HW4, Bayesian Modelling B 2016/17

Jonathan Rougier
School of Mathematics
University of Bristol UK

In the homeworks, questions with marks are officially 'exam-style', although you can expect any homework question to appear as an exam question, unless it is explicitly 'not examinable'.

Hand in Q2, Q3, and Q6.

1. Prove that convergence in mean square implies convergence in probability. [10 marks]

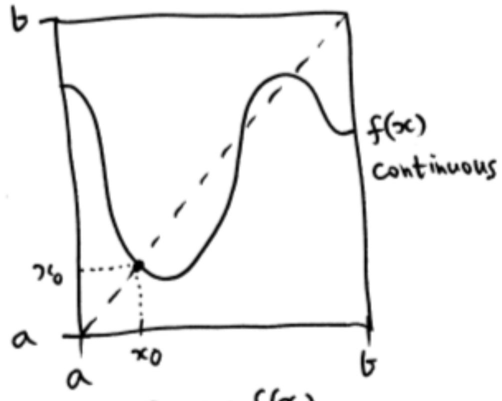
   Answer. X_n converges in probability to X exactly when $\Pr(|X_n - X|) \to 0$. X converges in mean square to X exactly when $\mathrm{E}\{(X_n - X)^2\} \to 0$. Hence

$$\begin{aligned}
\Pr(|X_n - X| < \epsilon) &= 1 - \Pr(|X_n - X| \geq \epsilon) \\
&= 1 - \Pr\{(X_n - X)^2 \geq \epsilon^2\} \\
&\geq 1 - \frac{\mathrm{E}\{(X_n - X)^2\}}{\epsilon^2} && \text{by Markov's inequality} \\
&\to 1 && \text{if } X_n \xrightarrow{\text{m.s.}} X
\end{aligned}$$

   for all $\epsilon > 0$, as required.

2. State Brouwer's Fixed Point Theorem, and prove it, using a diagram, in the 1D case. Prove that every transition matrix P has at least one fixed point. [10 marks]

   Answer. Brouwer's FPT states that if $\mathcal{X}$ is a closed bounded (i.e. compact) convex subset of Euclidean space, and $f : \mathcal{X} \to \mathcal{X}$ is a continuous function, then there is at least one $x_0 \in \mathcal{X}$ for which $x_0 = f(x_0)$. Here is the diagram.



Let P be a $d \times d$ transition matrix, and let $\mathcal{X} = \{x \in \mathbb{R}^d : x_i \geq 0, \sum_i x_i = 1\}$. First, check that $\mathcal{X}$ is a convex subset of $\mathbb{R}^d$. Let $x, y \in \mathcal{X}$, $\alpha \in [0,1]$, and define $z := \alpha x + (1-\alpha)y$. Then $z_i \geq 0$, and

$$z^T \mathbf{1} = [\alpha x^T + (1-\alpha)y^T]\mathbf{1} = \alpha x^T \mathbf{1} + (1-\alpha)y^T \mathbf{1} = \alpha + (1-\alpha) = 1,$$

showing that $\sum_i z_i = 1$, and the $z \in \mathcal{X}$, as required.

Now define $f : x \mapsto P^T x$. It is easy to check that f is a continuous function (no need to do this!). Now show that $f(\mathcal{X}) \subset \mathcal{X}$. Let $y = f(x) = P^T x$, for $x \in \mathcal{X}$. First, $P \geq 0$ and $x \geq 0$, and so $y_i \geq 0$. Second,

$$\mathbf{1}^T y = \mathbf{1}^T (P^T x) = (P\mathbf{1})^T x = \mathbf{1}^T x = 1$$

and so $\sum_i y_i = 1$. Hence $y \in \mathcal{X}$. So f is continuous and $f : \mathcal{X} \to \mathcal{X}$, satisfying the conditions of the FPT. Thus there exists an x_0 satisfying $x_0 = P^T x_0$, or $x_0^T = x_0^T P$, which defines x_0 as a stationary distribution for P.

3. Suppose that $X_0, X_1, \ldots$ are IID. What does the transition matrix look like? What is the stationary distribution? Is it unique? [10 marks]

   Answer. In this case p_{ij}, the probability of moving from i to j, is the same for all i. Hence $P = \mathbf{1}\mu^T$ for some probability vector μ. Suppose that π is any stationary distribution. Then

$$\pi^T = \pi^T P = \pi^T (\mathbf{1}\mu^T) = (\pi^T \mathbf{1})\mu^T = \mu^T.$$

   So $\pi = \mu$ is the unique stationary distribution.

4. Let P and Q be transition matrices with a common stationary distribution π.

   (a) Prove that if $0 \leq \alpha \leq 1$, then $R = \alpha P + (1-\alpha)Q$ also has stationary distribution π. Explain how sampling from R can be implemented in practice. [10 marks]

   Answer. We start by checking that R is indeed a transition matrix as claimed. Clearly $r_{ij} \geq 0$, and also

$$R\mathbf{1} = [\alpha P + (1-\alpha)Q]\mathbf{1} = \alpha P\mathbf{1} + (1-\alpha)Q\mathbf{1} = \alpha\mathbf{1} + (1-\alpha)\mathbf{1} = \mathbf{1}.$$

Now we show that $\pi^T R = \pi^T$:

$$\pi^T R = \pi^T[\alpha P + (1-\alpha)Q] = \alpha\pi^T P + (1-\alpha)\pi^T Q = \alpha\pi^T + (1-\alpha)\pi^T = \pi^T,$$

as required.

Imagine tossing a coin Y with outcomes $\{h, t\}$, where $\Pr(Y = h) = \alpha$, and the tosses are independent of each other and of the chain. Suppose that at each step we toss this coin, and select transition matrix P for $Y = h$, and Q for $Y = t$. Then

$$\begin{aligned}\Pr(X_{t+1} = j \mid X_t = i) &= \sum_{y\in\{h,t\}} \Pr(X_{t+1} = j \mid X_t = i, Y = y) \cdot \Pr(Y = y) \quad \text{by the LTP} \\ &= p_{ij} \cdot \alpha + q_{ij} \cdot (1-\alpha) = r_{ij},\end{aligned}$$

where LTP is the Law of Total Probability. So R represents a random choice between P and Q.

(b) Prove that $R = PQ$ also has stationary distribution π. Explain how sampling from R can be implemented in practice. [10 marks]

Answer. Again, we start by checking that R is indeed a transition matrix as claimed. Clearly $r_{ij} \geq 0$, and also

$$R\mathbf{1} = (PQ)\mathbf{1} = P(Q\mathbf{1}) = P\mathbf{1} = \mathbf{1},$$

as required. Now we show that $\pi^T R = \pi^T$:

$$\pi^T R = \pi^T(PQ) = (\pi^T P)Q = \pi^T Q = \pi^T,$$

as required.

Imagine first moving the chain according to P, and then moving the chain according to Q. In this case, the two-step transition probabilities are

$$\begin{aligned}\Pr(X_{t+2} = j \mid X_t = i) &= \sum_k \Pr(X_{t+2} = j \mid X_t = i, X_{t+1} = k) \cdot \Pr(X_{t+1} = k \mid X_t = i) \quad \text{LTP} \\ &= \sum_k \Pr(X_{t+2} = j \mid X_{t+1} = k) \cdot \Pr(X_{t+1} = k \mid X_t = i) \\ &= \sum_k q_{kj} \cdot p_{ik} = (PQ)_{ij} = R_{ij}.\end{aligned}$$

So R represents a two-step Markov chain where the first move is according to P, and the second according to Q.

5. Suppose that a transition matrix P can be written in terms of two smaller transition matrices P_1 and P_2 as

$$P = \begin{pmatrix} P_1 & \mathbf{0} \\ \mathbf{0} & P_2 \end{pmatrix}.$$

Show that P is not irreducible, and that it has multiple stationary distributions. [10 marks]

Answer. P is irreducible if it is possible to get from any i to any j in a finite number of steps. However, in this case it is impossible for a chain that starts in the first block to visit the second, and impossible for a chain that starts in the second block to visit the first. Therefore P is not irreducible.

Let μ_1 and μ_2 be stationary distributions of P_1 and P_2. Let $\pi^T = [\alpha\mu_1^T, (1-\alpha)\mu_2^T]$ for some $\alpha \in [0,1]$. Note that $\pi_i \geq 0$ and

$$\pi^T \mathbf{1} = [\alpha\mu_1^T, (1-\alpha)\mu_2^T]\mathbf{1} = \alpha\mu_1^T \mathbf{1}_1 + (1-\alpha)\mu_2^T \mathbf{1}_2 = \alpha + (1-\alpha) = 1.$$

Hence π is a probability vector. Now compute

$$\pi^T P = [\alpha\mu_1^T, (1-\alpha)\mu_2^T] \begin{pmatrix} P_1 & \mathbf{0} \\ \mathbf{0} & P_2 \end{pmatrix} = \begin{pmatrix} \alpha\mu_1^T P_1 & (1-\alpha)\mu_2^T P_2 \end{pmatrix} = [\alpha\mu_1^T, (1-\alpha)\mu_2^T] = \pi^T.$$

Hence π is a stationary distribution for P. However, the value α was arbitrary, and hence there are an uncountable number of stationary distributions for P, one for each element of $\alpha \in [0,1]$.

6. Consider the transition matrix

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

How would you demonstrate that this matrix is irreducible and aperiodic? How could you approximate its unique stationary distribution using R? [10 marks]

Answer. A sufficient condition for P to be irreducible and aperiodic is $p_{ij}(n) > 0$ for all ij for some n, where $p_{ij}(n)$ is the probability of moving from i to j in exactly n steps. But as $p_{ij}(n) = [P^n]_{ij}$, we can check for this by computing large powers of P. Here is some R code which you can paste in:

```
P <- matrix(c(
  0, 1, 0,   0,
  0, 0, 0.5, 0.5,
  0, 0, 0,   1,
  1, 0, 0,   0), byrow = TRUE, nrow = 4)
Q <- diag(1, 4) # Id matrix
for (i in 1:100)
  Q <- P %*% Q # Q = P^100
print(all(Q > 0)) # TRUE
```

If you inspect Q you will see:

```
> Q
          [,1]      [,2]      [,3]      [,4]
[1,] 0.2857144 0.2857157 0.1428568 0.2857131
[2,] 0.2857131 0.2857144 0.1428579 0.2857146
[3,] 0.2857135 0.2857128 0.1428580 0.2857157
[4,] 0.2857157 0.2857135 0.1428564 0.2857144
```

which looks a lot like $P^{100} = Q \approx \mathbf{1}\,(2,2,1,2)/7$. We should not be surprised because we have proved that $|p_{ij}(n) - \pi_j| \le \lambda^n \cdot c$ for some $\lambda \in (0,1)$. Thus all the rows of P^n are converging to the same unique stationary distribution; in this case, the distribution $\pi^T = (2,2,1,2)/7$. So the crude way to find π is just to raise P to a very large power, and use the first (or any) row.