pfsl17(15).tex
Lecture 17. 23.11.2015

V. MARKOV CHAINS

Markov chains are our first examples of *stochastic processes* (Ch. VI below: stochastic = random; process = unfolding with time). For most of Ch. V, both time and state will be *discrete*. Because the essence of the Markov property is that all one needs to know is where one is rather than how one got there so far as predicting the future is concerned, the probabilistic structure is determined by the *transition mechanism* (from 'where now' to 'where next time'). In the discrete case, this is expressed by a *matrix*. So the more elaborate machinery of Ch. VI is not needed here. Doing Markov chains now has two advantages:
(i) it gives earlier familiarity with Markov chains, important in Statistics for *Markov chain Monte Carlo* (MCMC);
(ii) when we begin general stochastic processes (Ch. VI), we already have a rich store of examples to hand.

1. Notation and Examples.

Recall that a *Markov process* in discrete time is a stochastic process $X = (X_n)$ with

$$P(X_n \in A | X_m, B) = P(X_n \in A | X_m)$$

for time $m < n$, where B denotes an event depending on values of X for time $< m$ (think of m here as the present, $n > m$ as in the future, and B as in the past). In words: the conditional probability of the future given the present and the past is that same as that of the future given the present only. That is, where you are is all that counts, not how you got there.

The values taken by the process $X = (X_n)$ may be discrete or continuous. The discrete case is easier, so we begin with it. The X-values form a finite or countable set, $\{x_n\}$. It is usually possible to disregard the precise values x_n and replace them by *labels*, n. Usually the label set will be the natural numbers $\mathbb{N}$, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$, $\mathbb{N}_n := \{1, \dots, n\}$, $\mathbb{N}_n^0 := \mathbb{N}_n \cup \{0\}$, or $\mathbb{Z}$, depending on context. In general, write E_k for state k.
Example: Simple random walk on $\mathbb{Z}$: the label set is $\mathbb{Z}$, and so is the value set.

It is conventional to refer to a Markov process with both time and state discrete as a *Markov chain*. To describe such a Markov chain, we need the *transition probabilities* $P(X_{n+1} = j | X_n = i)$. We confine ourselves here,

for simplicity, to the most important special case, when these transition probabilities are *stationary* – do not depend on n:

$$p_{ij} := P(X_{n+1} = i | X_n = j) = P(i \to j),$$

in an obvious notation. We assemble these transition probabilities (p_{ij}) into a *transition (probability) matrix*

$$P := (p_{ij})$$

(the matrix whose (i, j) element is p_{ij}). Similarly, we define the *n-step* transition probabilities

$$p_{ij}^{(n)} := P(X_{m+n} = j | X_m = i) = P(i \to j \text{ in } n \text{ steps})$$

(by stationarity, this does not depend on m), and form the *n-step transition (probability) matrix* $P(n) := (p_{ij}^{(n)})$:

$$P^{(n)} := (p_{ij}^{(n)}).$$

Note. 1. Here i and j run through the possible states of the chain. Usually, these will be labelled $\{1, 2, \ldots, N\}$ in the finite case, $\{1, 2, \ldots\}$ in the (countably) infinite case. It pays to keep the notation flexible, to cover both cases.
2. Much of what we will cover applies to both finite and infinite chains. Finite chains have certain special properties (VII.4). We are also much more familiar with finite matrices than with infinite ones. Bear in mind that in the infinite case, matrices and sums over states are both infinite.
3. A matrix is called *stochastic* if its entries are non-negative and sum to 1. The transition probability matrix $P = (p_{ij})$ of a Markov chain is stochastic, as

$$p_{ij} = P(i \to j) \geq 0,$$

$$\sum\nolimits_j p_{ij} = \sum\nolimits_j P(i \to j) = 1$$

– as the chain has to go somewhere. Infinite matrices are difficult in general, but stochastic matrices are much simpler, and are often no harder to handle than finite matrices.
Theorem (Chapman-Kolmogorov equations).

$$P^{(n)} = P^n :$$

the n-step transition probability matrix is the nth matrix power of the (1-step) transition probability matrix.
Proof. For $n = 2$:

$$\begin{aligned} p_{ij}^{(2)} &= P(i \to j \text{ in 2 steps}) \\ &= \sum\nolimits_k P(i \to k \to j) \\ &= \sum\nolimits_k P(i \to k \text{ on first step}).P(k \to j \text{ on second step} | i \to k \text{ on first step}) \\ &= \sum\nolimits_k P(i \to k).P(k \to j), \end{aligned}$$

using the Markov property in the second term. This says that

$$p_{ij}^{(2)} = \sum\nolimits_k p_{ik} p_{kj},$$

the (i, j) element of the second matrix power P^2.

For the general case we can use induction on the power n. Alternatively, we can argue as follows. The probability of going from i to j in n steps is, summing over all possible paths from i to j in n steps,

$$p_{ij}^{(n)} = \sum\nolimits_{k_1,\ldots,k_{n-1}} P(i \to k_1).P(k_1 \to k_2 | i \to k_1).P(k_2 \to k_3 | i \to k_1 \to k_2)$$

$$\ldots P(k_{n-1} \to j | i \to k_1 \to \ldots \to k_{n-1}),$$

by iterated conditional expectation. Using the Markov property,, the RHS simplifies to

$$p_{ij}^{(n)} = \sum\nolimits_{k_1,\ldots,k_{n-1}} P(i \to k_1).P(k_1 \to k_2).P(k_2 \to k_3) \ldots P(k_{n-1} \to j).$$

The LHS is the (i, j) element of $P^{(n)}$, while the RHS is the (i, j) element of the nth matrix power P^n of P. Since this holds for all i and j, the two matrices are equal, as required. //

This result is vital. It shows one of the great advantages of Markov chain theory – that it is perfectly adapted to the theory of matrices and Linear Algebra, which is very well developed.
Note. The result is named after Sydney CHAPMAN (1888-1970), an English applied mathematician (paper of 1928) and Kolmogorov (paper of 1931).
Initial distribution. Suppose that the position at time $t = 0$ is random, with

$$p_i := P(X_0 = i).$$

Form the *row-vector*

$$p := (p_0, p_1, \ldots).$$

Then

$$\begin{aligned} P(X_n = j) &= \sum_i P(X_n = j \ \& \ X_0 = i) \\ &= \sum_i P(X_0 = i) P(X_n = j | X_0 = i) \\ &= \sum_i p_i p_{ij}^{(n)} \\ &= (pP^{(n)})_j. \end{aligned}$$

That is, the *row-vector* $pP^{(n)} = pP^n$ gives the distribution of the chain at time n.

Note. 1. Because it is natural to specify where we are at one time (at i with probability p_i), and then where we go to next (go from i to j with probability p_{ij}), it is *row*-vectors, rather than column-vectors, that are more useful in Markov chain theory.

This is worth bearing in mind, as in Linear Algebra the convention is often adopted that vectors are *column*-vectors (by default – i.e., unless otherwise specified), in which case one needs to use a transpose sign (A^T denotes the transpose of a matrix A) to obtain a row-vector. This is actually unnecessary here: vectors, row or column, are special cases of matrices, and it is better not to clutter things up with unnecessary transpose signs.

2. Precisely for this reason, one sometimes sees p_{ji} used for what we call p_{ij}, as in e.g. [M], Ch. 3: Markov processes.

Beware of this if using this otherwise excellent book!

Stationary distribution.

Suppose that the initial distribution π satisfies the linear equations

$$\pi P = \pi. \qquad (SD)$$

Then by above, its distribution after one step is $\pi P = \pi$. Similarly, its distribution after n steps is

$$\pi P^{(n)} = \pi P^n = \pi P.P^{n-1} = \pi P^{n-1} = \pi P^{n-2} = \ldots = \pi P = \pi :$$

the distribution stays the same for all time. Such a distribution is called *stationary*, or *invariant*, or an *equilibrium distribution*. We shall return to such distributions later, when we shall see that they are (under broad conditions) *limiting distributions*, to which the chain settles down as time passes.