1
Observed Markov Chains

1.1 Introduction

This book studies finite state processes in discrete time. The simplest such process is just a sequence of independent random variables which at each time takes any one of the possible values in its state space with equal probability. The canonical probability space for such a process is the space of sequences of outcomes. The following chapter starts by describing this probability space and constructing on it an appropriate measure. A modified construction then gives a probability space on which not all outcomes have equal probability. A Radon–Nikodym derivative is then defined so the sequence of outcomes is no longer independent, rather the probability of the following state depends on the present state. That is, we construct a Markov chain. This construction of a Markov chain from first principles is not given in other treatments of the subject. The semi-martingale representation of the chain is also given.

The following first few chapters contruct Markov chains and hidden Markov chains from first principles. Estimation algorithms are derived. Semi-Markov chains and hidden semi-Markov chains are introduced and discussed from Chapter 9 onwards.

1.2 Observed Markov chain models

Suppose $\{X_k\,; k = 1, 2, \dots\}$ is a sequence of random quantities taking values in some set $\mathcal{S}$.

We say $\{X_k\,; k = 1, 2, \dots, L\}$ is a *Markov chain* if the following prop-

erties hold:

$$P(X_k = x_k | X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1}) = P(X_k = x_k | X_{k-1} = x_{k-1})$$

for each $k \geq 1$ and for all $x_1, x_2, \ldots, x_k$.

This is also called an M1 model. The iid model is called an M0 model. In an Mq model, a Markov chain of order q:

$$P(X_k = x_k | X_1 = x_1, X_2 = x_2, \ldots, X_{k-1} = x_{k-1}) = P(X_k = x_k | X_{k-q} = x_{k-q}, \ldots, X_{k-1} = x_{k-1})$$

for each $n \geq 1$ and for all $x_1, x_2, \ldots, x_k$.

It can be shown that a Markov chain of higher order can be reduced to a 1-step Markov chain, so they are of limited independent interest, at least theoretically.

In fact for an $M2$ chain $\{X_n\}$, where we have

$$P(X_k = x_k | X_1 = x_1, \ldots, X_{k-1} = x_{k-1}) = P(X_k = x_k | X_{k-2} = x_{k-2}, X_{k-1} = x_{k-1}),$$

we obtain an M1 chain if we set

$$Y_k = \begin{pmatrix} X_k \\ X_{k-1} \end{pmatrix}$$

for $k = 2, 3, \ldots, L$.

In our models, the state space S, (the set of values that each term of the chain can take), is finite corresponding to the number of elements in an alphabet. If S has N elements, it is convenient to let S consist of the N standard unit vectors e_i, $i = 1, 2, \ldots, N$, in $\mathbb{R}^N$. Here $e_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^N$. Then the elements of $S = \{e_1, \ldots, e_N\}$ are in one-to-one correspondence with an alphabet $\mathcal{Q}$ having N elements.

From now on we assume $\mathcal{S} = \{e_1, e_2, \ldots, e_N\}$. Associated with a time-homogeneous Markov chain we have well-defined transition probabilities:

$$p_{ij} \equiv p_{e_i, e_j} = P(X_k = e_j | X_{k-1} = e_i).$$

These have the same values for each $k = 2, 3, \ldots, L$. (That is what homogeneous means here.) This is the convention used by many probabilists. However, shall use the convention used by those who work with HMMs and write

$$a_{ji} = P(X_k = e_j | X_{k-1} = e_i).$$

For the matrix (p_{ij}) the row sums are all 1, while for the matrix (a_{ji}) the column sums are all 1.

Along with the transition probabilities, we also need initial probabilities

$$\pi_j \equiv \pi_{e_j} = P(X_1 = e_j)$$

for each $e_j \in S$.

We can then write down the probability of any sample path

$$\mathbf{x} = (x_1, x_2, \ldots, x_L)$$

of

$$\mathbf{X} = (X_1, X_2, \ldots, X_L)$$

as

$$P(\mathbf{x}) \equiv P(\mathbf{X} = \mathbf{x}) = \pi_{x_1} \times \prod_{k=1}^{L} p_{x_{k-1}, x_k} .$$

To calibrate this kind of model, we need some 'training data'. This could mean that we have M sequences of length L as sample paths from our model. We could then make estimates

$$\hat{\pi}_j = \frac{\text{number of times } e_j \text{ occurs as } X_1}{M}$$

and

$$\widehat{a}_{ji} = \widehat{p}_{ij} = \frac{\text{number of times } e_j \text{ follows } e_i}{\text{number of times anything follow } e_i} .$$

We shall show that expressions like these are maximum likelihood estimators of these quantities.

In many other applications, there is often only one observed sample path of a Markov chain. In other words, when k of X_k represents time, we only have one observation history. It is not the case that different spectators in this world see different histories (even though they may report events as if that were the case). One observed sequence is the case with financial and economic data, or other tracking signals. In genomics we can often have more than one sample path from the same model. These could be obtained from a DNA molecule by selecting out several subsequences of length L.

In this section, we shall discuss the construction and estimation of observed Markov chain models.

We shall consider a Markov chain X taking values in a finite set $\mathcal{S}$. We have not specified $\mathcal{S}$, except that it has N elements, say. It does not really matter what the objects in $\mathcal{S}$ are as long as we know how to put them in one-to-one correspondence with some alphabet.

As above it is most convenient to identity the elements of $\mathcal{S}$ with the standard unit vectors $e_1, e_2, \ldots, e_N$ in $\mathbb{R}^N$. This means that

$$e_i = (0, 0, \ldots, 1, 0, \ldots, 0)^\top$$

where the 1 is in the ith place and $\top$ denotes the transpose.

We shall first construct a Markov chain $\{X_k; k = 0, 1, 2, \ldots\}$ taking values in a finite state space $\mathcal{S} = \{e_1, e_2, \ldots, e_N\}$. This Markov chain will be defined on a canonical probability space $(\Omega, \mathcal{F}, P)$, which we shall now describe.

Note that with this notation we have the representations:

$$1 = \sum_{i=1}^{N} \langle X_k, e_i \rangle$$

and for any real-valued functions $g(X_k)$

$$g(X_k) = \langle g, X_k \rangle$$

where $g = (g_1, g_2, \ldots, g_N)$ and $g_i = g(e_i)$.

For $u, v \in \mathbb{R}^N$ write $\langle u, v \rangle = u_1 v_1 + \cdots + u_N v_N$, the usual inner product of $\mathbb{R}^N$. We have just used this notation and will continue to use this notation.

1.3 Notation

We introduce some notation to be used in this section.

The sample space Ω will consist of all sequences of

$$\omega = (\omega_0, \omega_1, \omega_2, \ldots)$$

where $\omega_i \in \mathcal{S}$ for each $i \geq 0$.

A σ-algebra on Ω is a family of subsets $\mathcal{F}$ of Ω which satisfies:

(1) $\Omega \in \mathcal{F}$;
(2) if $A \in \mathcal{F}$ then the complement $A^c \in \mathcal{F}$;
(3) if $A_1, A_2, \ldots$ are all in $\mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Consider the family $\mathcal{F}^A$ of subsets of Ω of the form:

$$\{\omega \in \Omega \,|\, \omega_{i_k} = e_{i_k}, k = 1, 2, \ldots, l\}, \tag{1.1}$$

where $i_1 < i_2 < \cdots < i_l$ and $e_{i_1}, e_{i_2}, \ldots, e_{i_k}$ are elements of $\mathcal{S}$.

The σ-algebra $\mathcal{F}$ we shall consider on Ω will be the smallest σ-algebra generated by all the sets in $\mathcal{F}^A$.

Elements in $\mathcal{F}$ will be called *events*.

Once we have assigned a probability $P(B)$ to each event $B \in \mathcal{F}^A$, then it can be extended to an event of $F \in \mathcal{F}$, by first expressing F as a disjoint union of such sets:

$$F = \bigcup_{i=1}^{\infty} A_i\,, \quad A_i \in \mathcal{F}^A, \text{ with } A_i \cap A_j = \emptyset \text{ for } i \neq j$$

and then defining $P(F)$ by

$$P(F) = \sum_{i=1}^{\infty} P(A_i).$$

There are many ways that a probability can be assigned to an event.

The probability function has the defining properties:

(i) $P(\Omega) = 1$,
(ii) $P(A^c) = 1 - P(A)$, where $A^c = \Omega \setminus A$ for any $A \in \mathcal{F}$,
(iii) if $\{A_n\} \subset \mathcal{F}$ are disjoint, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)\,.$$

The canonical process The canonical process $\{X_k\}$ is defined on Ω by

$$X_k(\omega) = \omega_k \quad \text{for each } \omega \in \Omega$$

for $k = 0, 1, 2, \ldots$ The statistical properties of $\{X_k\}$ will depend on the probability P defined on $\mathcal{F}$.

We let $\mathcal{F}_n \subset \mathcal{F}$ be the collection all subsets of Ω generated by the events A with $i_k \leq n$ in (1.1) for $n = 0, 1, 2, \ldots$. Then we have

$$\Omega = \bigcup_{A \in \mathcal{F}_n} A\,.$$

Note that $\mathcal{F}_n$ is the σ-algebra generated by $X_0, X_1, \ldots, X_n$. This means that knowing the elements of $\mathcal{F}_n$ is equivalent to knowing $X_0, X_1, \ldots, X_n$. The increasing family of σ-algebras $\{\mathcal{F}_n\}$ is called a *filtration* on Ω.

We shall call $\{X_n\}$ a *Markov chain* if it has the following property:

$$P(X_{n+1} = e_j \,|\, \mathcal{F}_n) = P(X_{n+1} = e_j \,|\, X_n).$$

Here the left-hand side is a conditional probability depending on the

entire past history of the process $\{X_k \,|\, k = 0, 1, \ldots, n\}$ while on the right-hand side the conditional probability depends only on the knowledge of X_n.

This implies that we can define transition probabilities

$$a_{ji} = P(X_{n+1} = e_j \,|\, X_n = e_i)$$

where

$$\sum_{j=1}^{n} a_{ji} = 1\,.$$

This is the case for a (time-)homogeneous Markov chain, as the matrix of probabilities (a_{ji}) does not depend on n. However, the transition probabilities could depend on n so

$$a_{ji}(n) = P(X_{n+1} = e_j | X_n = e_i).$$

Many of the results below extend to the situation. We would then write $a_{ji}(n)$ in place of a_{ji} for all i, j.

As noted earlier some probabilists write

$$p_{ij} = P(X_{n+1} = e_j \,|\, X_n = e_i).$$

However, we shall not follow this practice as there are some distinct advantages using the above notation which is that used in Elliott et al. (1995) and in other papers.

It is also possible to define Markov chains of higher-order $M \geq 2$. The Markov chain we have just described is the usual one and has order 1. For an order-2 chain we would instead have the condition

$$P(X_{n+1} = e_j \,|\, \mathcal{F}_n) = P(X_{n+1} = e_j \,|\, X_n,\, X_{n-1}).$$

As these higher-order Markov chains are used in genomic modelling, we shall describe their representation as an order-1 Markov chain with an extended state space.

1.4 Construction of Markov chains

The reference model We say that we have the reference model when the probability is specified by

$$\overline{P}(B) = \frac{1}{N^l}$$

for events $B \in \mathcal{F}^A$, of the form (1.1).

We shall write $\overline{P}$ and $\overline{\mathbf{E}}$ to indicate probabilities and expectations using this probability.

Properties of the reference model

Property 1: We have

$$\begin{aligned}\overline{P}(X_k = e_j) &\equiv \overline{P}(\{\omega \in \Omega \,|\, X_k(\omega) = e_j\}) = \overline{P}(\{\omega \in \Omega | \, \omega_k = e_j\}) \\ &= \frac{1}{N}\end{aligned}$$

for each k and e_j.

This means that each X_k has the same distribution, and this is the uniform distribution, assigning equal probabilities to the occurrence of each state in $\mathcal{S}$.

Property 2: The terms of $\{X_k\}$ are independent.

To show this let $k < l$. Then

$$\begin{aligned}\overline{P}(X_k = e_j, X_l = e_i) &\equiv \overline{P}(\{\omega \in \Omega | \, X_k(\omega) = e_j, X_l(\omega) = e_i\}) \\ &= \overline{P}(\{\omega \in \Omega | \, \omega_k = e_j, \omega_l = e_i\}) \\ &= \frac{1}{N^2} \\ &= \overline{P}(X_k = e_j)\, \overline{P}(X_l = e_i)\end{aligned}$$

This means that X_k and X_l are independent for any k, l and so the sequence $\{X_n\}$ is a uniformly iid (**i**ndependent, **i**dentically **d**istributed) sequence.

An iid non-uniform model Let $q_1, q_2, \ldots, q_N \geq 0$ so that

$$\sum_{i=1}^{N} q_i = 1\,.$$

We now construct a probability P on $(\Omega, \mathcal{F})$ so that the $\{X_n\}$ are iid with

$$P(X_n = e_j) = q_j \quad \text{for } j = 1, \ldots, N.$$

Construction At each time n the Markov chain value X_n is just one of the unit vector elements e_i in its state space $\{e_1, e_2, \ldots, e_N\}$.

We shall often use the identity

$$\sum_{j=1}^{N} \langle X_n,\, e_j \rangle = 1$$

for any $n = 0, 1, 2, \ldots$ from time to time without further explanation. Inserting this identity into an argument from time to time is often a useful trick.

For $l = 0, 1, \ldots$, define

$$\overline{\lambda}_l = N\langle q, X_l\rangle$$

where $q = (q_1, \ldots, q_N)^\top$.

Lemma 1.1 *Recall $\overline{E}$ refers to the reference probability $\overline{P}$ defined above.*

(i) $\overline{\mathbf{E}}[\,\overline{\lambda}_0] = 1$,
(ii) $\overline{\mathbf{E}}[\,\overline{\lambda}_l | \,\mathcal{F}_{l-1}] = 1$ *for* $l \geq 1$.

Proof For (i), we have

$$\begin{aligned}\overline{\mathbf{E}}[\,\overline{\lambda}_0] = \overline{\mathbf{E}}[N\langle q, X_0\rangle] &= \overline{\mathbf{E}}\left[\sum_{i=1}^N \langle X_0, e_i\rangle\, N\langle q, X_0\rangle\right] \\ &= \overline{\mathbf{E}}\left[\sum_{i=1}^N \langle X_0, e_i\rangle\, N\,\langle q, e_i\rangle\right] = \sum_{i=1}^N N\, q_i \cdot \overline{\mathbf{E}}\,[\langle X_0, e_i\rangle] \\ &= \sum_{i=1}^N N q_i \cdot \frac{1}{N} = 1.\end{aligned}$$

For (ii), we have

$$\overline{\mathbf{E}}[\,\overline{\lambda}_l | \mathcal{F}_{l-1}] = \overline{\mathbf{E}}[N\langle q, X_l\rangle |\, \mathcal{F}_{l-1}] = \overline{\mathbf{E}}[N\langle q, X_l\rangle] = 1$$

where we used the fact that under $\overline{P}$ the $\{X_k\}$ are independent and so

$$\overline{\mathbf{E}}[N\langle q, X_l\rangle |\, \mathcal{F}_{l-1}] = \overline{\mathbf{E}}[N\langle q, X_l\rangle]$$

and the last equality follows as in $l = 0$. □

We now introduce a new probability on $(\Omega, \mathcal{F})$.

Write

$$\overline{\Lambda}_n = \prod_{l=0}^n \overline{\lambda}_n = \overline{\lambda}_0 \cdot \overline{\lambda}_1 \cdots \overline{\lambda}_n\,. \tag{1.2}$$

We define the new probability P by requiring that

$$\frac{dP}{d\overline{P}}\bigg|_{\mathcal{F}_n} = \overline{\Lambda}_n\,.$$

This simply means that if $A \in \mathcal{F}_n$, then

$$P(A) = \overline{\mathbf{E}}[\,\overline{\Lambda}_n\, I(A)]\,. \tag{1.3}$$

We note that if $A \in \mathcal{F}_n$, then $A \in \mathcal{F}_{n+1}$ also. This leads to two definitions of $P(A)$ depending on whether we use $\overline{\Lambda}_n$ or $\overline{\Lambda}_{n+1}$ in (1.3). However, we have the following result:

Lemma 1.2 *The definition of P is well defined. That is,*

$$\overline{\mathbf{E}}[\,\overline{\Lambda}_n\, I(A)] = \overline{\mathbf{E}}[\,\overline{\Lambda}_m\, I(A)]$$

for any $A \in \mathcal{F}_n$ and $m > n$.

Proof We first note that

$$\overline{\mathbf{E}}[\,\overline{\Lambda}_m \,|\, \mathcal{F}_n] = \overline{\Lambda}_n\,.$$

This follows from Lemma 1.1, because

$$\begin{aligned}
\overline{\mathbf{E}}[\,\overline{\Lambda}_m \,|\, \mathcal{F}_n] &= \overline{\mathbf{E}}[\,\overline{\mathbf{E}}[\,\overline{\Lambda}_m \,|\, \mathcal{F}_{m-1}]\,|\,\mathcal{F}_n] \\
&= \overline{\mathbf{E}}[\,\overline{\Lambda}_{m-1}\overline{\mathbf{E}}[\,\overline{\lambda}_m \,|\, \mathcal{F}_{m-1}]\,|\,\mathcal{F}_n] \\
&= \overline{\mathbf{E}}[\,\overline{\Lambda}_{m-1} \,|\, \mathcal{F}_n] \\
&= \overline{\mathbf{E}}[\,\overline{\Lambda}_{m-2} \,|\, \mathcal{F}_n] \\
&= \cdots \\
&= \overline{\mathbf{E}}[\,\overline{\Lambda}_n \,|\, \mathcal{F}_n] \\
&= \overline{\Lambda}_n\,.
\end{aligned}$$

Then for $A \in \mathcal{F}_n$

$$\begin{aligned}
\overline{\mathbf{E}}[\,\overline{\Lambda}_m\, I(A)] &= \overline{\mathbf{E}}[\,\overline{\mathbf{E}}[\,\overline{\Lambda}_m\, I(A) \,|\, \mathcal{F}_n]\,] \\
&= \overline{\mathbf{E}}[\,\overline{\mathbf{E}}[\,\overline{\Lambda}_m \,|\, \mathcal{F}_n]\, I(A)\,] \\
&= \overline{\mathbf{E}}[\,\overline{\Lambda}_n\, I(A)\,]
\end{aligned}$$

and we are done. □

Now let B be an event in $\mathcal{F}^A$ so $B \in \mathcal{F}_n$ for some $n \geq 0$. We then define

$$P(B) = \overline{\mathbf{E}}[\,\overline{\Lambda}_n\, I(B)\,],$$

and by Lemma 1.2, this is well defined. Suppose $F \in \mathcal{F}$ is of the form

$$F = \bigcup_{j=1}^{\infty} A_j$$

for disjoint events $\{A_j\}$, $A_j \in \mathcal{F}^A$, (we could let many of the $A_j = \emptyset$). We then set

$$P(F) = \sum_{j=1}^{\infty} P(A_j)\,.$$

Properties of the iid non-uniform model We now investigate the statistics of $\{X_k\}$ under P.

Property 1: We have $P(X_k = e_j) = q_j$ for each k, j.

Proof For $k \geq 0$ and $j \in \{1, 2, \ldots, N\}$, let $A = \{\omega \in \Omega |\, \omega_k = e_j\} \in \mathcal{F}_k$, then

$$\begin{aligned}
P(X_k = e_j) &= P\left(\{\omega \in \Omega |\, X_k(\omega) = e_j\}\right) \\
&= P(A) \\
&= \overline{\mathbf{E}}[\overline{\Lambda}_k\, I(A)] \\
&= \overline{\mathbf{E}}\left[\,\overline{\mathbf{E}}[\overline{\Lambda}_k\, I(A)|\,\mathcal{F}_{k-1}]\right] \\
&= \overline{\mathbf{E}}\left[\,\overline{\Lambda}_{k-1}\,\overline{\mathbf{E}}[\overline{\lambda}_k\, I(A)|\,\mathcal{F}_{k-1}]\right] \\
&= \overline{\mathbf{E}}\left[\,\overline{\Lambda}_{k-1}\,\overline{\mathbf{E}}[\overline{\lambda}_k\, I(A)]\right] \\
&= \overline{\mathbf{E}}[\overline{\lambda}_k\, I(A)]\,\overline{\mathbf{E}}[\,\overline{\Lambda}_{k-1}] \\
&= \overline{\mathbf{E}}[\overline{\lambda}_k\, I(A)],
\end{aligned}$$

where we note that $\overline{\lambda}_k\, I(A)$ depends only on the values of X_k and so is independent under $\overline{P}$ of $\mathcal{F}_{k-1}$.

We also used

$$\overline{\mathbf{E}}[\overline{\Lambda}_{k-1}] = \overline{\mathbf{E}}[\overline{\Lambda}_{k-1}\, I(\Omega)] = P(\Omega) = 1.$$

Continuing the calculation,

$$\begin{aligned}
\overline{\mathbf{E}}[\overline{\lambda}_k\, I(A)] &= \overline{\mathbf{E}}\left[\sum_{i=1}^{N} \langle X_k, e_i\rangle\, N\langle X_k, q\rangle\, I(X_k = e_j)\right] \\
&= \overline{\mathbf{E}}\left[\sum_{i=1}^{N} \langle X_k, e_i\rangle\, N\langle e_i, q\rangle\, I(e_i = e_j)\right] \\
&= \overline{\mathbf{E}}\left[\langle X_k, e_j\rangle\, N q_j\right] \\
&= \frac{1}{N} \cdot N q_j \\
&= q_j
\end{aligned}$$

where we used

$$\overline{\mathbf{E}}[\langle X_k, e_j\rangle] = \overline{P}(X_k = e_j) = \frac{1}{N}\,.$$