last revised: 26 January 2009

1 Markov Chains

A Markov chain process is a simple type of stochastic process with many social science applications. We'll start with an abstract description before moving to analysis of short-run and long-run dynamics. This chapter also introduces one sociological application – social mobility – that will be pursued further in Chapter 2.

1.1 Description

Consider a process that occupies one of n possible states each period. If the process is currently in state i, then it will occupy state j in the next period with probability $P(i,j)$. Crucially, transition probabilities are determined entirely by the current state – no further "history dependence" is permitted – and these probabilities remain fixed over time. Under these conditions, this process is a Markov chain process, and the sequence of states generated over time is a Markov chain.

To restate this somewhat more formally, suppose that the possible states of the process are given by the finite set

$$S = \{1, 2, \ldots, n\}.$$

Let $s_t \in S$ denote the state occupied by the process in period $t \in \{0, 1, 2, \ldots\}$. Further suppose that

$$Prob(s_{t+1} = j \mid s_t = i) \;=\; P(i,j)$$

where $P(i,j)$ is parameter fixed for each pair of states $(i,j) \in S \times S$. As this notation makes explicit, the probability of moving to each state j in period $t+1$ depends only on the state i occupied in period t. Given these assumptions, the process is a Markov chain process, and the sequence $(s_0, s_1, s_2, \ldots)$ is a Markov chain.

The parameters of a Markov chain process can thus be summarized by a *transition matrix*, written as

$$P = \begin{bmatrix} P(1,1) & P(1,2) & \ldots & P(1,n) \\ P(2,1) & P(2,2) & \ldots & P(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ P(n,1) & P(n,2) & \ldots & P(n,n) \end{bmatrix}.$$

Of course, because the elements of this matrix are interpreted as probabilities, they must be non-negative. That is,

$$P(i,j) \geq 0 \quad \text{for all } i,j \in S.$$

Further, each row of P must be a *probability vector*, which requires that

$$\sum_{j \in S} P(i,j) \;=\; 1 \quad \text{for all } i \in S.$$

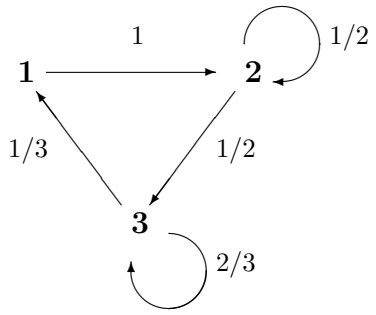
Consequently, while the transition matrix has n^2 elements, the Markov chain process has only $n(n-1)$ free parameters.

To make this description more concrete, consider an example (drawn from Kemeny et al, 1966, p 195). A Markov process has 3 states, with the transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 1/3 & 0 & 2/3 \end{bmatrix}.$$

In words, if the process is currently in state 1, it always transitions to state 2 in the next period. If the process is in state 2, it remains in state 2 with probability 1/2, and transitions to state 3 with probability 1/2. Finally, if the process is in state 3, it remains in state 3 with probability 2/3, and moves to state 1 with probability 1/3.

Beyond the matrix specification of the transition probabilities, it may also be helpful to visualize a Markov chain process using a *transition diagram*. Below is the transition diagram for the 3×3 transition matrix given above.



The nodes of the transition diagram correspond to the possible states of the process (labeled in boldface); the directed edges indicate possible transitions between states. If a transition is possible from state i to state j, the directed edge from node i to node j is labeled with the probability $P(i,j)$. A loop (an edge from some node to itself) indicates the possibility that the process continues to occupy the same state next period. By convention, no edge is drawn from node i to node j if $P(i,j) = 0$.

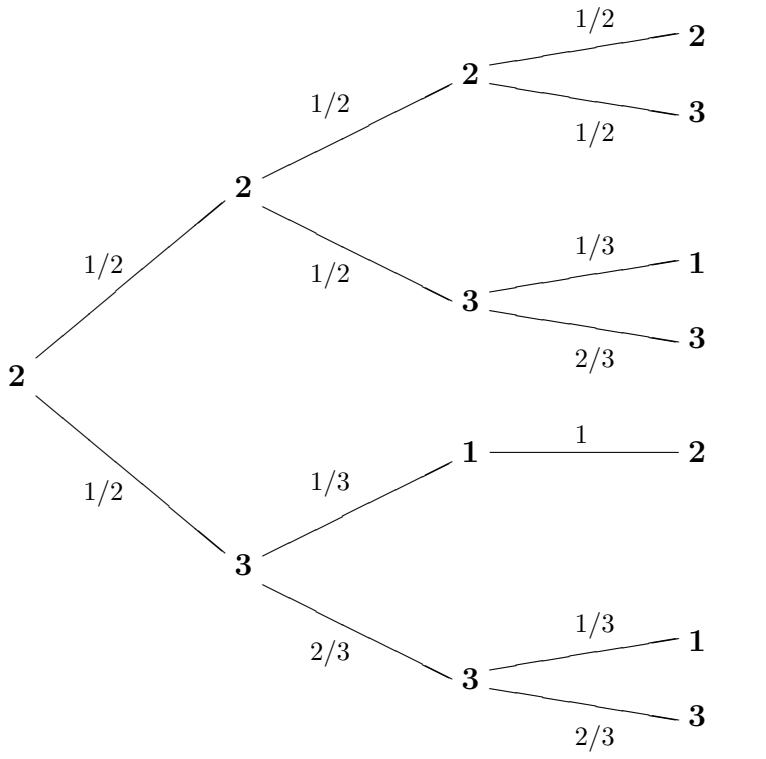
1.2 Analysis using a probability tree

Having specified a Markov chain process, we might now tackle the following sorts of questions: If the process begins in state i in period 0, what is the probability that

the process will be in state j in period 1? in period 2? in period t? in the long run, as t becomes very large? We'll soon learn simple matrix methods for answering these questions. But it might be instructive first to approach these questions by drawing a *probability tree diagram*.

The initial node of a probability tree (aptly called the "root") is given by the state i occupied in period 0. The first set of "branches" are given by the edges indicating possible transitions from state i. These edges are labeled with the probabilities $P(i, j)$ just as in the transition diagram. From each new node – each state j that the process may occupy in period 1 – we may then draw another set of branches indicating possible transitions from that state. This creates another set of nodes – every state j that the process may occupy in period 2 – which would each sprout another set of branches. Continuing in this fashion, we could (at least in principle) construct a probability tree to describe the future of the Markov chain over the first t periods (for any number of periods t). A particular Markov chain generated by the process corresponds to a single path through the tree, starting at the initial (period 0) node and ending at one of the terminal (period t) nodes.

To illustrate, let's return to the 3-state example above, and suppose the process starts in state 2 in period 0. The probability tree for the next 3 periods is shown below.



Having assumed (arbitrarily) that the process started in state 2 in period 0, we could also construct similar diagrams with state 1 or state 3 as the initial node. (See Kemeny et al, 1966, p 196 for the diagram starting from state 1. I'll leave the diagram for state 3 as an exercise.)

Using the probability tree diagram, we can now begin to answer the questions posed above. By assumption, the process starts in state 2 in period 0. Following one of the two branches from this initial node, the process then occupies either state 2 (with probability 1/2) or state 3 (with probability 1/2) in period 1. Of course, we could have obtained those probabilities directly from the second row of the transition matrix without the aid of the tree diagram. But for period 2, the diagram becomes more useful, showing that the process has reached state 1 with probability

$$(1/2)(1/3) \;=\; 1/6,$$

has reached state 2 with probability

$$(1/2)(1/2) \;=\; 1/4,$$

and has reached state 3 with probability

$$(1/2)(1/2) + (1/2)(2/3) \;=\; 7/12.$$

Importantly, as reflected in this last computation, there are two different paths that the process could have taken from the initial state to the destination state. In particular, the process could have transitioned from state 2 to state 3 to state 3 (with probability $(1/2)(1/2) = 1/4$) or it could have transitioned from state 2 to state 2 to state 3 (with probability $(1/2)(2/3) = 1/3$). To obtain the probability that the process has transitioned from state 2 to state 3 through *any* path, we add these two probabilities together (hence $1/4 + 1/3 = 7/12$). It is also important to note that

$$1/6 + 1/4 + 7/12 \;=\; 1,$$

which reflects the fact that the process must occupy *some* state in period 2. More formally, we say that the probability vector

$$\begin{bmatrix} 1/6 & 1/4 & 7/12 \end{bmatrix}$$

constitutes a *probability distribution* over states of the process in period 2.

In a similar way, we can use the tree diagram to obtain the probability distribution over states in period 3. The process has reached state 1 with probability

$$(1/2)(1/2)(1/3) + (1/2)(2/3)(1/3) \;=\; 7/36,$$

has reached state 2 with probability

$$(1/2)(1/2)(1/2) + (1/2)(1/3)(1) \;=\; 7/24,$$

and has reached state 3 with probability

$$(1/2)(1/2)(1/2) + (1/2)(1/2)(2/3) + (1/2)(2/3)(2/3) \;=\; 37/72.$$

The computations reflect 2 possible paths (of length 3) from state 2 to state 1, 2 possible paths (of length 3) from state 2 to state 2, and 3 possible paths (of length 3) from state 2 to state 3.

1.3 Analysis using matrix algebra

While probability trees offer an intuitive first approach for analyzing Markov chains, it is obvious that this method would quickly become impractical as the number of states (n) or time periods (t) becomes large. Fortunately, there is a simple matrix approach involving iterated multiplication of the transition matrix. Raising the matrix P to the power t, we obtain the matrix P^t. Let element (i, j) of this matrix be denoted as $P^t(i, j)$. This element has the following interpretation: $P^t(i, j)$ is the probability that a process which begin in state i in period 0 will occupy state j in period t. Consequently, given initial state i, the probability distribution over states after t periods is given by the ith row of P^t matrix.

Having so quickly summarized the matrix approach, perhaps an illustration would be useful. Returning again to the 3-state example, we can use Matlab to determine the (probabilistic) future of the chain after 2, 3, 4, 5, 10, or 100 periods.

```
>> P = [0 1 0; 0 1/2 1/2; 1/3 0 2/3]   % transition matrix

P =
       0      1.0000       0
       0      0.5000    0.5000
    0.3333       0      0.6667

>> P^2   % period 2

ans =
     0        0.5000    0.5000
    0.1667    0.2500    0.5833
    0.2222    0.3333    0.4444

>> P^3   % period 3

ans =
    0.1667    0.2500    0.5833
    0.1944    0.2917    0.5139
    0.1481    0.3889    0.4630
```

```
>> P^4   % period 4

ans =
    0.1944    0.2917    0.5139
    0.1713    0.3403    0.4884
    0.1543    0.3426    0.5031

>> P^5   % period 5

ans =
    0.1713    0.3403    0.4884
    0.1628    0.3414    0.4958
    0.1677    0.3256    0.5067

>> P^10   % period 10

ans =
    0.1666    0.3335    0.4998
    0.1666    0.3334    0.5000
    0.1667    0.3332    0.5001

>> P^100   % period 100

ans =
    0.1667    0.3333    0.5000
    0.1667    0.3333    0.5000
    0.1667    0.3333    0.5000
```

To reconcile these computations with our probability tree analysis, which assumed that the process was initially in state 2, consider the second row of the P^2 and P^3 matrices. Consistent with our previous analysis of period 2, we find

$$P^2(2,1) = .1667,\ P^2(2,2) = .25,\ \text{and}\ P^2(2,3) = .5833.$$

And consistent with our analysis of period 3, we find

$$P^3(2,1) = .1944,\ P^3(2,2) = .2917,\ \text{and}\ P^3(2,3) = .5139.$$

But moving beyond the results of our probability tree analysis, we can learn much more from the preceding matrix computations. Suppose that the process was initially in state 1. To use the probability tree approach, we would need to draw a new tree (with state 1 as the root). But given these computations, we can simply inspect the first row of the matrix P^t to determine the probability distribution over states in period t. Similarly, without drawing a new probability tree with state 3 as the root, we can inspect the third row of P^t to determine the probability distribution over states in period t. Furthermore, while the probability tree method is impractical when the number of periods is large, it is trivial (using matrix algebra software) to compute probability distributions for any number of periods t.

Having seen how to use matrix algebra, we should also consider why this approach works. The argument is already implicit in our reconciliation of the matrix computations with the probability tree diagram. But restated somewhat more formally, the argument proceeds by induction. Consider element (i, j) of the P^2 matrix. Because this element is found by multiplying row i of P by column j of P, we obtain

$$P^2(i, j) = \sum_{k \in S} P(i, k)P(k, j).$$

The kth term of this sum gives the probability that the process transitions from state i to state k to state j over the course of two periods. Summing over all k, we obtain the probability that the process moves from state i in period 0 to state j in period 2 through any intermediate state in period 1. Now consider element (i, j) of the P^3 matrix. Because $P^3 = P^2P$, this element can be found by multiplying row i of P^2 by column j of P, and we thus obtain

$$P^3(i, j) = \sum_{k \in S} P^2(i, k)P(k, j).$$

The kth term in this sum gives the probability that the process transitions from state i in period 0 to state k in period 2 through any intermediate state, and then transitions from state k to state j. Summing over all k, we obtain the probability that the process moves from state i in period 0 to state j in period 3 through any intermediate states in periods 1 and 2. Because this argument could be extended to cover every subsequent period, we have established that $P^t(i, j)$ is the probability that the process moves from state i in period 0 to state j in period t through any intermediate states in periods 1 through $t - 1$.

Returning to the matrix computations, it is interesting (perhaps even surprising) to find that the rows of the P^t matrix become more similar as t becomes large. We'll return to this "convergence" result below. But first, let's consider another example which introduces a first sociological application of Markov chains.

1.4 Social mobility

Sociologists have long been interested in *social mobility* – the transition of individuals between social classes defined on the basis of income or occupation. Some research has focused on *intergenerational* mobility from parent's class to child's class, while other research has examined *intragenerational* mobility over an individual's life course. Chapter 2 will explore this several aspects of this topic in greater depth. But here, we'll develop a simple hypothetical example.

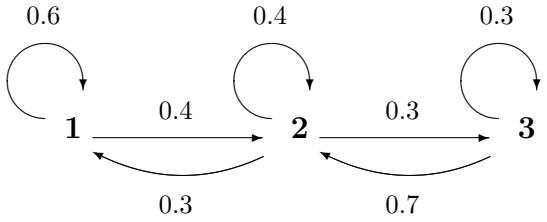
Consider a society with three social classes. Each individual may belong to the lower class (state 1), the middle class (state 2), or the upper class (state 3). Thus, the social class occupied by an individual in generation t may be denoted by $s_t \in \{1, 2, 3\}$. Further suppose that each individual in generation t has exactly one child

in generation $t+1$, who has exactly one child in generation $t+2$, and so on.[1] Finally, suppose that intergenerational mobility is characterized by a (3×3) transition matrix which does not change over time. Under these conditions, a single "family history" – the sequence of social classes $(s_0, s_1, s_2, \ldots)$ – is a Markov chain.

To offer a numerical example, suppose that intergenerational mobility is described by the transition matrix

$$P = \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.7 & 0.3 \end{bmatrix}$$

which may, in turn, be represented by the transition diagram below.



Thus, a child with a lower-class parent has a 60% chance of remaining in the lower class, has a 40% chance to rise to the middle class, and has no chance to reach the upper class. A child with a middle-class parent has a 30% chance of falling to the lower class, a 40% chance of remaining middle class, and a 40% chance of rising to the upper class. Finally, a child with an upper-class parent have no chance of falling to the lower class, has a 70% chance of falling to the middle class, and has a 30% chance of remaining in the upper class.

While mobility across one generation can thus be read directly from the transition matrix, how can we compute the "life chances" of subsequent generations? That is, how does the parent's class affect the class attained by the grandchild, greatgrandchild, etc? Having assumed that social mobility is a Markov chain process, we can again obtain the answers through matrix algebra.

```
>> P = [.6 .4 0; .3 .4 .3; 0 .7 .3]

P =
    0.6000    0.4000         0
    0.3000    0.4000    0.3000
```

[1] Two assumptions are implicit here. First, we are assuming that fertility rates do not vary across social classes. Second, we are ignoring complications that arise because individuals actually have two parents (who might themselves have come from different social classes). To use demographers' terminology, we are considering a "one-sex" (rather than "two-sex") model. While the present model offers a useful starting point, we will begin to address differential reproduction and two-sex models in Chapters 3 and xx.

```
       0      0.7000    0.3000

>> P^2

ans =
    0.4800    0.4000    0.1200
    0.3000    0.4900    0.2100
    0.2100    0.4900    0.3000

>> P^3

ans =
    0.4080    0.4360    0.1560
    0.3270    0.4630    0.2100
    0.2730    0.4900    0.2370

>> P^4

ans =
    0.3756    0.4468    0.1776
    0.3351    0.4630    0.2019
    0.3108    0.4711    0.2181

>> P^5

ans =
    0.3594    0.4533    0.1873
    0.3400    0.4606    0.1995
    0.3278    0.4654    0.2068

>> P^10

ans =
    0.3447    0.4589    0.1965
    0.3441    0.4591    0.1968
    0.3438    0.4592    0.1970

>> P^100

ans =

    0.3443    0.4590    0.1967
    0.3443    0.4590    0.1967
    0.3443    0.4590    0.1967
```

Given the assumed transition matrix, it is impossible for the children of lower-class parents to transition directly to the upper class. But these computations show that grandchildren of these parents have a 12% of reaching the upper class, that great-grandchildren have greater than 15% chance, and that in the long-run (after many generations) there's a nearly 20% chance.

1.5 The limiting distribution

In both of the numerical examples we've considered, the rows of the P^t matrix become increasingly similar as t rises, and become identical as t becomes very large. This "convergence" result may be described in several ways. We may say that, after a Markov chain process has run for many periods, the current state of the chain does not depend on the initial state of the chain. Alternatively, we may say that the *limiting distribution* – the probability distribution over states in the long run – is independent of the initial state.

To define the limiting distribution more precisely, suppose that the initial state of the chain is characterized by the row vector $\mathbf{x}_0$. For instance, given a Markov process with 3 states, and assuming that the process begins in state 2, this vector is

$$\mathbf{x}_0 \;=\; \begin{bmatrix} 0 & 1 & 0 \end{bmatrix},$$

which indicates that the process is initially in state 2 with probability 1 (and the other states with probability 0). Having specified the initial condition in this fashion, the probability distribution over states in period 1 is given by

$$\mathbf{x}_1 \;=\; \mathbf{x}_0 P,$$

the probability distribution in period 2 is given by

$$\mathbf{x}_2 \;=\; \mathbf{x}_1 P \;=\; (\mathbf{x}_0 P)P \;=\; \mathbf{x}_0 P^2,$$

and the probability distribution in period 3 is given by

$$\mathbf{x}_3 \;=\; \mathbf{x}_2 P \;=\; (\mathbf{x}_1 P)P \;=\; ((\mathbf{x}_0 P)P)P \;=\; \mathbf{x}_0 P^3.$$

By induction, the probability distribution for period t is given by

$$\mathbf{x}_t \;=\; \mathbf{x}_0 P^t.$$

By definition, the limiting distribution is the probability distribution $\mathbf{x}_t$ as t becomes very large (i.e., as t approaches the "limit" of infinity).

In the examples we have seen, the limiting distribution does not depend on the initial condition $\mathbf{x}_0$. Of course, because we have considered only two examples, we should not assume that this result will hold for every Markov chain process. Indeed, it turns out that this result is guaranteed only if the transition matrix satisifies the condition described in the following

Definition.[2] *A square, non-negative matrix A is* ***primitive*** *if and only if there exists some $t \geq 1$ such that every element of A^t is positive (i.e., $A^t(i,j) > 0$ for all i, j).*

[2]Some authors – notably Kemeny and Snell (1960) – refer to a primitive matrix as a *regular* matrix. In any case, this condition should not be mistaken for *irreducibility*, a weaker condition that will be discussed in Chapter 7.

Given our matrix computations, it is easy to see that both of the transition matrices we've considered are primitive. In the first example, all elements of the P^t matrix are positive for every $t \geq 3$. In the second (social mobility) example, all elements of the P^t matrix are positive for every $t \geq 2$. Having offered this definition, the "convergence" result can now be stated precisely as the following

Theorem.[3] *Consider a Markov chain process for which the transition matrix P is primitive, and the initial state of the chain is given by $\mathbf{x}_0$ so that the probability distribution in period t is given by the vector $\mathbf{x}_t = \mathbf{x}_0 P^t$. As t becomes very large, $\mathbf{x}_t$ converges to the unique probability vector $\mathbf{x}$ such that $\mathbf{x} = \mathbf{x}P$.*

Thus, if the transition matrix is primitive, the probability distribution $\mathbf{x}_t$ converges to the limiting distribution $\mathbf{x}$ as t becomes large. Moreover, the limiting distribution $\mathbf{x}$ does not depend on the initial condition $\mathbf{x}_0$.

1.6 Solving algebraically for the limiting distribution

We have already seen how to solve numerically for the limiting distribution. Raising the (primitive) transition matrix P to some sufficiently high power t, every row of P^t is equal to $\mathbf{x}$. But it is useful to recognize that the condition

$$\mathbf{x} = \mathbf{x}P$$

is a simultaneous equation system that can also be solved algebraically. To illustrate, consider again the social mobility example. The limiting distribution is determined by the condition

$$\begin{bmatrix} \mathbf{x}(1) & \mathbf{x}(2) & \mathbf{x}(3) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(1) & \mathbf{x}(2) & \mathbf{x}(3) \end{bmatrix} \begin{bmatrix} 0.6 & 0.4 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.7 & 0.3 \end{bmatrix}$$

along with the requirement that $\mathbf{x}$ is a probability vector. Reverting to high-school algebra, the matrix condition can be rewritten as

$$\begin{aligned} \mathbf{x}(1) &= 0.6\ \mathbf{x}(1) + 0.3\ \mathbf{x}(2) \\ \mathbf{x}(2) &= 0.4\ \mathbf{x}(1) + 0.4\ \mathbf{x}(2) + 0.7\ \mathbf{x}(3) \\ \mathbf{x}(3) &= 0.3\ \mathbf{x}(2) + 0.3\ \mathbf{x}(3) \end{aligned}$$

and the probability vector requirement is

$$\mathbf{x}(1) + \mathbf{x}(2) + \mathbf{x}(3) = 1.$$

[3] See Kemeny and Snell (1960, Chapter 4) for a restatement and proof of this theorem. This theorem can also be viewed as a special case of the Perron-Frobenius Theorem, which is the central result necessary for understanding the long-run dynamics of linear systems. See Chapter 3 for more discussion.

Using the first and third equations, we obtain

$$\begin{aligned} \mathbf{x}(1) &= (3/4)\ \mathbf{x}(2) \\ \mathbf{x}(3) &= (3/7)\ \mathbf{x}(2) \end{aligned}$$

and substitution into the probability vector condition yields

$$(3/4)\ \mathbf{x}(2) + \mathbf{x}(2) + (3/7)\ \mathbf{x}(2) \ = \ 1.$$

We thus obtain

$$\begin{aligned} \mathbf{x}(1) &= 21/61 \approx 0.3443 \\ \mathbf{x}(2) &= 28/61 \approx 0.4590 \\ \mathbf{x}(3) &= 12/61 \approx 0.1967 \end{aligned}$$

which is consistent with our numerical computations.

This algebraic approach becomes essential when the elements of the transition matrix are specified symbolically (rather than numerically). For instance, consider the generic 2-state Markov chain process, with transition matrix

$$P \ = \ \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}.$$

Without specifying the parameters a and b we cannot solve numerically for the limiting distribution. But proceeding algebraically, it is easy to verify that the limiting distribution is given by

$$\mathbf{x} \ = \ \begin{bmatrix} b/(a+b) & a/(a+b) \end{bmatrix}.$$

Intuitively, an increase in the parameter a (holding b constant) implies that the system transitions more often from state 1 to state 2, and thus increases the probability of occupying state 2 (and decreases the probability of occupying state 1) in the long run. Conversely, an increase in b (holding a constant) increases the probability that the process occupies state 1 in the long run.

1.7 A "macro-level" interpretation

Throughout this chapter, we have so far maintained a "micro-level" interpretation of Markov chain processes. For instance, in the social mobility example, we adopted the perspective of a particular individual, and considered how that individual's social class affected the "life chances" of his child and grandchild and subsequent descendents. Thus, the initial condition $\mathbf{x}_0$ reflected the individual's class, and the probability vector $\mathbf{x}_t$ reflected the probability distribution over classes for a particular descendent in generation t.

But it is also possible to adopt a "macro-level" interpretation of this process. Given a large population in which each individual belongs to one of the social classes,

we may interpret $\mathbf{x}_0(i)$ as the share of the population in class i in generation 0. For instance, setting

$$\mathbf{x}_0 \;=\; \begin{bmatrix} 0.2 & 0.3 & 0.5 \end{bmatrix},$$

we are assuming that 20% of the population belongs to the lower class, that 30% belong to the middle class, and that 50% belong to the upper class. Assuming that each individual has one child, population dynamics are determined by the equation

$$\mathbf{x}_t \;=\; \mathbf{x}_0 P^t$$

and we know (from the preceding theorem) that the population distribution $\mathbf{x}_t$ will converge to the limiting distribution $\mathbf{x}$ as t becomes large. To illustrate, consider population dynamics over the next 15 generations (computed using a `for` loop in Matlab).

```
>> x = [.2 .3 .5]; P = [.6 .4 0; .3 .4 .3; 0 .7 .3];

>> for t = 0:15; disp(x*P^t); end

    0.2000    0.3000    0.5000

    0.2100    0.5500    0.2400

    0.2910    0.4720    0.2370

    0.3162    0.4711    0.2127

    0.3310    0.4638    0.2051

    0.3378    0.4615    0.2007

    0.3411    0.4602    0.1987

    0.3427    0.4596    0.1977

    0.3435    0.4593    0.1972

    0.3439    0.4592    0.1969

    0.3441    0.4591    0.1968

    0.3442    0.4590    0.1968

    0.3442    0.4590    0.1967

    0.3442    0.4590    0.1967

    0.3443    0.4590    0.1967

    0.3443    0.4590    0.1967
```

Thus, for this example, the population distribution has converged to the limiting distribution in less than 15 generations.

Given the micro-level interpretation, Markov chain processes are *stochastic*. Our knowledge of an individual's current class i in period 0 allows us to derive the probability that a descendent is in class i in period t. But obviously we cannot know the realization of the process – the particular social class that will occupied – in period t. In contrast, adopting the macro-level interpretation, Markov chain processes generate population dynamics that are *deterministic*. Intuitively, each individual in the population is associated with a different Markov chain, and we can ignore sampling variation because we are averaging across the realizations of many chains.[4]

1.8 Further reading

Bradley and Meek (1986, Chap 6) provide an informal introduction to Markov chains. The classic undergraduate text by Kemeny, Snell, and Thompson (1966) offers a somewhat more formal (but still very readable) introduction to Markov chains (see especially Chapters 4.13 and 5.7) as well as probability theory, matrix algebra, and other useful topics in discrete mathematics. See Kemeny and Snell (1960) for a more rigorous treatment of Markov chains.

[4]More formally, by assuming that the size of the population approaches infinity, we can invoke the Law of Large Numbers (see, e.g., Kemeny and Snell, 1960, p 73).