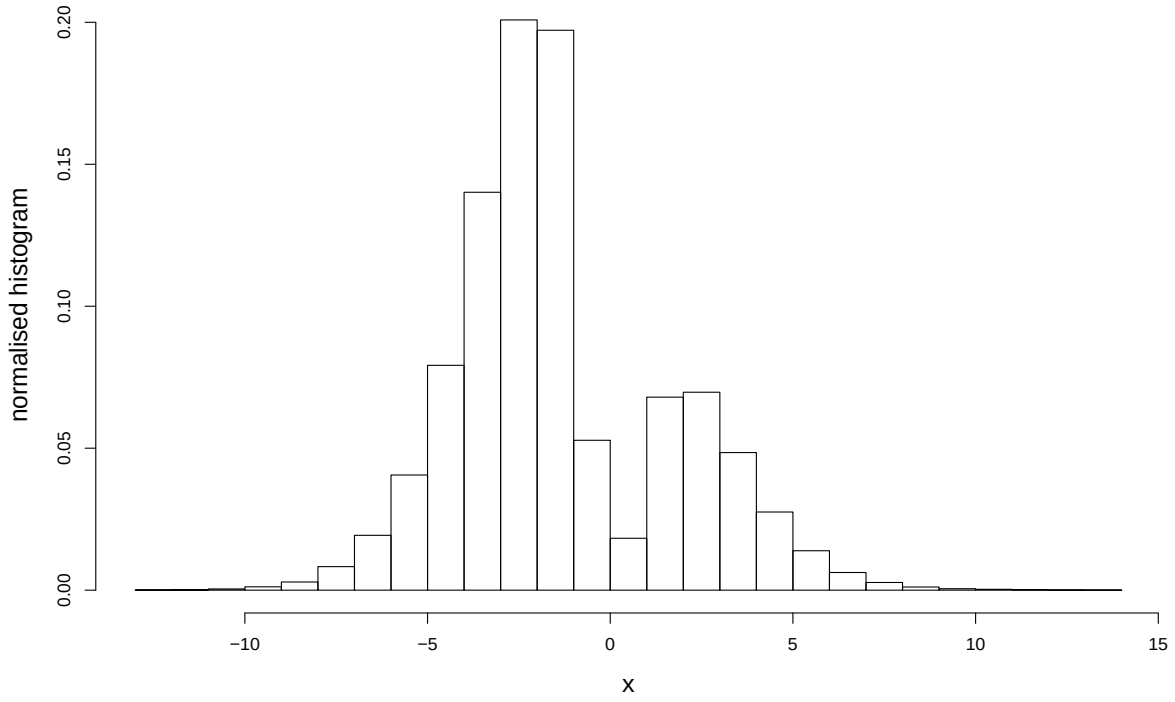

Figure 4: Histogram of the MCMC $X_{1001}, X_{1002}, \ldots, X_{\mathrm{sim}}$, after having trimmed away the first 1000 iterations as burn-in phase (I used $\mathrm{sim} = 1000 + 10^5$ *here).*

(f) The above MCMC scheme depends on the proposal distribution being symmetrical, that is, $Q_{i,j} = Q_{j,i}$ (as with proposals $i \pm 1$ being equally likely for the process that led to Figures 3 and 4 here). It is quite fruitful to allow non-symmetric proposal distribution, however, in many situations. Show that the revised scheme that uses

$$\mathrm{pracc}(i,j) = \min\Bigl(1, \frac{\pi_j Q_{j,i}}{\pi_i Q_{i,j}}\Bigr)$$

works properly, in the required sense that condition (B) of Exercise 7 is met.

9. Kolmogorov's forward and backward equations

The following may be considered a more compact version of some of the material in Ross's Sections 6.4–6.5. Consider a Markov process $X = \{X(t): t \geq 0\}$ on a finite state space, say $1, \ldots, k$, with certain transition probabilities

$$P_{i,j}(t) = \Pr\{X(t) = j \,|\, X(0) = i\} \quad \text{for } i, j = 1, \ldots, k.$$

Assume that there are transition rates (or intensities, or forces of transition) $\lambda_{i,j}$ such that

$$P_{i,j}(h) = \lambda_{i,j} h + o(h) \quad \text{for } j \neq i.$$

We write $P(t)$ for the $k \times k$ matrix of $P_{i,j}(t)$.

(a) Show that

$$P_{i,i}(h) = 1 - \lambda_i h + o(h), \quad \text{where} \quad \lambda_i = \sum_{j \neq i} \lambda_{i,j}.$$

(b) Explain why the matrix

$$R = P'(0) = \lim_{h \to 0} \{P(h) - I\}/h$$

exists, and that it consists of elements $\lambda_{i,j}$ outside the diagonal and of elements $-\lambda_1, \ldots, -\lambda_k$ along the diagonal. The R matrix is called the generator, or the infinitesimal generator, of the process.

(c) Prove the Chapman–Kolmogorov equations

$$P_{i,j}(s+t) = \sum_l P_{i,l}(s) P_{l,k}(t),$$

valid for each (i, j) and all s, t. Show that these k^2 equations neatly may be given in compact matrix form as

$$P(s+t) = P(s)P(t).$$

(d) Use $P(t+h) = P(t)P(h) = P(h)P(t)$ to show that

$$P'(t) = RP(t) = P(t)R.$$

This compact matrix equation corresponds to k^2 backward and k^2 forward differential equations. Write these in component fashion, i.e. for each $P'_{i,j}(t)$.

(e) Only in rare cases is it easy or possible to find closed form version of $P_{i,j}(t)$ based on solving these equations. There is however a neat way of expressing the solutions in general, using the matrix exponential function. Let us define

$$\exp(A) = I + A + A^2/2 + A^3/6 + A^4/24 + \cdots = \sum_{n=0}^{\infty} A^n/n!,$$

for any matrix A of size $k \times k$. Explain why this gives a finite matrix, i.e. that the series is convergent. Show that with this definition, $\exp(A+B) = \exp(A)\exp(B)$, for matrices A and B.

(f) Show that

$$P(t) = \exp(Rt) = I + \sum_{n=1}^{\infty} B^n t^n/n!$$

in fact satisfies the differential equations of point (d). This is accordingly the solution, giving at least each $P_{i,j}(t)$ numerically, in any practical situation with known (typically estimated) transition rates $\lambda_{i,j}$.

10. Brownian motion

Here we briefly study the Brownian motion process. We say that $X = \{X(t): t \geq 0\}$ is a Brownian motion, or a Wiener process, provided

(i) $X(0) = 0$;

(ii) $X(t) - X(s) \sim \mathrm{N}(0, \sigma^2(t-s))$ for each interval (s,t); and

(iii) all increments are independent, i.e. if $t_1 < \cdots < t_k$, then the increments $D_j = X(t_j) - X(t_{j-1})$ are independent.

(a) Let $s < t < u$. Assess the distribution of

$$X(u) - X(s) = \{X(t) - X(s)\} + \{X(u) - X(t)\}$$

in two different ways, from the definitions, and make sure these two viewpoints or assessments agree.

(b) Show that a tentative definition of a different process, where point (ii) is replaced by $X(t) - X(s) \sim \mathrm{N}(0, |t-s|^\alpha)$, with perhaps $\alpha = \frac{1}{2}$, would not work, as it would lead to logical incoherence – only the value $\alpha = 1$ is acceptable.

(c) Show that the random integral $A = \int_0^1 X(t)\,\mathrm{d}t$ is normal with mean zero and variance $\sigma^2/3$. Hint: start working with $A_n = n^{-1}\sum_{i=1}^n X(i/n)$, and use the fact that A_n converges in distribution to A.

(d) For a suitably high n, perhaps $n = 1000$, generate a process X_n by letting

$$X_n(t) = (\sigma/\sqrt{n}) \sum_{i: i/n \leq t} Y_i,$$

in which $Y_1, Y_2, \ldots$ are independent and standard normal. Show that X_n is 'almost' a Brownian motion – it is normal, with zero mean, and independent increments, but the variance of $X_n(t) - X_n(s)$ is not exactly, only approximately, equal to $\sigma^2(t-s)$. In the limit, though, as n grows, this X_n becomes a Brownian motion. I have used this trick to simulate 'almost' Brownian motions in Figure 5 (using $n = 10^4$, actually).

(e) Simulate some Brownian motions in your computer, and display them on your screen, using the trick above; cf. the code below. See Figure 5.

```
tval = seq(0,1,by=1/nn)
yy = c(0,rnorm(nn))
Brown = (1/sqrt(nn))*cumsum(yy)
matplot(tval, Brown, type="l")
```

(f) Show that $\mathrm{cov}\{X(s), X(t)\} = \sigma^2 \min(s,t)$, and that $\mathrm{corr}\{X(s), X(t)\} = \sqrt{s/t}$ for $s \leq t$.

(g) Show that a linear combination of Brownian motions is another Brownian motion.

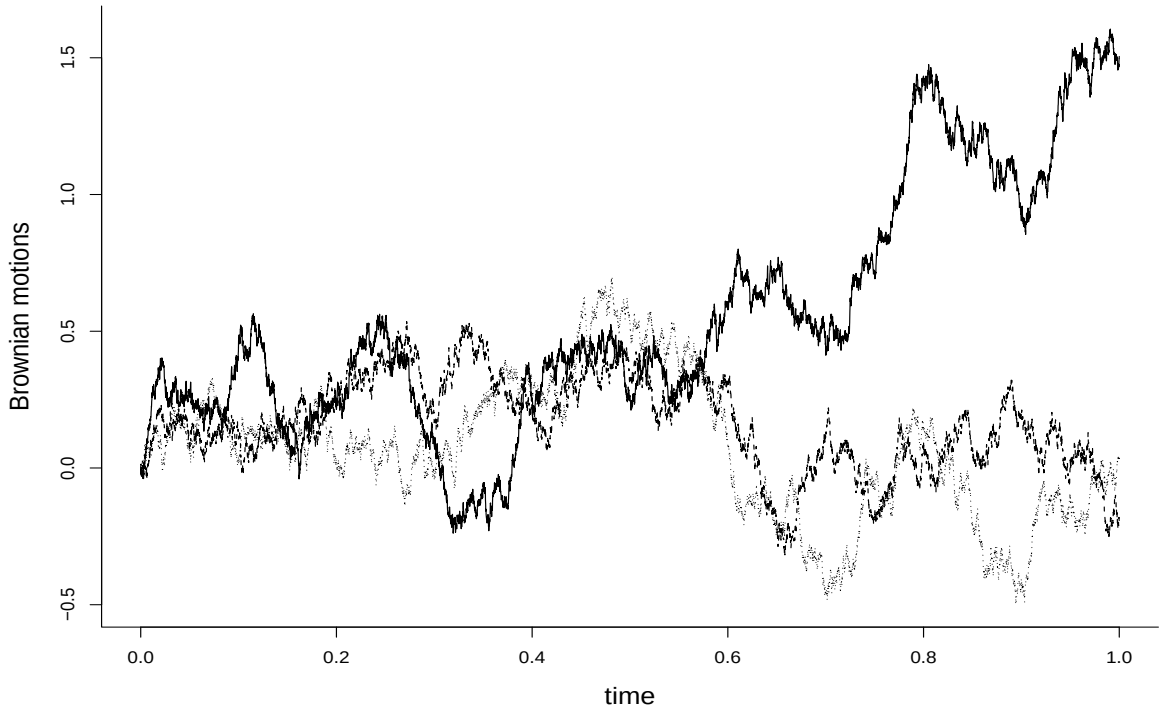

Figure 5: Three simulated Brownian motions over the time interval $[0, 1]$*, with* $\sigma = 1$*.*

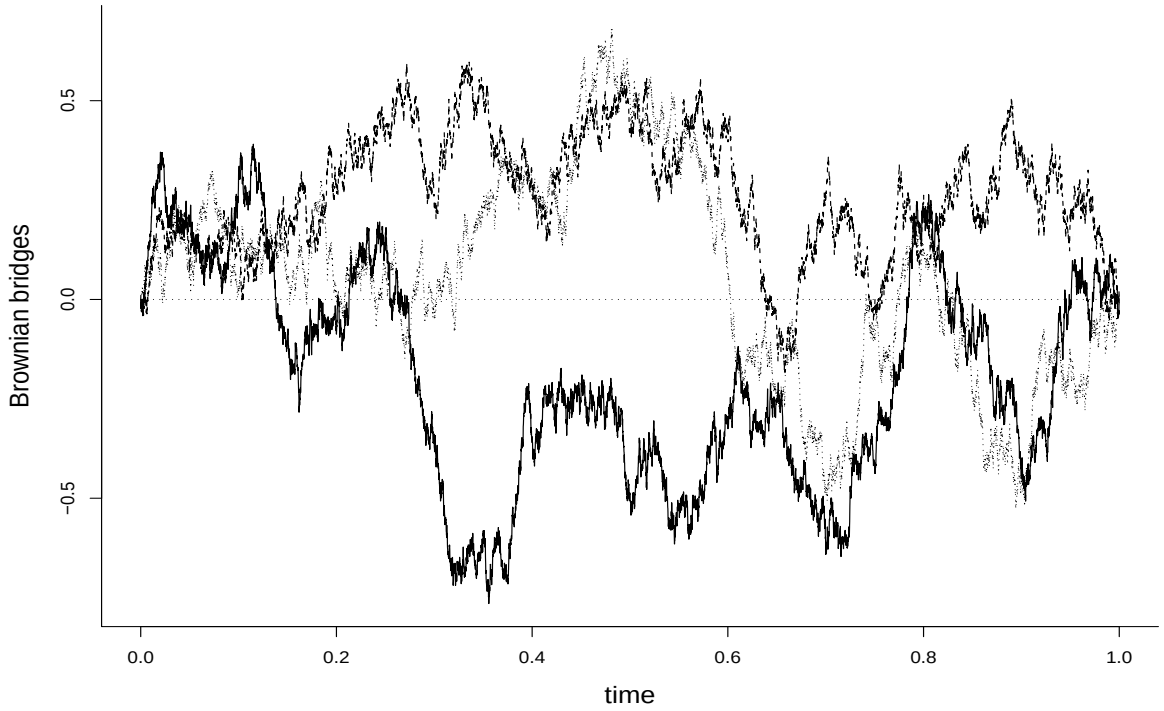

Figure 6: Three simulated Brownian bridges over the time interval $[0, 1]$*, with* $\sigma = 1$*.*

11. Some famous relatives of Brownian motion

Just like the normal distribution is a popular building tool or LEGO-brikke for constructing various other distributions and models, the Brownian motion is also such a LEGO-brikke for building other useful processes.

(a) We say that $X = \{X(t)\colon t \geq 0\}$ is a normal process if all finite-dimensional distributions are multinormal. This is equivalent to saying that all linear combinations $\sum_{j=1}^{m} c_j X(t_j)$ are normal. Argue in general that such a normal process is uniquely determined when we have specified

$$\text{mean function } m(t) = \mathrm{E}\, X(t) \quad \text{and} \quad \text{covariance function } K(s,t) = \operatorname{cov}\{X(s), X(t)\}.$$

Thus Brownian motion is the normal process with mean function $m(t) = 0$ and covariance function $K(s,t) = \sigma^2 \min(s,t)$, for example.

(b) The Brownian bridge is the process $X^0 = \{X^0(t)\colon t \in [0,1]\}$ defined by conditioning Brownian motion on ending in $X(1) = 0$. I have simulated three such in Figure 6. Show that X^0 has mean function zero and covariance function $\sigma^2 s(1-t)$ for $s \leq t$. Show in fact that the process $Y(t) = X(t) - tX(1)$ has these properties, so this is one way to represent Brownian bridges.

(c) Brownian motion becomes more and more variable as time goes by. A process with stable variation is

$$A(t) = \frac{X(t)}{\sqrt{t}}.$$

Show that this process has constant variance.

(d) A relative of this last process is the Ornstein–Uhlenbeck process, defined as

$$U(t) = \frac{X(\exp(2t))}{\exp(t)},$$

over the full real line. Show that its variance is constant and that the correlation function is $\exp(-|t-s|)$.

(e) Geometric Brownian motion is the process $Z(t) = \exp\{X(t)\}$. Find the mean, variance and covariance function for this process.