Genetics

Markov chains

Markov chains

We conduct an experiment with a set of r outcomes,

$$S = \{S_1, \dots, S_r\}.$$

The experiment is repeated n times (with n large, potentially infinite).

The system has no memory: the next state depends only on the present state.

The probability of S_j occurring on the next step, given that S_i occurred on the last step, is

$$p_{ij} = p(S_j | S_i).$$

Suppose that S_i is the current state, then one of $S_1, \dots, S_r$ must be the next state; therefore,

$$p_{i1} + p_{i2} + \cdots + p_{ir} = 1, \quad 1 \le i \le r.$$

(Note that some of the p_{ij} can be zero, all that is needed is that $\sum_{j=1}^{r} p_{ij} = 1$ for all i.)

Markov chain

Definition

An experiment with finite number of possible outcomes $S_1, \ldots, S_r$ is repeated. The sequence of outcomes is a *Markov chain* if there is a set of r^2 numbers $\{p_{ij}\}$ such that the conditional probability of outcome S_j on any experiment given outcome S_i on the previous experiment is p_{ij}, i.e., for $1 \leq i, j \leq r$, $n = 1, \ldots,$

$$p_{ij} = \Pr(S_j \text{ on experiment } n+1 | S_i \text{ on experiment } n).$$

The outcomes $S_1, \ldots, S_r$ are the *states*, and the p_{ij} are the *transition probabilities*. The matrix $P = [p_{ij}]$ is the *transition matrix*.

Transition matrix

The matrix

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{pmatrix}$$

has

- nonnegative entries, $p_{ij} \geq 0$
- entries less than 1, $p_{ij} \leq 1$
- row sum 1, which we write

$$\sum_{j=1}^{r} p_{ij} = 1, \quad i = 1, \ldots, r$$

or, using the notation $\mathbb{1}^T = (1, \ldots, 1)$,

$$P\mathbb{1} = \mathbb{1}$$

Simple Mendelian inheritance

A certain trait is determined by a specific pair of genes, each of which may be two types, say G and g.

One individual may have:

- GG combination (*dominant*)
- Gg or gG, considered equivalent genetically (*hybrid*)
- gg combination (*recessive*)

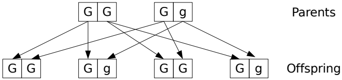
In sexual reproduction, offspring inherit one gene of the pair from each parent.

Basic assumption of Mendelian genetics

Genes inherited from each parent are selected at random, independently of each other. This determines probability of occurrence of each type of offspring. The offspring

- of two *GG* parents must be *GG*,
- of two *gg* parents must be *gg*,
- of one *GG* and one *gg* parent must be *Gg*,
- other cases must be examined in more detail.

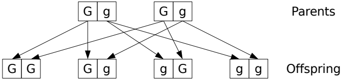
GG and *Gg* parents



Offspring has probability

- $\frac{1}{2}$ of being *GG*
- $\frac{1}{2}$ of being *Gg*

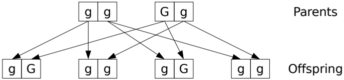
Gg and *Gg* parents



Offspring has probability

- $\frac{1}{4}$ of being *GG*
- $\frac{1}{2}$ of being *Gg*
- $\frac{1}{4}$ of being *gg*

gg and *Gg* parents



Offspring has probability

- $\frac{1}{2}$ of being *Gg*
- $\frac{1}{2}$ of being *gg*

General case

Let $p_i(n)$ be the probability that the state S_i will occur on the n^{th} repetition of the experiment, $1 \leq i \leq r$.

Since one the states S_i must occur on the n^{th} repetition,

$$p_1(n) + p_2(n) + \cdots + p_r(n) = 1.$$

Let $p_i(n+1)$ be the probability that state S_i, $1 \leq i \leq r$, occurs on $(n+1)^{th}$ repetition of the experiment.

There are r ways to be in state S_i at step $n+1$:

1. Step n is S_1. Probability of getting S_1 on n^{th} step is $p_1(n)$, and probability of having S_i after S_1 is p_{1i}. Therefore, by multiplication principle, $P(S_i|S_1) = p_{1i}p_1(n)$.
2. We get S_2 on step n and S_i on step $(n+1)$. Then $P(S_i|S_2) = p_{2i}p_2(n)$.

..

r. Probability of occurrence of S_i at step $n+1$ if S_r at step n is $P(S_i|S_r) = p_{ri}p_r(n)$.

Therefore, $p_i(n+1)$ is sum of all these,

$$\begin{aligned} p_i(n+1) &= P(S_i|S_1) + \cdots + P(S_i|S_r) \\ &= p_{1i}p_1(n) + \cdots + p_{ri}p_r(n) \end{aligned}$$

Therefore,

$$\begin{aligned} p_1(n+1) &= p_{11}p_1(n) + p_{21}p_2(n) + \cdots + p_{r1}p_r(n) \\ &\vdots \\ p_k(n+1) &= p_{1k}p_1(n) + p_{2k}p_2(n) + \cdots + p_{rk}p_r(n) \qquad (1) \\ &\vdots \\ p_r(n+1) &= p_{1r}p_1(n) + p_{2r}p_2(n) + \cdots + p_{rr}p_r(n) \end{aligned}$$

In matrix form

$$p(n+1) = p(n)P, \quad n = 1, 2, 3, \ldots \tag{2}$$

where $p(n) = (p_1(n), p_2(n), \ldots, p_r(n))$ is a (row) probability vector and $P = (p_{ij})$ is a $r \times r$ *transition matrix*,

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ & & & \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{pmatrix}$$

So, what we have is

$$(p_1(n+1), \ldots, p_r(n+1)) =$$

$$(p_1(n), \ldots, p_r(n)) \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ & & & \\ p_{r1} & p_{r2} & \cdots & p_{rr} \end{pmatrix}$$

It is easy to check that this gives the same expression as (1).

For our genetic model..

Consider a process of continued matings.

- Start with an individual of known or unknown genetic character and mate it with a hybrid.
- Assume that there is at least one offspring; choose one of them at random and mate it with a hybrid.
- Repeat this process through a number of generations.

The genetic type of the chosen offspring in successive generations can be represented by a Markov chain, with states GG, Gg and gg. So there are 3 possibles states $S_1 = GG$, $S_2 = Gg$ and $S_3 = gg$.

We have

$\nearrow$	GG	Gg	gg
GG	0.5	0.5	0
Gg	0.25	0.5	0.25
gg	0	0.5	0.5

The transition probabilities are thus

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Stochastic matrices

Definition (Stochastic matrix)

The nonnegative $r \times r$ matrix M is *stochastic* if $\sum_{j=1}^{r} a_{ij} = 1$ for all $i = 1, 2, \ldots, r$.

Theorem

Let M be a stochastic matrix M. Then all eigenvalues λ of M are such that $|\lambda| \leq 1$. Furthermore, $\lambda = 1$ is an eigenvalue of M.

To see that 1 is an eigenvalue, write the definition of a stochastic matrix, i.e., M has row sums 1. In vector form, $M\mathbb{1} = \mathbb{1}$. Now remember that λ is an eigenvalue of M, with associated eigenvector v, iff $Mv = \lambda v$. So, in the expression $M\mathbb{1} = \mathbb{1}$, we read an eigenvector, $\mathbb{1}$, and an eigenvalue, 1.

Long "time" behavior

Let $p(0)$ be the initial distribution (row) vector. Then

$$\begin{aligned} p(1) &= p(0)P \\ p(2) &= p(1)P \\ &= (p(0)P)P \\ &= p(0)P^2 \end{aligned}$$

Iterating, we get that for any n,

$$p(n) = p(0)P^n$$

Therefore,

$$\lim_{n\to+\infty} p(n) = \lim_{n\to+\infty} p(0)P^n = p(0) \lim_{n\to+\infty} P^n$$

Additional properties of stochastic matrices

Theorem

If M, N are stochastic matrices, then MN is a stochastic matrix.

Theorem

If M is a stochastic matrix, then for any $k \in \mathbb{N}$, M^k is a stochastic matrix.

Regular Markov chain

Definition (Regular Markov chain)

A regular Markov chain is one in which P^k is positive for some integer $k > 0$, i.e., P^k has only positive entries, no zero entries.

Definition

A nonnegative matrix M is primitive if, and only if, there is an integer $k > 0$ such that M^k is positive.

Theorem

A Markov chain is regular if, and only if, the transition matrix P is primitive.

Important result for regular Markov chains

Theorem

If P is the transition matrix of a regular Markov chain, then

1. *the powers P^n approach a stochastic matrix W,*
2. *each row of W is the same (row) vector $w = (w_1, \ldots, w_r)$,*
3. *the components of w are positive.*

So if the Markov chain is regular,

$$\lim_{n\to+\infty} p(n) = p(0) \lim_{n\to+\infty} P^n = p(0)W$$

Left and right eigenvectors

Let M be an $r \times r$ matrix, u, v be two column vectors, $\lambda \in \mathbb{R}$. Then, if

$$Mu = \lambda u,$$

u is the (right) eigenvector corresponding to λ, and if

$$v^T M = \lambda v^T$$

then v is the left eigenvector corresponding to λ. Note that to a given eigenvalue there corresponds one left and one right eigenvector.

The vector w is in fact the left eigenvector corresponding to the eigenvalue 1 of P. (We already know that the (right) eigenvector corresponding to 1 is $\mathbb{1}$.)

To see this, remark that, if $p(n)$ converges, then $p(n+1) = p(n)P$, so w is a fixed point of the system. We thus write

$$wP = w$$

and solve for w, which amounts to finding w as the left eigenvector corresponding to the eigenvalue 1.

Alternatively, we can find w as the (right) eigenvector associated to the eigenvalue 1 for the transpose of P,

$$P^T w^T = w^T$$

Now remember that when you compute an eigenvector, you get a result that is the eigenvector, to a multiple.

So the expression you obtain for w might have to be normalized (you want a probability vector). Once you obtain w, check that the norm $\|w\|$ defined by

$$\|w\| = w_1 + \cdots + w_r$$

is equal to one. If not, use

$$\frac{w}{\|w\|}$$

Back to genetics..

The Markov chain is here regular. Indeed, take the matrix P,

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

and compute P^2:

$$P^2 = \begin{pmatrix} \frac{3}{8} & \frac{1}{2} & \frac{1}{8} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \end{pmatrix}$$

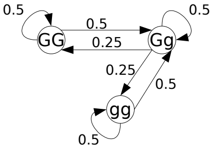
As all entries are positive, P is primitive and the Markov chain is regular.

Another way to check regularity:

Theorem

A matrix M is primitive if the associated connection graph is strongly connected, i.e., that there is a path between any pair (i, j) of states, and that there is at least one positive entry on the diagonal of M.

This is checked directly on the transition graph



Compute the left eigenvector associated to 1 by solving

$$(w_1, w_2, w_3) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} = (w_1, w_2, w_3)$$

$$\frac{1}{2}w_1 + \frac{1}{4}w_2 = w_1 \tag{3a}$$
$$\frac{1}{2}w_1 + \frac{1}{2}w_2 + \frac{1}{2}w_3 = w_2 \tag{3b}$$
$$\frac{1}{4}w_2 + \frac{1}{2}w_3 = w_3 \tag{3c}$$

From (3a), $w_1 = w_2/2$, and from (3c), $w_3 = w_2/2$. Substituting these values into (3b),

$$\frac{1}{4}w_2 + \frac{1}{2}w_2 + \frac{1}{4}w_2 = w_2,$$

that is, $w_2 = w_2$, i.e., w_2 can take any value. So $w = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.

Markov chains

A simple genetic model

Repetition of the process

Regular Markov chains

Absorbing Markov chains

Mating with a *GG* individual

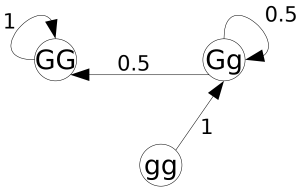
Suppose now that we conduct the same experiment, but mate each new generation with a *GG* individual instead of a *Gg* individual. Transition table is

$\nearrow$	GG	Gg	gg
GG	1	0	0
Gg	0.5	0.5	0
gg	0	1	0

The transition probabilities are thus

$$P = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

New transition graph



Clearly:

- we leave *gg* after one iteration, and can never return,
- as soon as we leave *Gg*, we can never return,
- can never leave *GG* as soon as we get there.

Absorbing states, absorbing chains

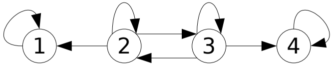
Definition
A state S_i in a Markov chain is *absorbing* if whenever it occurs on the n^{th} generation of the experiment, it then occurs on every subsequent step. In other words, S_i is absorbing if $p_{ii} = 1$ and $p_{ij} = 0$ for $i \neq j$.

Definition
A Markov chain is said to be absorbing if it has at least one absorbing state, and if from every state it is possible to go to an absorbing state.

In an absorbing Markov chain, a state that is not absorbing is called *transient*.

Some questions on absorbing chains

Suppose we have a chain like the following:



1. Does the process eventually reach an absorbing state?
2. Average number of times spent in a transient state, if starting in a transient state?
3. Average number of steps before entering an absorbing state?
4. Probability of being absorbed by a given absorbing state, when there are more than one, when starting in a given transient state?

Reaching an absorbing state

Answer to question 1:

Theorem

In an absorbing Markov chain, the probability of reaching an absorbing state is 1.

Standard form of the transition matrix

For an absorbing chain with k absorbing states and $r-k$ transient states, the transition matrix can be written as

$$P = \begin{pmatrix} \mathbb{I}_k & \mathbf{0} \\ R & Q \end{pmatrix}$$

with following meaning,

	Absorbing states	Transient states
Absorbing states	$\mathbb{I}_k$	$\mathbf{0}$
Transient states	R	Q

with $\mathbb{I}_k$ the $k \times k$ identity matrix, $\mathbf{0}$ an $k \times (r-k)$ matrix of zeros, R an $(r-k) \times k$ matrix and Q an $(r-k) \times (r-k)$ matrix.

The matrix $\mathbb{I}_{r-k} - Q$ is invertible. Let

- $N = (\mathbb{I}_{r-k} - Q)^{-1}$ be the *fundamental matrix* of the Markov chain
- T_i be the sum of the entries on row i of N
- $B = NR$.

Answers to our remaining questions:

2. N_{ij} is the average number of times the process is in the jth transient state if it starts in the ith transient state.
3. T_i is the average number of steps before the process enters an absorbing state if it starts in the ith transient state.
4. B_{ij} is the probability of eventually entering the jth absorbing state if the process starts in the ith transient state.

Back to genetics..

The matrix is already in standard form,

$$P = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \mathbb{I}_1 & \mathbf{0} \\ R & Q \end{pmatrix}$$

with $\mathbb{I}_1 = 1$, $\mathbf{0} = (0 \;\; 0)$ and

$$R = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} \qquad Q = \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix}$$

We have

$$\mathbb{I}_2 - Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} \frac{1}{2} & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ -1 & 1 \end{pmatrix}$$

so

$$N = (\mathbb{I}_2 - Q)^{-1} = 2 \begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 2 & 1 \end{pmatrix}$$

Then

$$T = N\mathbb{1} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

and

$$B = NR = \begin{pmatrix} 2 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$