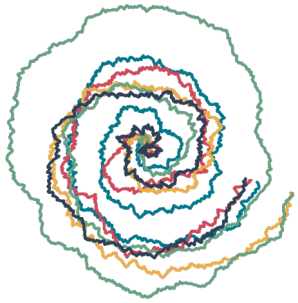


Stochastic Processes

STAT 4370/5370

Théo Michelot

September 18, 2023

Table of contents

Preface

These are the notes for the course STAT 4370/5370 Stochastic Processes at Dalhousie University. For other resources, such as the syllabus and lecture slides, please consult the Brightspace page for the course.

Note: The material in these notes is not original, and much of it has been adapted from other references on stochastic processes, to which I am greatly indebted. There are many general books on stochastic processes, as well as on specific topics covered in each chapter of this course, including the following references.

- Ross (2019) and Grimmett and Stirzaker (2020) are classic texts on probability and stochastic processes, and they provide a rigorous mathematical treatment of most topics covered in this course.
- Dobrow (2016) and Korosteleva (2022) provide a more application-oriented description, including many great examples, as well as R code for implementation.
- Norris (1998) is a great reference about Markov chains specifically.
- Zucchini, MacDonald, and Langrock (2017) give a relatively non-technical introduction to hidden Markov models, with a focus on application and implementation.

All errors in these notes are my own, however; please let me know when you find one! You can contact me at: `theo.michelot@dal.ca`.

1 Background

1.1 Some motivation

A stochastic process describes the evolution of some phenomenon over time. "Stochastic" is generally used as interchangeable with "random", indicating that the processes include an element of chance or unpredictability. Note that this does not imply that we have no information whatsoever about the dynamics: a stochastic process typically defines some structure, while allowing for some randomness around that. This makes stochastic processes very widely applicable: in many physical, biological, financial (and other) systems, we know what kind of changes to expect, but not to the point that we could predict their evolution exactly. For example, we might assume that a population increases exponentially, but we couldn't predict precisely each individual birth and death. This is the stochastic component.

The study of stochastic processes can be viewed as an application of probability theory, and we will use them as probability models. A probability model is a way to define how different quantities are related probabilistically, i.e., through probability distributions rather than deterministic relationships. Our main focus will be on describing such processes as mathematical objects, study their properties, and understand what types of real-world situations they can represent.

Defining a model is a separate question to the problem of statistical analysis and inference, which typically refers to the estimation of the parameters of a probability model from observed data. This will only be a secondary focus in some of the chapters of this course, but it is a very important topic in practice, as it is necessary to use stochastic processes to understand and predict things around us.

Example 1.1: Population modelling

Thomas Austin was an English settler in Australia, who released 29 rabbits there in 1859 for recreational hunting. The non-native species thrived and, within years, the population grew to several millions. We can view the number of rabbits as a random process, where births depend on the reproduction rate and deaths depend on predation, disease, etc. Each birth and death is unpredictable, but we might be able to model the overall population fairly well. We will talk about birth-death processes, which formalise the ideas presented here, in a later chapter.

It is common to model population growth as exponential, but we might not expect it to follow an exponential curve exactly (e.g., some years might be warmer than others). These irregularities can also be captured by the stochastic process. Figure 1.1 shows five example simulated populations starting from 19 individuals, all assuming the same birth and death rates. The five simulations look different because of the randomness, but they all follow an exponential growth.

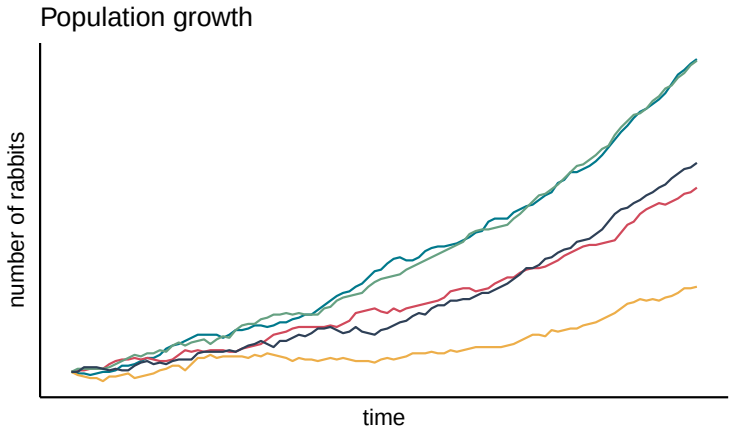


Figure 1.1: *Simulated example of rabbit population dynamics. The five lines correspond to five simulations, all with the same initial number (19 rabbits) and the same birth and death rates.*

In a situation like this, we might be interested to answer questions such as:

- how long do we expect it will take to reach 100,000 rabbits?
- what is the probability that the rabbit population will die out?

Example 1.2: SIR model

SIR models are very common in the study of infectious diseases. The three letters stand for "susceptible", "infectious", and "recovered", because the model assumes that each person is in one of those three categories. The dynamics are usually parameterised in terms of a transmission rate (susceptible to infectious) and a recovery rate (infectious to recovered). These parameters are linked to the R_0 value, which is often used to summarise how infectious a disease is. Note that we can think of the "recovered" group as also including those who have died from the disease.

Figure 1.2 shows an example of how the number of infected people might change through time, according to an SIR model. It starts with a fast increase in the number of infections, as the number of susceptible people is still large (i.e., nobody has immunity yet). After some time, after a large part of the population has either died or gained immunity, the number of infectious people starts decreasing, eventually towards zero.

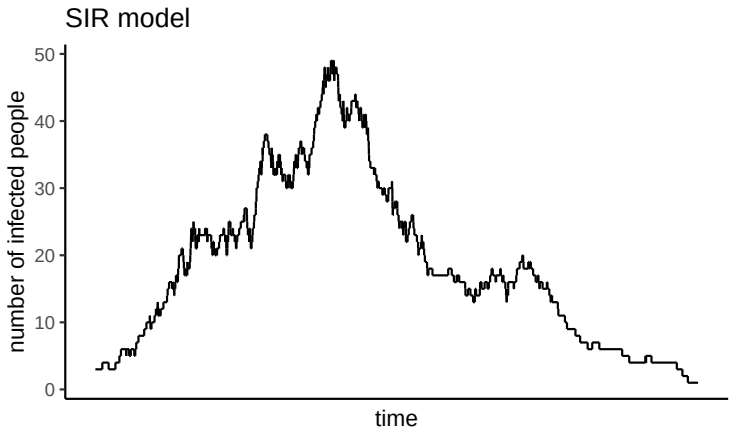


Figure 1.2: *Simulated example of SIR disease spread model, for a population of 300 individuals.*

We could use a model like this to investigate many important questions, for example:

- when will the number of infected people peak?
- how high can we expect the peak to be?
- how long will it take for 70% of the population to recover?

Of course, this is a very simple model, and there are many extensions, e.g., where immunity does not last forever, or where it is possible to gain immunity through vaccination rather than infection. Many of the models used to analyse the spread of Covid-19 were based on extensions of the SIR model.

Example 1.3: Quality of my lectures

Assume that I am not a very consistent instructor, and that some of the lectures I give are Good and some of them are Bad. If a given lecture is Good, then there is a 90% chance that the next lecture will be Good too, and a 10% chance that it will be Bad. If a lecture is Bad, there is a 50% chance that the next lecture will be Good or Bad. In this case, unlike the population and disease spread examples above, there isn't a numeric variable that we can quantify through time. Instead, the variable of interest is "quality of a lecture", which is binary ("Good"/"Bad"). A simulated example might look something like (Good, Good, Good, Bad, Good, Bad, Bad, Good, Good, …)

This is also a problem that we can study using stochastic processes. For example, we might be interested to know:

- if I give a Good lecture today, what is the probability that I give a Good lecture again in a week?
- in the long run, what will be the proportion of Good and Bad lectures?

We could also think about situations in which stochastic processes are not appropriate. For example, classical mechanics can describe the position of planets over time very precisely, and so there is no need for a probabilistic model.

We will start by reviewing some results from probability theory, that will be useful in later chapters.

1.2 Probability review

Probability is a mathematical tool used to describe randomness. The real-world interpretation of probability and randomness has been a long-standing philosophical question that we will not discuss in this course. Whether we consider that randomness is a feature of the world or that it emerges from our (lack of) knowledge does not affect our study of probability models.

1.2.1 Sample space, events, and so on

Although it is most often implicit in practice (and in this course), the description of probability usually starts from the notion of an **experiment**, i.e., some procedure with a set of possible **outcomes**. For example, the experiment could be measuring the length of an object (outcome = positive number), throwing a die (outcome from 1 to 6), a general election (outcome $\in$ {Conservative Party wins, Liberal Party wins, NDP wins, ... }), etc.

The **sample space** Ω is the set of all possible outcomes of the experiment, and an element $\omega \in \Omega$ is called an outcome. A subset of Ω, $A \subset \Omega$, is called an **event**. In the general election example, {Liberal Party wins, NDP wins} is an event.

We can assign a **probability** $\Pr(A)$ to each event A. Mathematically, probability is defined by three axioms, that you can easily find online. In practice, we will think of the probability in the common sense, as a measure of uncertainty of the event. Because events are sets, set theory is often a convenient way to think about probability, e.g., using Venn diagrams. For example, this helps formalise ideas such as the intersection of two events ("A and B"), the union of two events ("A or B"), the complement of an event ("not A"), etc.

1.2.2 Random variables

The building block of probability models is the **random variable**, which can be informally defined as a function that associates a numeric value to each possible outcome of a random experiment. We do not know what value the random variable will take before we measure it, but we can quantify the probability of it taking different values. These probabilities define the **distribution** of the random variable.

If the range of values that the random variable can take is a countable set (such as the integers), then we say that it is a **discrete** random variable. The distribution of a discrete random variable associates a probability to each number in the range. For example, consider the number of emails that you receive on a given day, which is defined over $\{0, 1, 2, ... \}$. We don't know what the number will be in advance, but we might be able to give a probability to each possible value

(e.g., the probability of getting 0 emails is 0.01, the probability of getting 1 email is 0.05, and so on). Note that a countable set can be finite (e.g., $\{1, 2, 3, 4, 5, 6\}$) or infinite (e.g., $\mathbb{N}$).

If the range is uncountably infinite (such as the real line), then the random variable is **continuous**. The probability that a continuous variable take any given value in the sample space is zero, but we can instead use the probability of falling within some interval. For example, consider the random variable that measures the height of a given student in this class, defined over $\mathbb{R}^+$ (or arguably some subset of it). Its distribution could be used to make statements such as: the probability of being between 150 and 160 cm is 0.3, the probability of being between taller than 240 cm is 10^{-10}, etc.

Terminology and notation

Let X be a random variable with range $\mathcal{S}$. We use the following notation:

- $\Pr(X = x)$ is the probability that X equals $x \in \mathcal{S}$ (particularly useful for a discrete random variable);
- $\Pr(x_1 \leq X \leq x_2) = \Pr(X \in [x_1, x_2])$ is the probability that X is between $x_1 \in \mathcal{S}$ and $x_2 \in \mathcal{S}$ (particularly useful for a continuous random variable).

Discrete case:
When X is a discrete random variable, its probability distribution (or simply "distribution") is a vector $u = (u_0, u_1, \dots)$, of length the size of $\mathcal{S}$. Each entry is defined as $u_i = \Pr(X = s_i)$, where s_i is the i^{th} element in $\mathcal{S}$.

When viewed as a function of x, we call $\Pr(X = x)$ the **probability mass function** of X.

Continuous case:
When X is a continuous random variable, we denote as $f_X(x)$ the **probability density function** of X evaluated at x. That is, f_X is the non-negative function defined by

$$\Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\, dx.$$

In some of the later chapters, we might use notation like $f(X = x)$ instead, because it makes some long equations easier to read, but it makes the dependence of the function f on the random variable X implicit.

By definition, the sum of the probabilities of all possible outcomes must equal 1, which implies the following constraints on the probability distribution,

- Discrete case:

$$\sum_{x \in \mathcal{S}} \Pr(X = x) = 1;$$

- Continuous case:

$$\int_S f_X(x)\ dx = 1.$$

It is important to distinguish between a random variable and a realisation from this random variable. The former is defined by a probability distribution, whereas the latter is a single numeric value. It is common to use an uppercase letter to denote a random variable ("X") and a lowercase letter to denote its realisations ("x").

1.2.3 Basic definitions and properties

The concepts presented above can be used to describe the distribution of a single random variable. To represent real-world systems, however, we usually need several random variables, and we need a way to express relationships between them using the language of probability. In this context, we distinguish between three different types of distributions:

- the joint distribution (multivariate distribution of the random variables);
- the marginal distribution (distribution of one random variable regardless of the other ones);
- the conditional distribution (distribution of one random variable when the other random variables are known).

In this section, we describe conditional probability and conditional distributions in the context of two random variables. Several definitions and properties are presented in three different forms: for two events, for two discrete random variables, and for two continuous random variables. Note that the three formulas need not be memorised separately, as it is easy to go from one to another.

Definition 1.1

For two events A and B, the **conditional probability** of A given B is

$$\Pr(A \mid B) = \frac{\Pr(A, B)}{\Pr(B)},$$

where $\Pr(A, B)$ is the (joint) probability of events A and B (sometimes denoted as $\Pr(A \cap B)$).

In practice, the "events" that we will study will always be statements about the value of a random variable, e.g., $X = 1$ or $1.5 \leq X \leq 2$. This leads us to define the conditional distribution of a random variable given another random variable. For the continuous case, we must first define the joint density function of two random variables.

Definition 1.2

Let X and Y be two continuous random variables with range the real line. The **joint probability density function** of X and Y is the non-negative function $f_{X,Y}$ defined as

$$\Pr(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) dv du,$$

for all $x, y \in \mathbb{R}$.

Definition 1.3

Let X and Y be two random variables.

- If X and Y are discrete, the **conditional probability mass function** of Y given $X = x$ is
$$\Pr(Y = y \mid X = x) = \frac{\Pr(X = x, Y = y)}{\Pr(X = x)}$$
- If X and Y are continuous, the **conditional probability density function** of Y given $X = x$ is
$$f_{Y|X}(y \mid X = x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Example 1.4: Conditional mass function

For two discrete random variable, one way to understand the connection between the joint and conditional distributions is by writing them as a table. Consider the random variables X and Y with joint distribution defined by the following table.

	$\mathbf{X = 0}$	$\mathbf{X = 1}$	$\mathbf{X = 2}$
$\mathbf{Y = 0}$	0.2	0	0.3
$\mathbf{Y = 1}$	0.1	0.3	0.1

Each entry gives the probability of some combination of values for X and Y, e.g., $\Pr(X = 0, Y = 0) = 0.2$, $\Pr(X = 1, Y = 0) = 0$, and so on. All entries sum to 1, indicating that these are the only possible combinations for X and Y.

We can sum over the rows (respectively, columns) to get the marginal distribution of X (respectively, Y). The marginal distribution is just the unconditional distribution of the random variable. So, for X, we get the probability distribution $(0.3, 0.3, 0.4)$ by summing over the rows of the table.

The conditional distribution of Y given X is obtained by considering just one column of the table at a time, and dividing it by the sum of its entries to obtain a vector of probabilities that sum to 1. So, for example, the conditional distribution of Y given $X = 2$ is

$$\left(\frac{0.3}{0.3 + 0.1}, \frac{0.1}{0.3 + 0.1}\right) = (0.75, 0.25).$$

Similarly, we could use the table to get the conditional distribution of X given Y. For example, given $Y = 0$, the conditional distribution of X is

$$\left(\frac{0.2}{0.2 + 0 + 0.3}, \frac{0}{0.2 + 0 + 0.3}, \frac{0.3}{0.2 + 0 + 0.3}\right) = (0.4, 0, 0.6).$$

Example 1.5: Conditional density function

For two continuous random variables, we can't write the joint distribution as a table, but we can visualise $f_{X,Y}$ as a heatmap. Figure 1.3 shows an example where X and Y jointly follow a bivariate normal distribution with correlation 0.8. Areas where the joint density function is high correspond to combinations of X and Y that are more likely, even though any given combination has probability zero. In this example, it is more likely for (X, Y) to be close to the origin than to $(2, -2)$, for example.

In the continuous case, the conditional distribution can be viewed as a slice through the joint distribution (analogous to taking one column at a time in the discrete example). So, for example, to get the conditional distribution of Y given $X = 1$, we take a slice of the joint density function at $x = 1$, and we normalise the resulting function so that it integrates to 1. The resulting conditional density function is shown in the right panel of Figure 1.3. In this case, there is a formula that gives us the conditional distribution, and it turns out that $Y \mid X$ follows a (univariate) normal distribution.

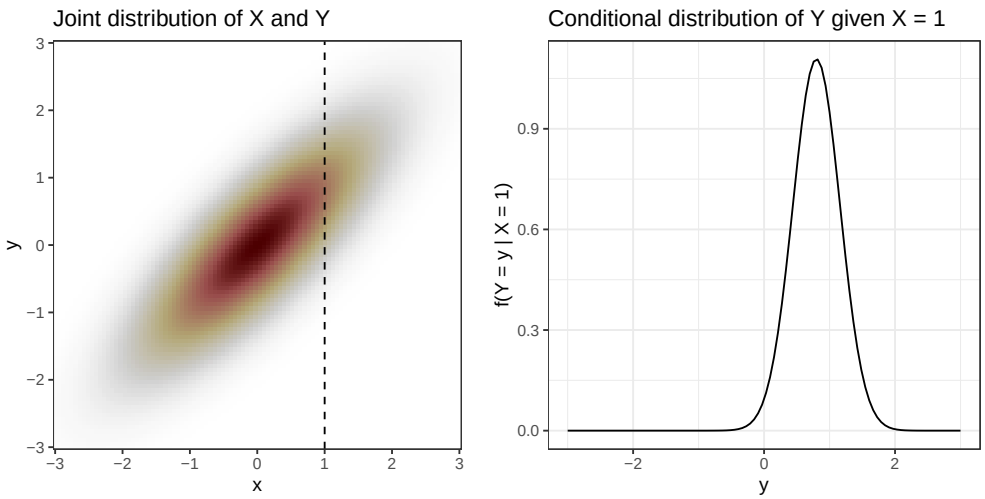


Figure 1.3: *Illustration of joint (left) and conditional (right) density functions for two random variables.*

The notion of "independence" of two random variables has an intuitive interpretation: it asserts that knowing one random variable gives us no information about the other one. For example, if a coin is flipped twice, we usually assume that knowing the outcome of the first flip does not help predict the outcome of the second flip. Two events A and B are said to be independent if $\Pr(A, B) = \Pr(A)\Pr(B)$, and an analogous definition exists for two random variables.

Definition 1.4

Let X and Y be two random variables with range $\mathcal{S}$. X and Y are **independent** if, for all $x, y \in \mathcal{S}$,

- Discrete case:
$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y);$$
- Continuous case:
$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

The definition of independence can also be rewritten in terms of conditional probability, which makes the connection to the common language definition of the term more explicit. If X and Y are independent then, in the discrete case,

$$\Pr(X = x \mid Y = y) = \Pr(X = x),$$

and, in the continuous case,

$$f_{X|Y}(x \mid Y = y) = f_X(x).$$

That is, the conditional distribution of $X \mid Y$ and the marginal distribution of X are the same. Independence is symmetric, so we could also rewrite this to find that the conditional distribution of $Y \mid X$ and the marginal distribution of Y are equal. In both cases, knowing one of the two random variables does not change the distribution of the other one.

The law of total probability is a useful result to obtain the marginal probability of an event, based on conditional probabilities.

Proposition 1.1: Law of total probability

Let $B_1, B_2, \ldots, B_K$ be events that partition the sample space. That is, the B_i are mutually exclusive (pairwise intersections are empty), and their union is equal to the sample space. Then, for any event A,

$$\Pr(A) = \sum_{k=1}^{K} \Pr(A \mid B_k) \Pr(B_k).$$

The law of total probability can also be written in terms of probability density/mass functions. Let $X \in \mathcal{S}_X$ and $Y \in \mathcal{S}_Y$ be two random variables. In the discrete case, we have

$$\Pr(Y = y) = \sum_{x \in \mathcal{S}_X} \Pr(Y = y \mid X = x) \Pr(X = x),$$

and, in the continuous case,

$$f_Y(y) = \int_{\mathcal{S}_X} f_{Y|X}(y \mid X = x) f_X(x) \, dx,$$

for all $y \in \mathcal{S}_Y$.

Example 1.6

Let's say that you go to sleep before 11pm 40% of the time, between 11pm and 1am 40% of the time, and after 1am 20% of the time. The probability that you arrive late to the morning lecture is 0.05 if you go to sleep before 11pm, 0.2 if you go to sleep between 11pm and 1am, and 0.5 if you go to sleep after 1am.

What is the probability that you will be late to the morning lecture on a given day?

We first need to introduce some mathematical notation. Let X be the random variable equal to 0, 1, or 2 if you go to sleep before 11pm, between 11pm and 1am, or after 1am, respectively. Let Y be the random variable equal to 0 if you are late and 1 if you are not. We can now write the information in the question as probability statements about X and Y:

$$\begin{cases} \Pr(X = 0) = 0.4 \\ \Pr(X = 1) = 0.4 \\ \Pr(X = 2) = 0.2 \end{cases} \quad \text{and} \quad \begin{cases} \Pr(Y = 0 \mid X = 0) = 0.05 \\ \Pr(Y = 0 \mid X = 1) = 0.2 \\ \Pr(Y = 0 \mid X = 2) = 0.5 \end{cases}$$

We are interested in $\Pr(Y = 0)$, which we can calculate using the law of total probability:

$$\begin{aligned}\Pr(Y = 0) &= \sum_{x=0}^{2} \Pr(Y = 0 \mid X = x) \Pr(X = x) \\ &= 0.05 \times 0.4 + 0.2 \times 0.4 + 0.5 \times 0.2 \\ &= 0.2\end{aligned}$$

That is, there is a 20% probability that you will be late on a given day.

1.3 Introducing stochastic processes

Definition 1.5

A **stochastic process** (X_t) is a collection of random variables, indexed by time t.

If t is defined over a countable set (e.g., $t \in \{0, 1, 2, \dots\}$), then we call (X_t) a **discrete-time** stochastic process.

If t is defined over an uncountable set (e.g., $t \in \mathbb{R}^+$), then (X_t) is a **continuous-time** stochastic process.

We will study discrete-time processes in Chapters 2 and 5, and continuous-time processes in Chapters 3 and 4. In the discrete-time case, we will generally use the letter n to index time, i.e., we will write the stochastic process as (X_n). Discrete-time processes usually rely on the assumption that the random variables in the sequence are on a regular time grid, and so we almost always define n over the non-negative integers, so that one time interval corresponds to one unit of time.

The simplest possible stochastic process is a sequence of independent, identically distributed random variables $\{X_1, X_2, X_3, \dots\}$, e.g., $X_t \sim N(0, 1)$. This is not a particularly interesting process, and we will usually focus on cases where there is some dependence between the X_t, as this is key to model the evolution of many real-life phenomena. We will discuss how the dependence can be modelled in later chapters.

Notation

We denote a stochastic process as (X_t), or X, or $(X_t)_{t \geq 0}$. This refers to the process, whereas X_t refers to the value of the process at time t.

We often visualise stochastic processes using line graphs with time on the x axis and the value of the process on the y axis, like in the population modelling example at the beginning of this

chapter. The line could for example be obtained by simulating from the stochastic process, or it could be a real data set that we would like to analyse using the stochastic process as a model. Just like it is important to distinguish between a random variable and the value it takes, it is important to separate a stochastic process and the line on that graph. We can think of the stochastic process as the recipe that tells us how to generate the lines. We call each line a realisation, or a sample path, from the stochastic process.

Terminology

A stochastic process is often also called a **random process**, and we will use these two terms interchangeably.

Another related concept is the **time series**. There is no universally-accepted definition of a time series, but the term is most often used to refer to a realisation from a discrete-time stochastic process. Sometimes, a stochastic process is called a time series process, but this is less common. We often talk of time series data, however, to refer to a series of observations made at regular time intervals (e.g., temperature over time).

Definition 1.6

The **state space** $\mathcal{S}$ of a stochastic process (X_t) is the set of values that X_t can take, i.e., $X_t \in \mathcal{S}$.

If $\mathcal{S}$ is countable, then we call (X_t) a **discrete-space** (or discrete-valued) stochastic process.

If $\mathcal{S}$ is uncountable, then we call (X_t) a **continuous-space** (or continuous-valued) stochastic process.

It is important to distinguish between the set over which t is defined, and the state space $\mathcal{S}$ over which X_t is defined. Whether each is discrete or continuous is unrelated, and should be assessed separately.

Example: We can think of situations where either space and/or time should be discrete or continuous.

	discrete time	**continuous time**
discrete space	chess	disease status
continuous space	stock price	particle movement

We will spend more time focusing on discrete-space stochastic processes, as they tend to be a little easier to study, but many of the results can be extended to continuous space.

Note that, while time is always a quantitative (discrete or continuous) variable, X_t can be:

1. qualitative; e.g., "sick"/"healthy", or position of piece on chess board;
2. quantitative discrete; e.g., number of students in a lecture;
3. quantitative continuous; e.g., daily temperature.

When the process is qualitative, we usually write $\mathcal{S} = \{0, 1, 2 \dots\}$ for simplicity; you can simply think of X_t as the random variable that associates a non-negative integer value to each possible category (e.g., "sick" is 0 and "healthy" is 1).

2 Discrete-time Markov processes

We first consider the case of a discrete-time stochastic process $(X_n)_{n\in\mathbb{N}}$, defined over a countable state space $\mathcal{S}$ (i.e., $X_n \in \mathcal{S}$). In this chapter, we denote the time index n rather than t as a reminder that time is discrete. We will later cover the case of an uncountable state space (Section 2.6), and continuous-time processes (Chapter 4).

2.1 Introduction

A key feature of a stochastic process is its dependence structure, i.e., how successive values of the process depend on each other. The simplest assumption would be that the X_n are independent random variables, but this is an unrealistic premise in many situations. The next simplest assumption would be that X_{n+1} is dependent on X_n, but not on previous values of the process (at least, not conditional on X_n). This is called the Markov assumption, and we will see that it captures the dependence of many real-world processes, despite its apparent simplicity.

2.1.1 Definition

Definition 2.1

A discrete-time stochastic process (X_n) is called a **Markov process** if it satisfies

$$\Pr(X_{n+1} = j \mid X_n, X_{n-1}, \dots, X_0) = \Pr(X_{n+1} = j \mid X_n),$$

for all $j \in \mathcal{S}$. This is called the Markov property or Markov assumption.

There are several ways we could describe this property in words:

- X_n contains all the information we need about the history of the process to determine the distribution of X_{n+1};
- the future is independent of the past, conditionally on the present;
- the process has no memory (the Markov property is sometimes called "memorylessness").

Example 2.1

1. Perhaps the most common textbook example of a Markov chain is a simplified weather model. Let's assume that the weather on a given day is either sunny or cloudy, and that tomorrow's weather depends on today's weather, but that is is independent of previous days (conditionally on today). This system can be modelled with a Markov chain with state space $\mathcal{S} = \{\text{sunny}, \text{cloudy}\}$.

2. Consider the following game. You repeatedly throw a die; if it falls on 6, you win \$10 and, if it falls on any other number, you lose \$1. Let X_n be the amount you have won (or lost) after n rounds, starting from $X_0 = 0$. The process (X_n) satisfies the Markov property because, if you know X_n, knowing $X_{n-1}, X_{n-2}, ...$ does not give you any extra information to predict what X_{n+1} will be. (Try to write the conditional distribution of X_{n+1} given X_n in this example.)

By abuse of language, the Markov property is sometimes described by saying that X_{n+1} only depends on X_n and not on previous values of the process $(X_{n-1}, X_{n-2}, ...)$. Note that, in this version, there is no explicit mention of the conditional nature of this statement. If we do not condition on X_t, then X_{n+1} is in fact dependent on X_{n-1}, as will become clearer later.

We often use the word "**state**" when describing Markov processes, and it can refer to two things:

- the state of the process at time n is its value X_n;
- the states of the process are elements of its state space (e.g., "sunny" and "cloudy" in the weather example).

We can then rephrase the Markov property as: the state of the process at time $n + 1$ is independent of its state at time $n - 1$ (and before), conditional on its state at time n.

In the definition of a Markov process that we gave above, the process is assumed to be defined over a countable set $\mathcal{S}$. An important special case is when $\mathcal{S}$ is finite, i.e., the process can take on a finite number of values. In such a case, we sometimes refer to a N-state Markov process for a process defined over a set of size N. For Markov processes over countable (finite or infinite) sets, it is often convenient to denote the state space using the integers, e.g., $\mathcal{S} = \{0, 1, 2\}$ or $\mathcal{S} = \mathbb{Z}$. Despite this notation, it is important to remember that, in many applications, the states are qualitative rather than quantitative, and the integers are merely indices to distinguish them. We can also sometimes label the states using letters or some other symbols to make this more explicit.

Example 2.2: random walks

Random walks are a broad class of stochastic processes, which are widely used in science. Here we consider two examples, to illustrate Markov processes with finite or infinite state spaces.

1. **Finite state space:** Consider the position of a player's token on a Monopoly board. The board has 40 squares in total and, at each round, the player moves between 2 and 12 squares forward based on the throw of two dice (ignoring other special rules). The board is a loop, so the 40th square is next to the 1st. If we let $X_n \in \{0, 1, \dots, 39\}$ be the position of the token after n rounds, then (X_n) is a Markov process.

2. **Countably infinite state space:** Let X_n be the number of pandas in the world on day n, and assume that, on a given day, the population can increase by 1 with probability p, decrease by 1 with probability q, or stay the same with probability $1 - p - q$. If the population decreases to 0, then it cannot change anymore. This defines a Markov process over the non-negative integers.

Terminology: Markov process and Markov chain

The term "Markov chain" is used very commonly, but its meaning is somewhat inconsistent. Depending on the source, it can refer to:

1. any Markov process;
2. a Markov process with discrete state space;
3. a Markov process in discrete time.

In these notes, we follow the first interpretation, and use Markov chain interchangeably with Markov process.

Due to the Markov property, the dynamics of a Markov process with countable state space can be specified in terms of the probabilities of moving between any two states over one time interval, given by $\Pr(X_{n+1} = j \mid X_n = i)$ for any $i, j \in \mathcal{S}$. If the state space is of finite size $|\mathcal{S}| = N$, there are N^2 such probabilities.

Definition 2.2

The one-step **transition probability** from state i to state j is

$$P_{ij} = \Pr(X_{n+1} = j \mid X_n = i)$$

for $i, j \in \mathcal{S}$.

The one-step **transition probability matrix** is

$$P = \begin{pmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ P_{20} & P_{21} & P_{22} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

We also call it the transition matrix.

Remark: The transition probabilities are subject to the following constraints.

- $P_{ij} \in [0, 1]$, for any $i, j \in \mathcal{S}$
- $\sum_{j \in \mathcal{S}} P_{ij} = 1$, for any $i \in \mathcal{S}$

The first point follows from the definition of probabilities, and the second point reflects the necessity that $X_{n+1} \in \mathcal{S}$ (i.e., there must be a $j \in \mathcal{S}$ for which $X_{n+1} = j$). A square matrix whose entries satisfy those two conditions is called a stochastic matrix (or sometimes a "right" stochastic matrix). Each row of a stochastic matrix is a (discrete) probability distribution.

If the state space of the Markov chain is finite, i.e., $|\mathcal{S}| = N < \infty$, then the transition probability matrix is an $N \times N$ matrix. In the case of an countable infinite state space, however, the matrix is infinite (i.e., it has an infinite number of rows and columns). Although this seems to complicate things, most matrix operations are still well defined for infinite matrices, and the results described below hold for any countable state space. In practice, the main challenge associated with the infinite state space is that we cannot write out the full matrix, either by hand or to store it in a computer.

In this chapter, we will assume that the transition probabilities do not depend on the time step n, i.e., they are constant through time.

Definition 2.3

A Markov chain is **time-homogeneous** if, for any $n = 0, 1, ...,$

$$\Pr(X_{n+1} = j \mid X_n = i) = \Pr(X_1 = j \mid X_0 = i)$$

where $i, j \in \mathcal{S}$.

That is, the probability of going from state i to state j does not depend on n.

It is common to represent a Markov chain as a **transition graph**, with one node for each state, and arrows showing all possible transitions. This can be viewed as a weighted graph, where the weight of each edge is the corresponding transition probability.

Example: Consider the 3-state Markov chain with transition probability matrix

$$P = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.3 & 0.7 \\ 0.1 & 0 & 0.9 \end{pmatrix}$$

If we label the three states as "a", "b", and "c", we can represent the transition structure of the process as shown in Figure 2.1.

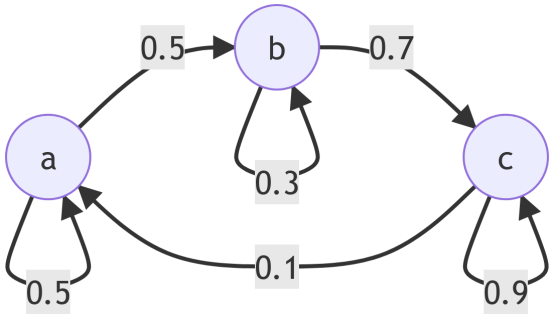


Figure 2.1: *Example transition graph of 3-state Markov chain*

A realisation from a Markov chain is a sequence of states for some set of time indices. For example, (a, a, b, c, c, c, c, c, c, a) is one possible realisation of the 3-state Markov chain shown in Figure 2.1 over 10 time steps. We can display those as time series graphs, with time along the x axis and state along the y axis, as long as we remember that the ordering of the states is often arbitrary. Four example realisations from the Markov chain of Figure 2.1 are shown in Figure 2.2.

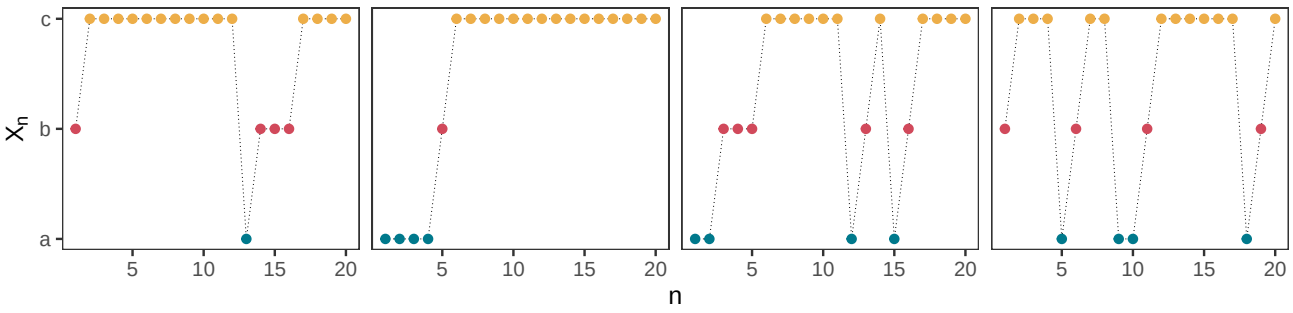


Figure 2.2: *Four realisations of a 3-state Markov process, where the initial state was chosen at random.*

Figure 2.2 makes it clear that, if we don't condition on X_n, then X_{n+1} is dependent on previous states. For example, if all we know is that $X_{n-1} = \text{c}$ (and we don't know X_n), this still gives us quite a bit of information about X_{n+1}. In this example, we know that X_{n+1} is most likely to also be c, because the process tends to stay in that state for many consecutive time steps.

2.1.2 Holding times

One way to understand the Markov assumption, and whether it is violated in a given context, is to think about how long the process spends in a given state (before switching to another state). Let D_i denote the number of consecutive time steps spent in state i, called the holding time (or dwell time). The event $D_i = 1$ corresponds to the situation where the process switches out of state i in the first time step, which has probability $1 - P_{ii}$ (i.e., one minus the probability of remaining in state i). The event $D_i = 2$ requires remaining in state i in the first time step (with probability P_{ii}) and switching out of state i in the second time step (with probability $1 - P_{ii}$), so it has probability $P_{ii}(1 - P_{ii})$. We can repeat this reasoning to find the general formula:

$$\Pr(D_i = k) = P_{ii}^{k-1}(1 - P_{ii}),$$

because $D_i = k$ means that the process remained in state i for $k - 1$ time steps, and then switched to another state.

This is the probability mass function of the geometric distribution with parameter $1 - P_{ii}$. Figure 2.3 shows the graph of this function for different values of the probability parameter $p = 1 - P_{ii}$. Although the decay rate of the distribution depends on the transition probability, its mode is always 1, i.e., the most likely holding time is 1 regardless of the transition probabilities.

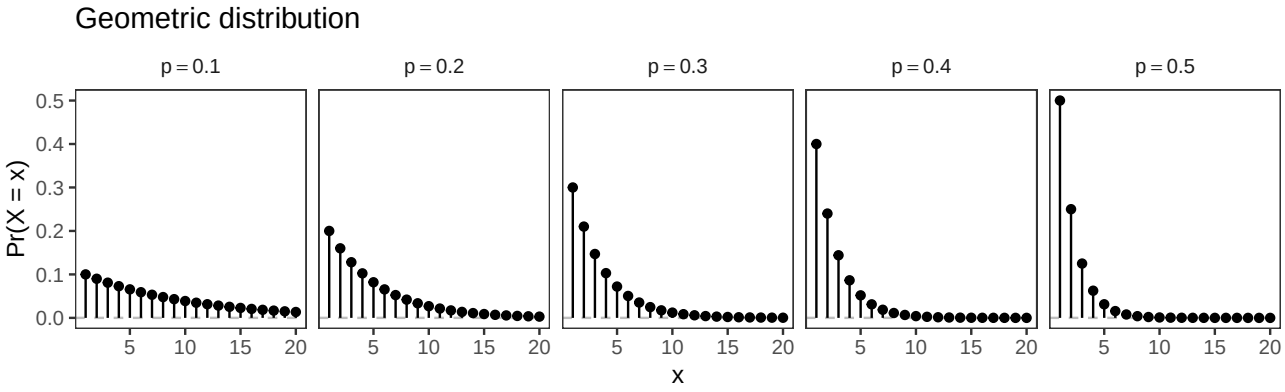


Figure 2.3: *Probability mass function of the geometric distribution, for different values of the probability parameter.*

The mean of the geometric distribution is the inverse of the probability parameter so, in the context of the Markov chain,

$$E[D_i] = \frac{1}{1 - \gamma_{ii}}$$

For example, the expected holding times for the Markov chain shown in Figure 2.1 are $E[D_1] = 1/(1-0.5) = 2$, $E[D_2] = 1/(1-0.3) = 1.43$, and $E[D_3] = 1/(1-0.9) = 10$. This is consistent with the patterns observed in the simulated realisations of Figure 2.2, where the process tends to spend much longer in state 3 than in states 1 and 2.

2.1.3 Higher-order dependence

The Markov property might seem like a strong assumption in many situations. After all, most real-world processes have very complex dependence structures. For example, tomorrow's weather likely depends on more than just today's weather. But it is important to remember that stochastic models, like any models, are only supposed to be an approximation. The Markov property turns out to be a pretty good approximation to many complex phenomena.

There are several ways to relax the Markov assumption while preserving some of the convenient mathematical properties of Markov chains.

Definition 2.4

A discrete-time stochastic process (X_n) is called a **p-th order Markov process** if it satisfies

$$\Pr(X_n = j \mid X_{n-1}, X_{n-2}, \dots, X_0) = \Pr(X_n = j \mid X_{n-1}, \dots, X_{n-p}),$$

for all $j \in \mathcal{S}$.

Higher-order Markov processes might seem considerably more flexible than (first-order) Markov processes, but they are also harder to implement. Fortunately, they can be written as first-order Markov processes with an expanded state space, such that all results in this chapter can be applied to them directly. To convert a p-th order Markov chain into a first-order Markov chain, we can define the new state space to be the set of all possible sequences of p states (let's call these new states "expanded states"). So, we will focus on first-order Markov processes, keeping in mind that they can be used very generally.

Example 2.3

For example, if we have a second-order Markov chain with states {A, B, C}, the expanded state space would be {AA, AB, AC, BA, BB, BC, CA, CB, CC}. Once we have defined the expanded state space, we can define a first-order Markov chain where each state is one of the expanded states (i.e., a sequence of p states). By expanding the state space in this way, we can convert a higher-order Markov chain into a first-order Markov chain, which is easier to study. Note that not all transitions are possible in this new first-order Markov chain; for example, the process cannot transition from AA to BA.

If we label the states from 1 to 9 in the order shown above, the transition probability

matrix for this example would be

$$P = \begin{pmatrix} P_{11} & P_{12} & P_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{24} & P_{25} & P_{26} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & P_{37} & P_{38} & P_{39} \\ P_{41} & P_{42} & P_{43} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{54} & P_{55} & P_{56} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & P_{67} & P_{68} & P_{69} \\ P_{71} & P_{72} & P_{73} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & P_{84} & P_{85} & P_{86} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & P_{97} & P_{98} & P_{99} \end{pmatrix},$$

where $P_{11} = \Pr(X_n = \text{AA} \mid X_{tn1} = \text{AA})$, $P_{12} = \Pr(X_n = \text{AB} \mid X_{n-1} = \text{AA})$, and so on.

This approach can in principle be used to represent any high-order Markov process, but the state space grows rapidly with the number of states and the order.

2.1.4 Simulating from a Markov process

Given an initial distribution and a transition probability matrix, we can simulate from a Markov chain by iteratively sampling from a categorical distribution, e.g., using the `sample()` function in R. The code chunk below shows an example simulation over 100 time steps, for a 3-state Markov chain with state space $\mathcal{S} = \{0, 1, 2\}$, initial distribution

$$u^{(0)} = (0.2, 0.2, 0.4)$$

and transition probability matrix

$$P = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.3 & 0.7 & 0 \\ 0 & 0.2 & 0.8 \end{pmatrix}$$

This outputs one realisation from the process, and changing the random seed would yield a different realisation.

```
# Set random seed for reproducibility
set.seed(67)

# Define parameters
n <- 100
u0 <- c(0.2, 0.2, 0.4)
```

```r
P <- matrix(c(0.8, 0.1, 0.1,
              0.3, 0.7, 0,
              0, 0.2, 0.8),
            nrow = 3, byrow = TRUE)

# Initialise
X <- rep(NA, length = n)
X[1] <- sample(1:3, size = 1, prob = u0)

# Loop over time steps
for(i in 2:n) {
    # Choose row of transition matrix based on previous state
    P_row <- P[X[i-1],]
    # Sample new state
    X[i] <- sample(1:3, size = 1, prob = P_row)
}

# Minus 1 to get states {0, 1, 2} rather than {1, 2, 3}
X - 1
```

```
 [1] 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 2 2 2 2 2 2 2 1 1 1 1 1 1
[38] 1 1 1 1 1 0 0 0 0 2 2 2 2 2 2 2 2 1 0 0 0 0 2 2 1 1 1 1 1 1 1 1 1 0 1 1 1
[75] 0 0 2 2 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

2.2 Looking into the future

The transition probabilities describe what happens to the process over one time interval. With this information, it seems that we should also be able to say something about the distribution of the process further into the future (although perhaps with less and less certainty).

2.2.1 Chapman-Kolmogorov Equations

Definition 2.5

The n-**step transition probability** from state i to state j is

$$P_{ij}^{(n)} = \Pr(X_{m+n} = j \mid X_m = i),$$

for any $m \geq 0$. It is the probability that the process will be in state j after n transitions, given that it started in state i.

The n-step transition probability matrix is denoted as

$$P^{(n)} = \begin{pmatrix} P_{00}^{(n)} & P_{01}^{(n)} & P_{02}^{(n)} & \cdots \\ P_{10}^{(n)} & P_{11}^{(n)} & P_{12}^{(n)} & \cdots \\ P_{20}^{(n)} & P_{21}^{(n)} & P_{22}^{(n)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Note that the transition probabilities that we defined previously are 1-step transition probabilities, i.e., we have $P^{(1)} = P$ and $P_{ij}^{(1)} = P_{ij}$. For any n, the n-step transition probabilities can be derived from the 1-step transition probabilities, and the Chapman-Kolmogorov equations provide this relationship.

Proposition 2.2 (Chapman-Kolmogorov equations)

The Chapman-Kolmogorov equations are

$$P_{ij}^{(n+m)} = \sum_{k \in \mathcal{S}} P_{ik}^{(n)} P_{kj}^{(m)} \tag{2.1}$$

for any $i, j \in \mathcal{S}$, and any $n, m \in \mathbb{N}$.

Equivalently, in matrix notation:

$$P^{(n+m)} = P^{(n)} P^{(m)}.$$

Proof

The Chapman-Kolmogorov equations can be viewed as an application of the law of total probability to the Markov chain.

$$\begin{aligned} P_{ij}^{(n+m)} &= \Pr(X_{n+m} = j \mid X_0 = i) \\ &= \sum_{k \in \mathcal{S}} \Pr(X_{n+m} = j, X_n = k \mid X_0 = i) && \text{(a)} \\ &= \sum_{k \in \mathcal{S}} \Pr(X_{n+m} = j \mid X_n = k, X_0 = i) \Pr(X_n = k \mid X_0 = i) && \text{(b)} \\ &= \sum_{k \in \mathcal{S}} \Pr(X_{n+m} = j \mid X_n = k) \Pr(X_n = k \mid X_0 = i) && \text{(c)} \\ &= \sum_{k \in \mathcal{S}} P_{kj}^{(m)} P_{ik}^{(n)} \\ &= (P^{(n)} P^{(m)})_{ij} \end{aligned}$$

Step (a) is the law of total probability, (b) uses the definition of conditional probability, and (c) uses the Markov property.

In particular, we have

$$P^{(2)} = P^{(1)}P^{(1)} = P^1P^1 = P^2,$$
$$P^{(3)} = P^{(2)}P^{(1)} = P^2P^1 = P^3,$$

and so on. By induction, we can prove the following result.

Corrolary

The n-step transition probability matrix of a Markov chain is

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

That is, the n-step transition probability matrix $P^{(n)}$ can be computed by multiplying the transition probability P by itself n times.

Note that this does *not* imply that $P_{ij}^{(n)} = P_{ij}^n$ for any i and j. This is generally not the case, and the full matrix needs to be taken to the power of n to obtain $P_{ij}^{(n)}$.

2.2.2 Marginal state distribution

The transition probabilities of the Markov chain define the conditional distribution of the state X_n given the state X_{n-1}. In some cases, we are interested in the marginal distribution (i.e., unconditional distribution) of X_n.

Notation

For a Markov chain (X_n) with state space $\mathcal{S}$, we denote as $u^{(n)}$ the marginal distribution of the random variable X_n, i.e.,

$$u_i^{(n)} = \Pr(X_n = i), \quad \text{for } i \in \mathcal{S}.$$

Note that we interpret $u^{(n)}$ as a row vector (which will be important for calculations later). By definition of probability distributions, we have $u_i^{(n)} \geq 0$ for all i, and $\sum_{i\in\mathcal{S}} u_i^{(n)} = 1$.

To derive the marginal distribution of X_n, we must fix the initial distribution of the chain, $u^{(0)}$. In practice, we often know what the initial value of the process is, so the initial distribution is set to a vector where all but one entries are zero.

Proposition 2.3

Let (X_n) be a Markov chain with initial distribution $u^{(0)}$ and transition probability matrix P. For all $n \geq 0$, the marginal distribution of X_n is $u^{(n)} = u^{(0)}P^n$.

Proof

$$
\begin{aligned}
u_j^{(n)} &= \Pr(X_n = j) \\
&= \sum_{i \in \mathcal{S}} \Pr(X_n = j \mid X_0 = i)\ \Pr(X_0 = i) \quad \text{(a)} \\
&= \sum_{i \in \mathcal{S}} P_{ij}^{(n)} u_i^{(0)} \quad \text{(b)} \\
&= (u^{(0)} P^n)_j,
\end{aligned}
$$

where (a) is the law of total probability, and (b) follows from the Chapman-Kolmogorov equations.

From this property, we can see that the Markov chain is fully specified by the initial distribution $u^{(0)}$ and the transition probability matrix P. That is, given those two parameters, the distribution of the chain can be computed at any time step $n \geq 0$.

To illustrate the idea of marginal distribution, we use the Markov chain defined by random moves of a piece on a chess board. The state space of the process is the list of squares on the board, i.e., there are $8 \times 8 = 64$ states, and the $64 \times 64 = 4096$ transition probabilities are defined by the rules for the chosen piece. We assume that the piece is moved at each time step to any of the allowed squares with equal probability. For example, Figure 2.4 shows the transition probability matrices for a king and a knight.

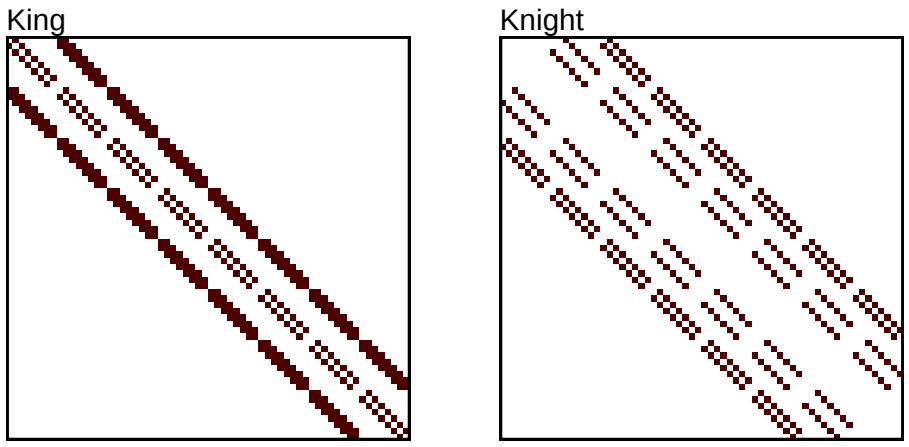


Figure 2.4: *Visualisation of transition probability matrices for king (left) and knight (right) on a chess board, where non-zero elements are shown in dark. Note that the way the squares of the board are ordered as states is arbitrary, so the matrix would look different if another convention was used.*

Given the starting position of the piece, we can use the last proposition to compute its distribution on the board after one move, two moves, and so on. The distribution is a vector of the probabilities of being in the different sequares of the board, which add up to 1. If we choose the square where the piece starts, the initial distribution is a vector of length 64, where 63 elements are set to zero (all expect the starting position). Then, we iteratively multiply that vector by the transition probability matrix to obtain the subsequent distributions.

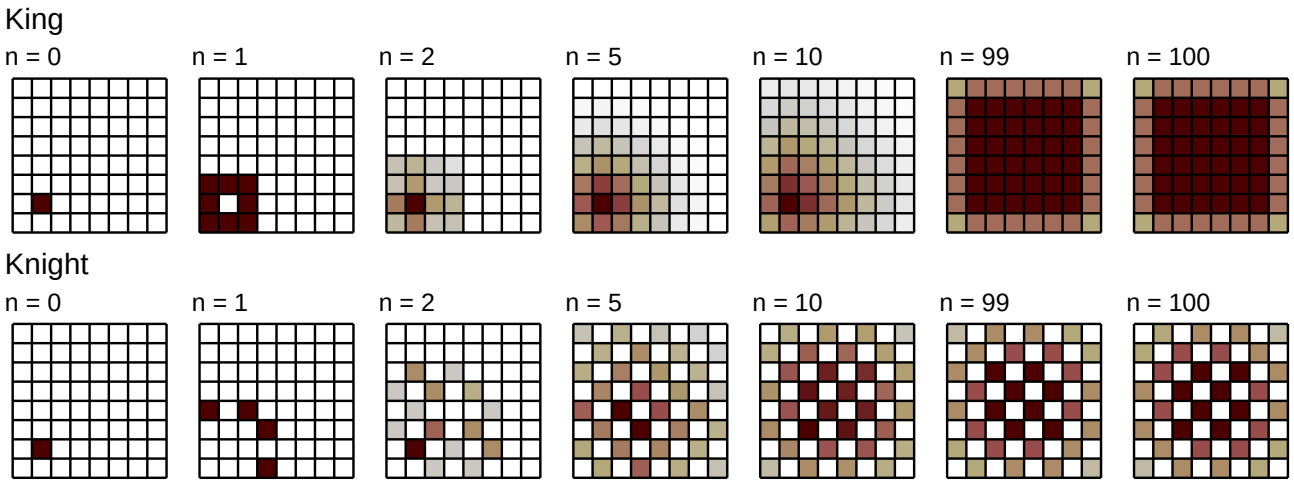


Figure 2.5: *Distribution of position on chess board in successive time steps, given some initial position and movement rules.*

Figure 2.5 shows the distributions of a king and a knight that start from some given square, and take an allowable move at random at each time step. The king has what we might call "diffusive" behaviour, and its distribution spreads over the board with time. After many time steps, it is almost equally likely to be in any of the non-edge squares of the board; the edge and corner squares are less likely because they are less connected. The distribution of the knight also spreads over the board with time, but it follows a different alternating pattern. This is because, due to its movement rules, a knight that's on a black square has to move to a white square at the next time, whereas a knight that's on a white square has to move to a black square. So, if the piece starts on a black square, all white squares have probability zero when n is even, and all black squares have probability zero when n is odd.

This example highlights several interesting phenomena that we will study in more detail in later sections. In particular, it seems like there is a key difference in the long-term behaviour of the Markov chain for the king and the knight: the distribution of the king stabilises as n grows, whereas the distribution of the knight does not.

2.3 Interstate travel

To describe the long-term behaviour of a Markov chain, we need to understand how states are related, i.e., how often the process travels from one state to another (not just over one time interval). In this section, we introduce several definitions that will become important in the next section to decide how the distribution of a given Markov chain evolves in the long run.

2.3.1 Communication and reducibility

Definition 2.6

We say that state j is **accessible** from state i if $P_{ij}^{(n)} > 0$ for some $n \geq 0$, i.e., if there is positive probability of travelling from i to j in a finite number of steps.

Definition 2.7

We say that two states i and j **communicate** if i is accessible from j and j is accessible from i, and we denote this as $i \leftrightarrow j$.

The relation of communication satisfies the following three properties.

1. Reflexivity: Every state communicates with itself.
2. Symmetry: If state i communicates with state j, then j communicates with i.
3. Transitivity: If state i communicates with state j, and state j communicates with state k, then i communicates with k.

A binary relation that satisfies these three properties is called an equivalence relation, and it can be used to divide the state space into equivalence classes.

Definition 2.8

Two states that communicate are in the same **class**. This creates a partition of the state space into communicating classes.

The transition graph of a Markov process can be used to identify communicating classes.

Example: Figure 2.6 shows a 5-state Markov process with states {A, B, C, D, E}, where the transition probability matrix is

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

This Markov process has two communicating classes: {A, B, C} and {D, E}. A and E are not in the same class because, while it is possible to travel from A to E (through B and C), it is not possible to travel from E to A.

Definition 2.9

A chain is **irreducible** if it only has one class, i.e., if all states communicate.

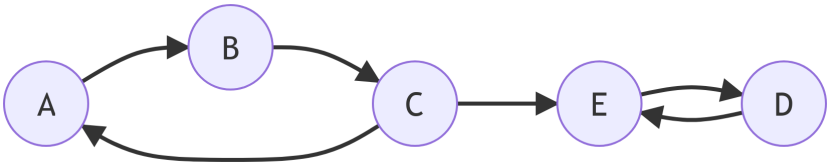


Figure 2.6: *Illustration of communicating classes.*

We will often focus on irreducible Markov chains later, when we study their long-term behaviour. It is a little more complicated to think about it for reducible processes, where we have to figure out which class the process will get stuck in (e.g., {E, D} in Figure 2.6), how long we can expect it will take to get stuck there, and so on. These are also problems that have been studied, but we will not cover them in this course.

A state $i \in \mathcal{S}$ is called an **absorbing state** if $P_{ii} = 1$, i.e., if the process that has reached i can never leave. If a Markov chain has an absorbing state, then it is not irreducible, because it is not possible to travel from the absorbing state to any other state. For example, we can represent the board game Snakes and Ladders as a Markov chain, in which the final square is an absorbing state (because the player stays there once they've reached it).

2.3.2 Transience and recurrence

We may want to know whether, starting in a given state, the process will ever return to it. We define the first passage time (or first hitting time) in state i as the first positive time at which the state is visited by the chain; that is, we define $\tau_i = \min\{\, n > 0 : X_n = i \,\}$, and set $\tau_i = \infty$ if $X_n \neq i$ for all $n > 0$. Further, let f_i be the probability that, starting in state i, the chain will ever return to state i, i.e.,

$$f_i = \Pr(\tau_i < \infty \mid X_0 = i).$$

Definition 2.10

State i is called **recurrent** if $f_i = 1$, i.e., if the process will return to state i with probability 1.

The state is called **transient** if $f_i < 1$, i.e., if there is a non-zero probability that the process will never visit i again.

Example: In the Markov process with transition graph shown in Figure 2.7, A and B are transient states, and C, D and E are recurrent state. If the process starts in A or B, there is a non-zero probabilility that it will never return (if it transitions to C). If it starts in C, D or E, the process will revisit the state with probability 1.

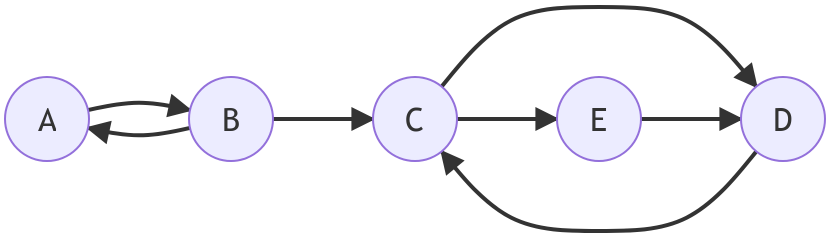


Figure 2.7: *Illustration of recurrent (C, D, E) and transient (A, B) states.*

We can show that a recurrent state i will be visited infinitely many times by the process, if it starts in i. Indeed, in that case, there is a probability 1 that the process will return to state i after some number of transitions, by definition of a recurrent state. Once it returns to i, the process is back to where it started, and again there is a probability 1 that the process will visit i a third time. We can repeat this reasoning to show that the process will infinitely return to any recurrent state i.

We can rewrite this statement in terms of transition probabilities. For any $n \geq 0$, define the indicator variable

$$I_n = \begin{cases} 1 & \text{if } X_n = i, \\ 0 & \text{otherwise,} \end{cases}$$

such that the total number of time steps spent in state i is $\sum_{n=0}^{\infty} I_n$. Then,

$$\begin{aligned} E\left[\sum_{n=0}^{\infty} I_n \,\middle|\, X_0 = i\right] &= \sum_{n=0}^{\infty} E\left[I_n \mid X_0 = i\right] \\ &= \sum_{n=0}^{\infty} \Pr(X_n = i \mid X_0 = i) \\ &= \sum_{n=0}^{\infty} P_{ii}^{(n)} \end{aligned}$$

This leads to an alternative definition for recurrence and transience.

Proposition 2.4

State i is recurrent if and only if $\sum_{n=0}^{\infty} P_{ii}^{(n)} = \infty$.

State i is transient if and only if $\sum_{n=0}^{\infty} P_{ii}^{(n)} < \infty$.

Note that, when the state space is finite, transient states only exist for reducible Markov chains, i.e., when there is no possible path from one state to another. However, when the state space is (countably) infinite, it is possible to have a transient state in an irreducible chain.

Example 2.4: transience with infinite state space

In an irreducible transient Markov chain, there is a positive probability of travelling from any state to any state, but there is also a positive probability of never visiting any given state again. One example is the Markov chain over $\mathcal{S} = \mathbb{N}$, where, for each state i,

$$\begin{cases} \Pr(X_{n+1} = i + 1 \mid X_n = i) = 0.8 \\ \Pr(X_{n+1} = i - 1 \mid X_n = i) = 0.2 \end{cases}$$

That is, at each time step, the process has a 80% probability of moving one unit to the right, and a 20% probability of moving one unit to the left. (It cannot move to the left when $X_n = 0$, so instead there is a 20% probability of not moving in this case.)

All states communicate, because we can in principle go between any two states in $\mathcal{S}$, so this is an irreducible Markov chain. However, because the probability of moving to the right is larger than 0.5, the chain will move towards larger and larger states, and there is a positive probability that it will never visit 0 again, for example. A realisation from this chain over 100 time steps is shown in Figure 2.8.

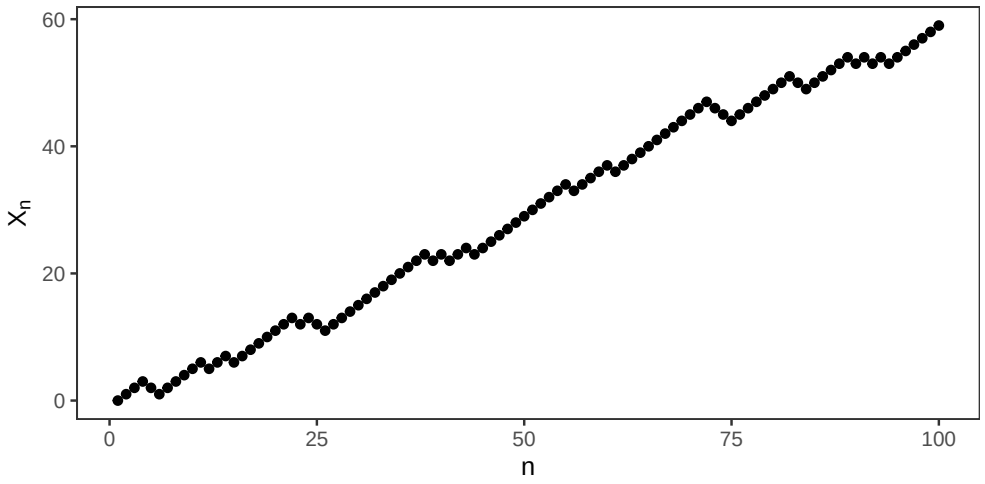


Figure 2.8: *Simulated realisation from example transient Markov chain.*

We end this section with a convenient property, which will allow us to talk about the transience and recurrence of communication classes, rather than individual states.

Proposition 2.5

Transience and recurrence are class properties:

1. if i is transient and $i \leftrightarrow j$, then j is transient;
2. if i is recurrent and $i \leftrightarrow j$, then j is recurrent.

2.3.3 Different types of recurrence

We denote as m_j the expected number of time steps it will take a chain that started in state j to return to state j, i.e.,

$$m_j = E[\tau_j \mid X_0 = j],$$

where τ_j is the first passage time defined in the previous section.

Definition 2.11

We say that a recurrent state j is **positive recurrent** if $m_j < \infty$, and **null recurrent** if $m_j = \infty$.

The distinction between positive and null recurrence only exists for Markov chains with infinite state spaces. If the state space is finite, then all recurrent states are positive recurrent. The notion of null recurrent state might be counter-intuitive: if there is positive probability that we revisit j, then how can the expected time until this happens be infinite? (Note that there are many other such examples where expectations defy our intuition, like the St Petersburg paradox). To understand, we can look at the definitions. We call state j recurrent if $f_j = 1$ where

$$\begin{aligned} f_j &= \Pr(\tau_j < \infty \mid X_0 = j) \\ &= \sum_{n=0}^{\infty} \Pr(\tau_j = n \mid X_0 = j) \end{aligned}$$

On the other hand, the expected return time m_j is

$$\begin{aligned} m_j &= E[\tau_j \mid X_0 = j] \\ &= \sum_{n=0}^{\infty} n \Pr(\tau_j = n \mid X_0 = j) \end{aligned}$$

Then, it's not too difficult to think of a situation where $f_j = 1$ but the expected return time is infinite. For example, if $\Pr(\tau_j = n \mid X_0 = j) = 6/(\pi n)^2$ for $n \geq 1$, then $f_j = 6/\pi^2 \sum 1/n^2$ converges to 1 (see "Basel problem"), while $m_j = 6/\pi^2 \sum 1/n$ diverges to ∞ (see "harmonic series"). More generally, whether or not $f_j = 1$ is not a good indicator of whether m_j is finite.

Just like for transience and recurrence, positive and null recurrence are class properties, so we can use those terms to refer to a class rather than just a state.

Proposition 2.6

Let $i, j \in \mathcal{S}$.

1. If i is a positive recurrent state and $i \leftrightarrow j$, then j is positive recurrent.
2. If i is a null recurrent state and $i \leftrightarrow j$, then j is null recurrent.

When all states communicate, i.e., when the chain is irreducible, we can go one step further

and use the terms to refer to the chain.

Definition 2.12

An irreducible Markov chain is called transient if at least one state is transient; it is called positive recurrent if at least one state is positive recurrent; and it is called null recurrent if at least one state is null recurrent.

In the definition above, saying that at least one state is transient or recurrent is equivalent to saying that every state is, because they are class properties.

Example 2.5: null recurrence

The textbook example of a null recurrent Markov chain is the random walk over $\mathcal{S} = \mathbb{Z}$; for any $i \in \mathbb{Z}$, the transition probabilities are

$$\begin{cases} \Pr(X_{n+1} = i+1 \mid X_n = i) = 0.5 \\ \Pr(X_{n+1} = i-1 \mid X_n = i) = 0.5 \end{cases}$$

The chain is recurrent because the probability of returning to any state is 1, but it is null recurrent because the expected return time is infinite. A simulated realisation from this process is shown in Figure 2.9.

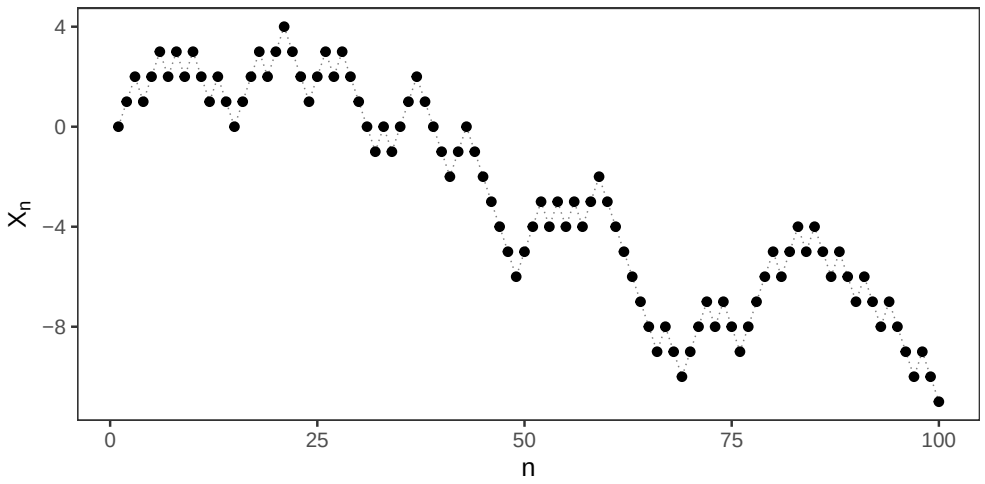


Figure 2.9: *Simulated realisation from example null recurrent Markov chain.*

Example 2.6: positive recurrence

Any irreducible Markov chain with finite state space is positive recurrent. For an example of a positive recurrent chain with infinite state space, consider the chain over $\mathcal{S} = \mathbb{N}$ with transition probabilities

$$\begin{cases} \Pr(X_{n+1} = i+1 \mid X_n = i) = 0.2 \\ \Pr(X_{n+1} = i-1 \mid X_n = i) = 0.8 \end{cases}$$

for any state $i > 0$, and, if $i = 0$, there is a 0.8 probability of remaining at 0, and a 0.2 probability of switching to 1.

Because zero is a dead end in this chain, and because it is more likely to go down than up, it will not suffer from the same divergence issues that we saw in the transient and null recurrent examples. This is a positive recurrent chain, and Figure 2.10 shows an example realisation.

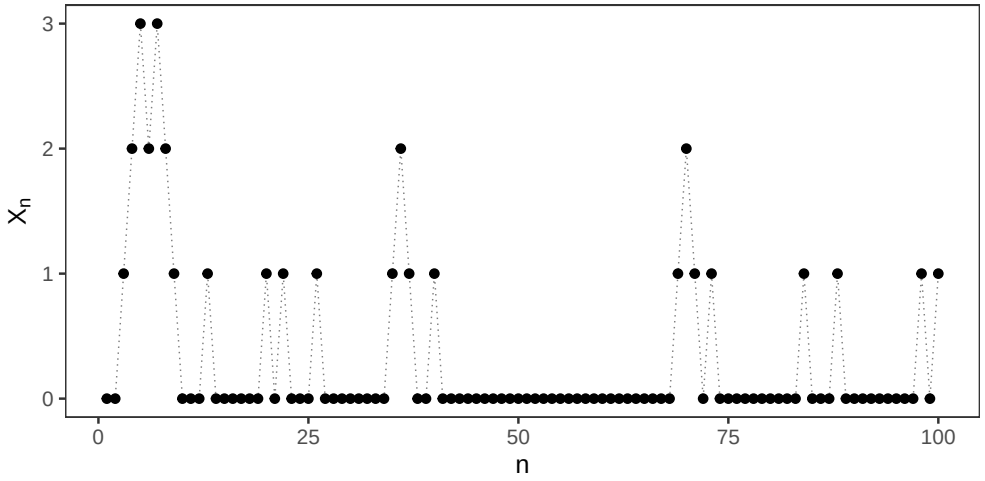


Figure 2.10: *Simulated realisation from example positive recurrent Markov chain.*

2.3.4 Periodicity

Definition 2.13

A state i has **period** d if the number of steps it takes the chain to return to i can only be a multiple of d. In other words, $d = \gcd\{n \in \mathbb{N}_{>0} : P_{ii}^{(n)} > 0\}$.

We say that i is **periodic** if $d(i) > 1$, and **aperiodic** if $d(i) = 1$.

It turns out that periodicity is also a class property, and so we also use the term to refer to a communicating class, or to an irreducible Markov chain.

Figure 2.11 shows an example periodic Markov chain. The only way to go from A to A is to go through B, C and D exactly once, so the period is 4. Figure 2.12 is the transition graph of a very similar Markov chain, but it has been modified by adding a non-zero probability of remaining in state B. Then, the period becomes 1 because the chain can take any number of steps to return to a state, i.e., the chain is aperiodic.

Another example is the difference between the king and knight in the chess example from Section 2.2.2. The king can take 1, 2, 3, or any number of steps to return to its initial position, whereas the knight can only return to a position after 2, 4, 6, or any even number of steps. The

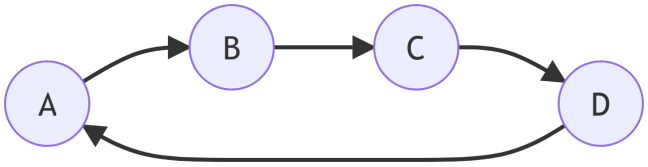


Figure 2.11: *Periodic Markov chain with period 4.*

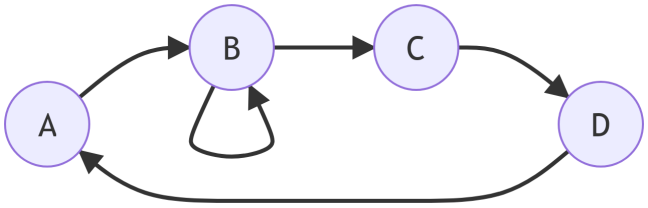


Figure 2.12: *Aperiodic Markov chain.*

king's process is aperiodic, whereas the knight's is periodic (with period 2).

2.4 Long-run properties

We are often interested in the long-run properties of the Markov chain, such as stability (does the system always converge to some distribution?) and long-run proportions (how much time does the process spend in each state on average?). We will link these questions to the concept of stationary distribution, and show how they can be answered in practice. In this section, we focus on irreducible aperiodic Markov chains. In this situation, the states must be either all positive recurrent, all null recurrent, or all transient.

2.4.1 Stationary distribution

Definition 2.15

Let (X_n) be a Markov chain with transition probability matrix P on $\mathcal{S}$, and consider the probability distribution π on $\mathcal{S}$. We say that π is a **stationary distribution** of (X_n) if

$$\pi P = \pi.$$

Equivalently, $\pi = (\pi_0, \pi_1, \dots)$ is a stationary distribution if it satisfies

$$\sum_{i \in \mathcal{S}} \pi_i P_{ij} = \pi_j, \quad \text{for any } j \in \mathcal{S}.$$

We use the term "stationary" because it represents an equilibrium. Indeed, if the initial distribution $u^{(0)}$ is a stationary distribution of the chain, the distributions at times $n = 1, 2, ...$ are

$$\begin{aligned} u^{(1)} &= u^{(0)}P = u^{(0)}, \\ u^{(2)} &= u^{(1)}P = u^{(0)}P = u^{(0)}, \\ u^{(3)} &= u^{(2)}P = u^{(0)}P = u^{(0)}, \end{aligned}$$

and so on. That is, if a Markov chain starts in its stationary distribution, then it will remain in the stationary distribution. The stationary distribution is also sometimes called the **equilibrium** or **invariant** distribution of the process.

Example 2.7: Calculating the stationary distribution "by hand"

Consider the two-state Markov chain with transition probability matrix

$$P = \begin{pmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{pmatrix}$$

where $0 < p_1, p_2 < 1$. Find its stationary distribution(s).

We are looking for a vector $\pi = (\pi_1, \pi_2)$ that satisfies

$$\begin{aligned} & \pi P = \pi \\ \Leftrightarrow \quad & \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \begin{pmatrix} 1-p_1 & p_1 \\ p_2 & 1-p_2 \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \end{pmatrix} \\ \Leftrightarrow \quad & \begin{cases} \pi_1(1-p_1) + \pi_2 p_2 = \pi_1 \\ \pi_1 p_1 + \pi_2(1-p_2) = \pi_2 \end{cases} \\ \Leftrightarrow \quad & \begin{cases} p_1\pi_1 = p_2\pi_2 \\ p_1\pi_1 = p_2\pi_2 \end{cases} \end{aligned}$$

The two equations are redundant, so we only keep one of them, and use the constraint

$\pi_1 + \pi_2 = 1$ to find the solution:

$$\begin{aligned}
&\begin{cases} p_1\pi_1 = p_2\pi_2 \\ \pi_1 + \pi_2 = 1 \end{cases} \\
\Leftrightarrow \quad &\begin{cases} p_1\pi_1 = p_2\pi_2 \\ \pi_2 = 1 - \pi_1 \end{cases} \\
\Leftrightarrow \quad &\begin{cases} p_1\pi_1 = (1 - \pi_1)p_2 \\ \pi_2 = 1 - \pi_1 \end{cases} \\
\Leftrightarrow \quad &\begin{cases} \pi_1(p_1 + p_2) = p_2 \\ \pi_2 = 1 - \pi_1 \end{cases} \\
\Leftrightarrow \quad &\pi = \begin{pmatrix} \frac{p_2}{p_1+p_2} & \frac{p_1}{p_1+p_2} \end{pmatrix}
\end{aligned}$$

So the Markov chain has a unique stationary distribution π given above.

In general, the derivation follows the same steps:

1. write the system of equations $\pi P = \pi$;
2. drop one of the equations;
3. use the row sum constraint to solve the system.

We will see that there are more efficient methods to derive the stationary distribution, which are particularly useful for models with a large number of states.

What is the connection between the long-term behaviour of the process and its stationary distribution? The following proposition gives us a link with the limiting distribution of the process, when it exists.

Proposition 2.7

Assume that $u_i^{(n)}$ has a limit for $n \to \infty$ for all $i \in \mathcal{S}$, and denote it as

$$\pi_i = \lim_{n\to\infty} u_i^{(n)}.$$

Then $\pi = (\pi_1, \pi_2, \dots)$ is a stationary distribution of the chain.

Proof

We will prove this proposition only in the case where $\mathcal{S}$ is finite, but it also applies to Markov chains with infinite state spaces.

We first show that π is a probability distribution over $\mathcal{S}$. We have

$$\begin{aligned}\sum_{j\in\mathcal{S}} \pi_j &= \sum_{j\in\mathcal{S}} \lim_{n\to\infty} u_j^{(n)} \\ &= \lim_{n\to\infty} \sum_{j\in\mathcal{S}} u_j^{(n)} \quad \text{(a)} \\ &= 1, \quad \text{(b)}\end{aligned}$$

where we can swap the limit and sum in (a) because $\mathcal{S}$ is finite, and (b) holds because $u^{(n)}$ is a probability distribution.

Moreover,

$$\begin{aligned}\pi_j &= \lim_{n\to\infty} \Pr(X_n = j) \\ &= \lim_{n\to\infty} \Pr(X_{n+1} = j) \\ &= \lim_{n\to\infty} \sum_{i\in\mathcal{S}} \Pr(X_n = i) P_{ij} \quad \text{(a)} \\ &= \sum_{i\in\mathcal{S}} \lim_{n\to\infty} \Pr(X_n = i) P_{ij} \quad \text{(b)} \\ &= \sum_{i\in\mathcal{S}} \pi_i P_{ij},\end{aligned}$$

where (a) is the law of total probability, and (b) follows from the finiteness of $\mathcal{S}$ as before. We can then conclude that π is a stationary distribution of the chain.

This result is not particularly useful in practice, because it assumes that we know what the limiting probabilities are, but it is a good starting point, and we will later see that we can go the other direction (from the stationary distribution, which we know how to compute, to the limiting probabilities).

The next theorem outlines the conditions under which an irreducible Markov chain has a stationary distribution, and the connection between the stationary distribution and the expected return time. Recall that the expected return time to state j is defined as $m_j = E[\tau_j \mid X_0 = j]$, where $\tau_j = \min\{n > 0 : X_n = j\}$ is the first passage time to state j.

Theorem 2.1

An irreducible Markov chain has a stationary distribution π if and only if it is positive recurrent. In this case, there is a unique stationary distribution, with elements

$$\pi_i = \frac{1}{m_i}, \quad \text{for } i \in \mathcal{S}.$$

We can interpret the connection to the return time as follows: the longer it takes to return to state i, the less time the chain will spend in state i overall. If the process visits state i every

m_i time steps, then it makes sense that the proportion of time spent at i is $1/m_i$.

2.4.2 Limiting probabilities

Theorem 2.2

Let (X_n) be an irreducible aperiodic Markov chain with n-step transition probabilities $P_{ij}^{(n)}$ for $i, j \in \mathcal{S}$. Then, we have

$$\lim_{n\to\infty} P_{ij}^{(n)} = \frac{1}{m_j}, \quad \text{for } i, j \in \mathcal{S}.$$

The limit does not depend on the starting state i so, equivalently, we can write

$$\lim_{n\to\infty} u_j^{(n)} = \frac{1}{m_j}, \quad \text{for } j \in \mathcal{S},$$

where $u^{(n)}$ is the probability distribution of X_n.

There are several important implications of this theorem:

1. If the chain is positive recurrent, then we saw in the previous section that $\pi_j = 1/m_j$ defines the unique stationary distribution of the process, so $\lim_{n\to\infty} u^{(n)} = \pi$. In this case, the limiting distribution of X_n and the stationary distribution coincide. This is very useful, because we usually know how to derive the stationary distribution of the process from the transition probability matrix.

2. If the chain is transient or null recurrent, then $m_j = \infty$, and so $\lim_{n\to\infty} u_j^{(n)} = 0$ for all $j \in \mathcal{S}$. The probability of being in any given state decreases to zero as time goes by.

In the previous theorem, we restricted our attention to aperiodic Markov chain because, in the case of periodic chains, the limit $\lim_{n\to\infty} \Pr(X_n = i)$ does not always exist. The following is a classic example of this situation.

Example 2.8: Periodic Markov chain

Convergence to equilibrium is only guaranteed when the chain is aperiodic. Consider the chain (X_n) with transition probability matrix

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

That is, the evolution of the chain is deterministic: it switches state at each time step.

(X_n) has $\pi = (0.5, 0.5)$ as a stationary distribution, because

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

Intuitively, if there is a 50% probability of being in each state at time n, there is still a 50% probability of being in each state at $n+1$.

However, we can show that there is no limiting distribution for this Markov chain, because

$$P^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P^3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad P^4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

and so on. The state distribution $u^{(n)}$ alternates between $(0, 1)$ and $(1, 0)$, and so the limit $\lim_{n\to\infty} u_i^{(n)}$ does not exist.

Another example is the difference between the random walk of a king and the knight on a chess board, as described in Section 2.2.2. We saw that the knight's distribution over the board does not converge as time goes by, because it has a different limit for even and odd n (leading to the alternating pattern in Figure 2.5). This is because its position follows a periodic chain with period 2. On the other hand, the king's chain is aperiodic, and it does have a limiting distribution.

2.4.3 Long-run proportions

Definition 2.16

Consider a Markov chain with state space $\mathcal{S}$. For any $i \in \mathcal{S}$, the i^{th} **long-run proportion** is the proportion of time that the Markov chain spends in the i^{th} state over the long run, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n} \mathbb{I}_{\{X_k = i\}},$$

where $\mathbb{I}$ is the indicator function, i.e.,

$$\mathbb{I}_{\{X_k = i\}} = \begin{cases} 1 & \text{if } X_k = i \\ 0 & \text{otherwise.} \end{cases}$$

Note that the question of finding the long-run proportions of a chain, i.e., to know how often each state will be active if we let it run for a long time, is separate from the question addressed in the previous section. Indeed, even in cases where the chain does not have a limiting distribution, the long-run proportions might exist. In the example of a periodic Markov chain where $p_{12} =$

$p_{21} = 1$, we saw that there is no limiting distribution, but we can compute the proportion of time spent in each state. Because the chain switches at every time step, the proportion of time in each state will converge to 50%. Likewise, the knight moving randomly around a chess board does have long-term proportions, even though it doesn't have a limiting distribution.

Theorem 2.3

If (X_n) is an irreducible Markov chain, then the long-run proportions are given by

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}_{\{X_k=i\}} = \frac{1}{m_i}, \quad \text{for } i \in \mathcal{S}.$$

This time, the result does not require aperiodicity. Based on results from the previous sections, this implies that:

1. For a positive recurrent chain, the long-run proportions coincide with the stationary distribution.
2. For a transient or null recurrent chain, the long-run proportions are zero.

We will not prove this result, but we can provide some intuition. If π'_i is the long-run proportion of time spent in state i, then the long-run proportion of transitions that go from i to j is $\pi'_i P_{ij}$. If we sum over all i, this becomes

$$\pi'_j = \sum_{i\in\mathcal{S}} \pi'_i P_{ij}$$

or, in matrix notation, $\pi' P = \pi'$. That is, π' is a stationary distribution of the Markov chain.

2.4.4 Calculating the stationary distribution

To compute the limiting distribution of a Markov chain or its long-run proportions, the most convenient approach is usually to find its stationary distribution. We describe two approaches to compute the stationary distribution of a Markov chain with finite state space (when it exists), one based on the eigendecomposition of the transition probability matrix, and one that only requires inverting a well-chosen matrix.

Method 1: eigendecomposition

Recall that an eigenvector (or right eigenvector) of the square matrix A is a non-zero vector x such that $Ax = \lambda x$ for some scalar λ; λ is called the eigenvalue associated with x. This looks a little similar to the definition of a stationary distribution, except the stationary distribution is a *left* eigenvector of the transition probability matrix (i.e., we left-multiply the transition matrix). Most eigenanalysis theory and software focuses on right eigenvector but, fortunately, there is a close connection between right and left eigenvectors. Indeed, x is a right eigenvector of A if and only if $x^\top$ is a left eigenvector of $A^\top$. You can see this using the general identity $(Ax)^\top = x^\top A^\top$.

Finding the stationary distribution of a Markov chain is therefore equivalent to finding the eigenvector of $P^\top$ associated with the eigenvalue 1. Note that eigenvectors are defined up to a multiplicative constant, because any multiple of an eigenvector is also an eigenvector. So, once the eigenvector is found, we divide each element by the sum of all elements, so find the vector that defines a valid probability distribution over $\mathcal{S}$.

Method 2: matrix inverse

Proposition 2.8

Let (X_n) be a Markov chain with transition probability matrix P and with $|\mathcal{S}| = N$ states. The probability distribution π is a stationary distribution of (X_n) if and only if $\pi(I - P + U) = 1$, where

- I is a $N \times N$ identity matrix;
- U is a $N \times N$ matrix of ones;
- 1 is a row vector of N ones.

Proof

We first show that $\pi U = 1$:

$$\begin{aligned}\pi U &= \begin{pmatrix}\pi_1 & \pi_2 & \cdots & \pi_N\end{pmatrix}\begin{pmatrix}1 & 1 & \cdots & 1\\ 1 & 1 & \cdots & 1\\ \vdots & \vdots & \ddots & \vdots\\ 1 & 1 & \cdots & 1\end{pmatrix}\\ &= \begin{pmatrix}\sum_k \pi_k & \sum_k \pi_k & \cdots & \sum_k \pi_k\end{pmatrix}\\ &= 1\end{aligned}$$

where $\sum_k \pi_k = 1$ because π is a probability distribution.

We can then use this result to prove the proposition:

$$\begin{aligned}&\pi(I - P + U) = 1\\ \Leftrightarrow\ &\pi I - \pi P + \pi U = 1\\ \Leftrightarrow\ &\pi - \pi P + 1 = 1\\ \Leftrightarrow\ &\pi - \pi P = 0\\ \Leftrightarrow\ &\pi = \pi P,\end{aligned}$$

i.e., π is a stationary distribution of (X_n).

When the stationary distribution exists, then $I - P + U$ is invertible, and the stationary

distribution can be calculated as

$$\pi = 1(I - P + U)^{-1}.$$

In practice, the matrix to invert is straightforward to compute, and there are many methods to invert a matrix, so this is convenient to implement a general approach to find the stationary distribution of a chain with finite state space.

Example 2.9

We illustrate the implementation of the two proposed methods to compute the stationary distribution in R. This example uses a 3 by 3 transition probability matrix, but the code would be virtually identical for a general N by N matrix.

```
# Transition probability matrix
P <- matrix(c(0.9, 0.05, 0.05,
              0, 0.7, 0.3,
              0.1, 0.1, 0.8),
            nrow = 3, byrow = TRUE)

###################################
## Method 1: eigendecomposition ##
###################################
# Get eigenvectors and eigenvalues
eig <- eigen(t(P))
# Keep vector associated with eigenvalue 1
eig_vec <- eig$vectors[,which(abs(eig$values - 1) < 1e-15)]
# Normalise vector to get a valid probability distribution
eig_vec / sum(eig_vec)
```

```
[1] 0.4 0.2 0.4
```

```
###################################
## Method 2: matrix inversion ##
###################################
# Define identity matrix and matrix of 1s
I <- diag(3)
U <- matrix(1, 3, 3)
# "solve" computes the matrix inverse, and "colSums" sums over columns
colSums(solve(I - P + U))
```

```
[1] 0.4 0.2 0.4
```

The two methods return the same solution.

2.5 Statistical Inference

So far, we have presented Markov chains as mathematical models, but we can also view them as data analysis tools. Given an observed time series of states, we might want to estimate the transition probabilities of the process, which can be used to determine its long-term properties.

2.5.1 Likelihood function

Let (X_n) be a Markov chain with transition probability matrix P. For an observed sequence $x_0, x_1, \dots, x_n$, the likelihood function for P is given by the joint probability $\Pr(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n)$. We can write

$$\begin{aligned}\Pr(X_0 = x_0, \dots, X_n = x_n) &= \Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &\quad \times \Pr(X_1 = x_1, \dots, X_{n-1} = x_{n-1}) \\ &= \Pr(X_n = x_n \mid X_{n-1} = x_{n-1}) \\ &\quad \times \Pr(X_0 = x_0, \dots, X_{n-1} = x_{n-1})\end{aligned}$$

where the first equality uses the definition of conditional probability, and the second equality comes from the Markov property. Using the same reasoning, we also have

$$\begin{aligned}\Pr(X_0 = x_0, \dots, X_{n-1} = x_{n-1}) &= \Pr(X_{n-1} = x_{n-1} \mid X_{n-2} = x_{n-2}, \dots, X_0 = x_0) \\ &\quad \times \Pr(X_0 = x_0, \dots, X_{n-2} = x_{n-2}) \\ &= \Pr(X_{n-1} = x_{n-1} \mid X_{n-2} = x_{n-2}) \\ &\quad \times \Pr(X_0 = x_0, \dots, X_{n-2} = x_{n-2})\end{aligned}$$

which we can combine with the previous equation to find

$$\begin{aligned}\Pr(X_0 = x_0, \dots, X_n = x_n) &= \Pr(X_n = x_n \mid X_{n-1} = x_{n-1}) \\ &\quad \times \Pr(X_{n-1} = x_{n-1} \mid X_{n-2} = x_{n-2}) \\ &\quad \times \Pr(X_0 = x_0, \dots, X_{n-2} = x_{n-2})\end{aligned}$$

We repeat this reasoning to find the joint probability as a product of transition probabilities

$$\Pr(X_0 = x_0, \dots, X_n = x_n) = \Pr(X_0 = x_0) \prod_{k=1}^{n} \Pr(X_k = x_k \mid X_{k-1} = x_{k-1})$$

The first term, $\Pr(X_0 = x_0)$, can either be treated as a parameter, or the first value can be viewed as deterministic, so $\Pr(X_0 = x_0) = 1$. Here, we follow the latter approach, which does not affect inference on the transition probabilities.

Finally, the likelihood function is the joint probability of the observed data, written as a function

of the parameter P:

$$L(P) = \prod_{k=1}^{n} \Pr(X_k = x_k \mid X_{k-1} = x_{k-1}) = \prod_{k=1}^{n} P_{x_{k-1},x_k}$$

It is convenient to notice that this product contains each transition probability P_{ij} as many times as there are transitions from state i to state j in the observed sequence. For $i, j \in \mathcal{S}$, let n_{ij} be the number of transitions from i to j. Then,

$$L(P) = \prod_{i \in \mathcal{S}} \prod_{j \in \mathcal{S}} P_{ij}^{n_{ij}}.$$

We often compute the log-likelihood, rather than the likelihood itself, because it is often easier to work with analytically and numerically (e.g., see "Parameter estimation" below). Taking the log, we find

$$\ell(P) = \log[L(P)] = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} n_{ij} \log(P_{ij})$$

2.5.2 Parameter estimation

The maximum likelihood estimator (MLE) of the transition probabilities in a Markov chain is obtained by counting the number of transitions between each pair of states in the observed data, and normalizing these counts to obtain the estimated probabilities.

Proposition 2.9

Let n_{ij} be the number of transitions from state i to state j in the observed data, and let $n_i = \sum_j n_{ij}$ be the total number of transitions from state i. Then the MLE of the transition probability P_{ij} from state i to state j is given by

$$\widehat{P}_{ij} = \frac{n_{ij}}{n_i}$$

To prove that this estimator is indeed the maximum likelihood estimator, we need to show that it maximises the likelihood function of the observed data.

The MLE is obtained by maximizing $L(P)$ (or, equivalently, $\ell(P)$) with respect to P, subject to the constraints that it is a stochastic matrix (i.e., entries are between 0 and 1, and rows sum to 1). This can be done using the method of Lagrange multipliers, which is a general approach to constrained optimisation problems. by noticing that each row of the transition probability matrix can be estimated separately. Specifically, we need to maximize the Lagrangian function

$\mathcal{L}$ with respect to the transition probability of interest, where

$$\mathcal{L}(P, \lambda) = \ell(P) - \sum_{m \in \mathcal{S}} \left\{ \lambda_m \left(\sum_{j \in \mathcal{S}} P_{mj} - 1 \right) \right\}$$

Taking the derivative of $\mathcal{L}$ with respect to P_{ij}, we obtain

$$\frac{\partial \mathcal{L}}{\partial P_{ij}} = \frac{n_{ij}}{P_{ij}} - \lambda.$$

To find the maximum likelihood estimator $\widehat{P}_{ij}$, we set the derivative to zero:

$$\begin{aligned} & \frac{n_{ij}}{\widehat{P}_{ij}} - \lambda = 0 \\ \Rightarrow \quad & \widehat{P}_{ij} = \frac{n_{ij}}{\lambda} \end{aligned}$$

We then use the row constraints on the transition probability matrix, and we find

$$\begin{aligned} & \sum_{j=1}^{N} \widehat{P}_{ij} = 1 \\ \Rightarrow \quad & \sum_{j=1}^{N} \frac{n_{ij}}{\lambda} = 1 \\ \Rightarrow \quad & \lambda = \sum_{j=1}^{N} n_{ij} \end{aligned}$$

Finally, putting everything together,

$$\widehat{P}_{ij} = \frac{n_{ij}}{\sum_{k \in \mathcal{S}} n_{ik}} = \frac{n_{ij}}{n_i}$$

In practice, given a sequence of observed states, we can then find the "best fitting" Markov chain by calculating those transition probabilities from the numbers of transitions in the data. Based on the estimates, we can then simulate from the process, or compute the stationary distribution to better understand its long-term emerging features.

2.6 Markov chains with uncountable state space

We now briefly turn to the case where the state space is uncountable, e.g., $\mathcal{S} = \mathbb{R}$ or $\mathcal{S} = [0, 1]$. Markov chains with uncountable state space have had great practical utility, as illustrated in the later section on Markov chain Monte Carlo methods.

A stochastic process (X_n) defined over an uncountable state space $\mathcal{S}$ is called a **Markov process** if

$$f_{X_{n+1}|X_n,\dots,X_0}(x_{n+1} \mid x_n, \dots, x_0) = f_{X_{n+1}|X_n}(x_{n+1} \mid x_n)$$

The idea is the same as for discrete state spaces: conditionally on the current state, future states are independent of past states. The only difference is that, now, the state takes on continuous rather than discrete values.

Example 2.10: Gaussian random walk

Consider the process (X_n) defined by $X_0 = 0$ and

$$X_{n+1} \mid X_n = x_n \sim N(x_n, 1),$$

for $n = 1, 2, \dots$. This process satisfies the Markov property, because the distribution of X_{n+1} can be written in terms of only X_n. Five example realisations of this Gaussian random walk are shown in Figure 2.13.

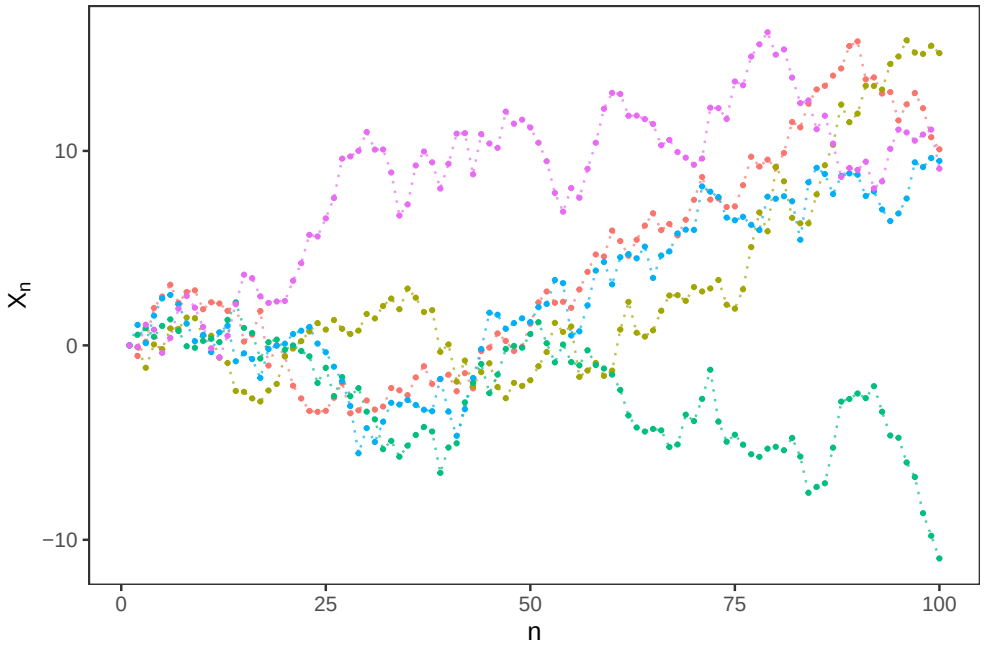


Figure 2.13: *Five simulated realisations of Gaussian random walk starting at* $X_0 = 0$.

In the uncountable case, the transition dynamics of the process cannot be written as a matrix (not even an infinite matrix). Instead, we define the transition kernel of the process, $K : \mathcal{S} \times \mathcal{S} \to [0, \infty)$, as

$$K(x, y) = f_{X_{n+1}|X_n}(y \mid x), \quad \text{for } x, y \in \mathcal{S}.$$

That is, $K(x, y)$ must be a probability density function with respect to y, such that $\int_{\mathcal{S}} K(x, y)\ dy = 1$. This function gives the probability density of each transition over continuous space. For example, in the Gaussian random walk described above, the transition kernel was given by the probability density function of the normal distribution.

Likewise, the initial distribution of the process must be written as a continuous probability distribution, with density function $f_{X_0} : \mathcal{S} \to [0, \infty)$. The marginal distribution of the process at subsequent time points can then be derived recursively as

$$f_{X_{n+1}}(x_{n+1}) = \int_{\mathcal{S}} f_{X_n}(x) K(x, x_{n+1}) \, dx$$

We say that the distribution $\pi : \mathcal{S} \to [0, \infty)$ is a stationary distribution of the Markov process (X_n) with transition kernel K if

$$\int_{\mathcal{S}} \pi(x) K(x, y) \, dx = \pi(y).$$

Because π is a probability distribution over $\mathcal{S}$, it is also subject to the constraint $\int_{\mathcal{S}} \pi(x) \, dx = 1$. The problem of deriving the stationary distribution of a Markov chain over an uncountable state space is the foundation of Markov chain Monte Carlo, a powerful method to sample from complex probability distributions (described in the Applications section below).

The different types of Markov chains that we studied in the countable case have to be refined to the uncountable case, with the notions of ϕ-irreducibility, Harris recurrence, and periodicity over a partition of the state space, but we will not describe these here.

2.7 Applications

2.7.1 Markov chain Monte Carlo

Consider the problem of Bayesian inference. If we denote as Z the random vector of data, and Θ the vector of parameters of interest, the focus of a Bayesian analysis is the posterior distribution

$$f_{\Theta|Z}(\theta \mid z) = \frac{f_{Z|\Theta}(z \mid \theta) f_{\Theta}(\theta)}{f_Z(z)}$$

where $f_{\Theta|Z}$ is called the posterior distribution, $f_{Z|\Theta}$ is the likelihood function, f_Θ is the prior distribution, and f_Z is the marginal distribution of the data. The likelihood function is determined by the choice of model formulation, and the prior distribution is chosen by the user based on prior knowledge about the parameters.

The marginal distribution in the denominator can be written as $f_Z(z) = \int f_{Z|\Theta}(z \mid \theta) f_\Theta(\theta) \, d\theta$, and it is not generally tractable except for simple model formulations. Because this term does not depend on the parameter θ, we can in principle evaluate the numerator over some grid of values of θ (plugging in the observed data for z), and approximate the constant on the denominator by summing them. However, this approach is not computationally feasible in cases where Θ is high-dimensional, which is very common in practice; many interesting statistical models have tens or hundreds of parameters to estimate.

An alternative approach to this problem is to *generate samples* from the posterior distribution $f_{\Theta|Z}$, rather than try to evaluate it directly. It turns out that a random sample from the distribution is all we need to approximate relevant summaries of the posterior distribution, e.g., the mean of the distribution (often used as point estimate), or its quantiles (often used to define interval estimates). Generating a random sample from a distribution is also a difficult task in general, though. Basic methods developed for this purpose include inverse transform sampling and rejection sampling, for example. These only work well in simple cases, e.g., when the cumulative distribution function can be evaluated, or for low-dimensional distributions.

Markov chain Monte Carlo (MCMC) is a general method which performs well in a wide range of situations, including for high-dimensional problems. The idea behind Markov chain Monte Carlo is to define a Markov chain over the parameter space, for which the stationary distribution is known to be the posterior distribution, i.e., $\pi(\theta) = f_{\Theta|Z}(\theta \mid z)$. Under some technical conditions analogous to those we described in the countable case in Section 2.4, the Markov chain is known to converge to the stationary distribution regardless of the starting point. A sample from the posterior distribution can then be obtained by simulating from the process (i.e., repeatedly sampling from its transition kernel).

2.7.2 Google

The PageRank algorithm is a method for ranking web pages based on their importance and relevance to a particular search query, which was originally used by Google. Although the original paper by Brin and Page (1998) did not make that connection, the algorithm uses a Markov chain to model the behaviour of a hypothetical web surfer who randomly clicks on links from one page to another.

The basic idea is to find the proportion of time spent on different web pages, assuming that the surfer will randomly click on links until they reach a dead end (i.e., a page with no outgoing links). To model this behavior as a Markov chain, we can represent each web page as a state in the chain, and the links between pages determine its transition probabilities. Then, assuming that the resulting process is irreductible, its stationary distribution can be computed to find the long-term proportion of time spent on each page.

This approach is superior to simply counting how many pages link to a given website, because this would be ignoring the fact that those pages don't all have the same weight. A link from a very popular page matters more, and this is captured by this random-surfer model.

As an example, we consider the web pages corresponding to the transition graph shown in Figure 2.14. Each node is a web page, and each arrow is a link. Note that there are two absorbing states, E and F, which means that the process is not irreducible. To solve this problem, the common solution is to assume that the surfer randomly opens a new page (from all existing pages) when they run out of links to click. (This means that we will work with the chain that has outgoing links from E and F to every other page.)

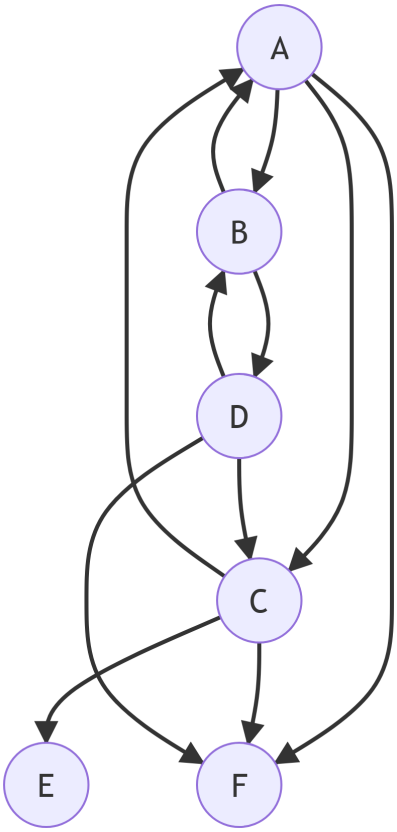


Figure 2.14: *Example transition graph for six webpages, where the arrows are links.*

The first step is to compute the transition probability matrix. The transition probability from i to j is zero if there is no link, and it is one over the number of outgoing links from i otherwise. Here, we start by defining an adjacency matrix, and then normalise the rows to get transition probabilities. Finally, we compute the stationary distribution using a result from Section 2.4.4.

```
# Define the adjacency matrix
N <- 6
A <-  matrix(c(0, 1, 1, 0, 0, 1,
               1, 0, 0, 1, 0, 0,
               1, 0, 0, 0, 1, 1,
               0, 1, 1, 0, 0, 1,
               0, 0, 0, 0, 0, 0,
               0, 0, 0, 0, 0, 0),
             nrow = N, ncol = N, byrow = TRUE,
             dimnames = list(LETTERS[1:N], LETTERS[1:N]))

# Remove sinks
A[which(rowSums(A) == 0),] <- 1

# Compute the transition probability matrix
P <- A / rowSums(A)
round(P, 3)
```

```
      A     B     C     D     E     F
A 0.000 0.333 0.333 0.000 0.000 0.333
B 0.500 0.000 0.000 0.500 0.000 0.000
C 0.333 0.000 0.000 0.000 0.333 0.333
D 0.000 0.333 0.333 0.000 0.000 0.333
E 0.167 0.167 0.167 0.167 0.167 0.167
F 0.167 0.167 0.167 0.167 0.167 0.167
```

```
# Compute the PageRank scores
v <- colSums(solve(diag(N) - P + matrix(1, N, N)))
round(v, 3)
```

```
    A     B     C     D     E     F
0.194 0.167 0.167 0.139 0.111 0.222
```

A person randomly clicking on the links on these six pages will therefore spend roughly 19% of their time on page A, 17% on page B, and so on. These proportions can be used as weights to rank the pages in terms of their overall importance.

2.7.3 N-gram models (predictive text)

An N-gram is a sequence of N consecutive words, and they can be used as states of a Markov chain to predict text input. Using a large corpus of text, we can count how many times each possible sequence of N words is followed by any other possible sequence of N words. (As you can imagine, this corresponds to a Markov chain with many states, especially for large N.) These give us transition probabilities of a Markov chain, and simulating from it produces new text. For small N, the result might be quite crude, but we can obtain more realistic text as N increases. Your phone could use this procedure to suggest the next word when you are typing a text, for example.

Here are some examples of text generated from a 2-gram, 3-gram, and 4-gram model, where the corpus used for estimation was the Wikipedia page about Canada. The initial state of the chain was set to "Canada" in the 2-gram model, "Canada is" in the 3-gram model, and "Canada is a" in the 4-gram model.

2-gram examples:

- *Canada (later Quebec) and French explorer Samuel de Champlain arrived at least 14,000 years prior to governance emphasizing multiculturalism, which considered by 1855. The country in Quebec, cultural identity and trading networks. Canada for families receiving social service programs. In addition to refer not without conflict, European Canadians' early interactions with the northeast, on the 55th parallel in the assimilation of 1791 divided the top 100 km2 (3,855,100 sq mi) of the national healthcare systems in international affairs, with France's overseas collectivity of adults having at Tadoussac along the commissioner of deaths in the OECD…*
- *Canada is 81.1 years. Beginning in the world, after the Canadas into French-speaking population declined by the Indigenous cultures had a variety of adults self-report having attained at the largest area subject to ensure reasonably uniform standards of past colonial injustices and the provincial and post-secondary. Education in the top 100 km2 (3,855,100 sq mi) of Commons and a lack of Canada's GDP for all federal government to have the United States in Canada for quality of the world's northernmost settlement, Canadian horse, the border westward along the oldest post-secondary institution in international trade networks…*

3-gram examples:

- *Canada is generally divided into primary education, followed by secondary education and post-secondary. Education in Canada was formed as a middle power for its role in assisting European coureur des bois and voyageurs in their explorations of the Truth and Reconciliation Commission of Canada in 2008. This includes recognition of past colonial injustices and settlement agreements and betterment of racial discrimination issues, such as the 1976 Summer Olympics, the 1988 Winter Olympics, and the relatively flat Canadian*

Prairies in the Canada Health Act of 1984 and is reflected in its folklore, literature, music, art, and music…

- *Canada is a French Canadian culture that is distinct from English Canadian culture. Canada's approach to governance emphasizing multiculturalism, which is applied by the Parliament of the Canadian Rockies, the Coast Mountains, and the Mount Edziza volcanic complex. Canada is experiencing an increase in healthcare expenditures due to a combination of the continent during the North American colonies through Confederation, Canada was Norse explorer Leif Erikson. In approximately 1000 AD, the Norse built a small short-lived encampment that was occupied sporadically for perhaps 20 years at L'Anse aux Meadows on the $1 coin, the Arms of Canada…*

4-gram examples:

- *Canada is a federation composed of 10 federated states, called provinces, and three federal territories. In turn, these may be grouped into four main regions: Western Canada, Central Canada, Atlantic Canada, and Northern Canada (Eastern Canada refers to Central Canada and Atlantic Canada together). Provinces and territories have responsibility for social programs such as healthcare, education, and welfare, as well as administration of justice (but not criminal law). Together, the provinces collect more revenue than the federal government, a rarity among other federations in the world…*
- *Canada is a country in North America. Its ten provinces and three territories extend from the Atlantic Ocean in the west, the country encompasses 9,984,670 km2 (3,855,100 sq mi) of territory. Canada also has vast maritime terrain, with the world's longest coastline of 243,042 kilometres (151,019 mi). In addition to sharing the world's largest area of fresh water lakes. Stretching from the Atlantic Ocean to the north, and to the Pacific Ocean and northward into the Arctic Ocean, making it the world's second-largest country by total area, with the world's longest coastline of 243,042 kilometres (151,019 mi)…*

3 Poisson processes

We now turn to a different type of stochastic process, called a counting process. A counting process is useful to model the number or timing of events of interest, e.g., births in a hospital, goals scored by a hockey team, or earthquakes in a region.

Definition 3.1

A stochastic process $(N_t)_{t\geq 0}$ is called a **counting process** if:

1. for any $t \geq 0$, $N_t \geq 0$;
2. for any $t \geq 0$, N_t is integer-valued;
3. for any $0 \leq s \leq t$, $N_s \leq N_t$.

We interpret N_t as the total number of events that have occurred by time t.

Remarks:

- In the definition of a counting process, the time index t is defined over a continuous rather than discrete space. We can make that explicit by writing $(N_t)_{t\geq 0}$, but sometimes we omit it and simply denote the process as (N_t). We use the letter t to denote continuous time, in contrast with n for the discrete time index in the previous chapter.
- Here, we use the term "events" in the common sense, rather than in the technical sense from probability theory.

Example: The graph in Figure 3.1 shows an example counting process.

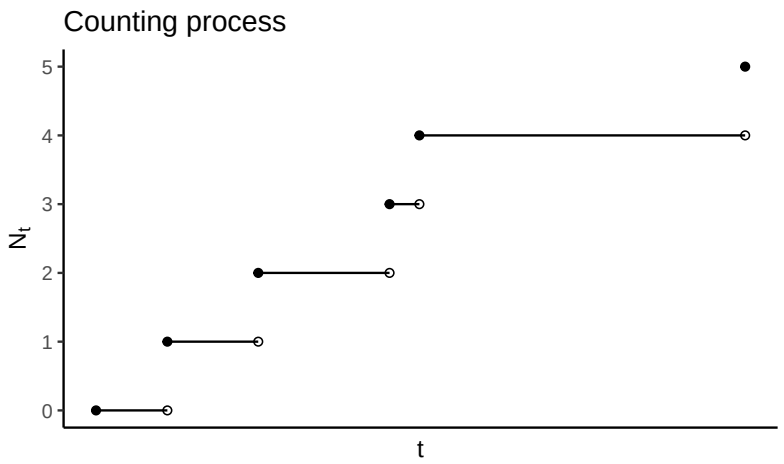


Figure 3.1: *Example of counting process.*

3.1 The Poisson process

3.1.1 Definition and terminology

Definition 3.2

We say that the stochastic process (N_t) has **independent increments** if, for any $0 \leq q < r \leq s < t$, the random variables $N_r - N_q$ and $N_t - N_s$ are independent.
We say that the stochastic process (N_t) has **stationary increments** if, for any $s, t > 0$, the random variables N_t and $N_{s+t} - N_s$ have the same distribution.

If the process (N_t) has independent and stationary increments, the random variable $N_t - N_s$ depends only on the length of the time interval $t - s$, and not on the specific values of s and t.

Definition 3.3

A counting process $(N_t)_{t \geq 0}$ is called a **Poisson process** with rate $\lambda > 0$ if:

1. $N_0 = 0$, i.e., no events have occurred yet at time 0;
2. (N_t) has independent and stationary increments;
3. $N_t \sim \text{Poisson}(\lambda t)$ for any $t > 0$, i.e.,

$$\Pr(N_t = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad \text{for } n \in \mathbb{N}.$$

There are at least three alternative definitions of the Poisson process, and we will discuss another one a little later.

From the definition of a Poisson process, we can see that the number of events over *any interval of length* t follows a Poisson distribution with rate λt. Indeed, because the increments of the process are stationary, N_t and $N_{s+t} - N_s$ have the same distribution for any $s, t > 0$, i.e.,

$$\Pr(N_{s+t} - N_s = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad \text{for } n \in \mathbb{N}.$$

As a reminder, Figure 3.2 shows the probability mass function of the Poisson distribution for five different values of the rate parameter λ.

The rate parameter λ of the process controls how often events happen: larger values of λ correspond to more frequent events. Remember that both the mean and variance of the Poisson distribution are equal to the rate parameter, so we know that, on average, λt events will take place in an interval of length t. In other words, the number of events is proportional to the rate parameter (λ) and to the length of the interval (t). It is intuitive that, on average, more events should take place over a longer interval; for a Poisson process, we expect twice as many events over an interval twice as long, and so on. The interpretation of λ is the average number

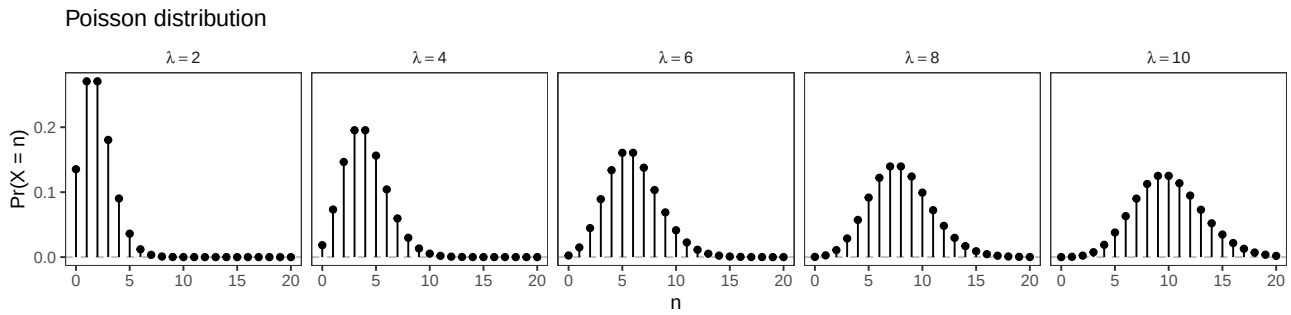


Figure 3.2: *Probability mass function of Poisson distribution for different values of the rate parameter λ.*

of events over an interval of length 1.

Example 3.1

Assume that the number of customers who arrive at Superstore follows a Poisson process with an average rate of $\lambda = 2$ customers per minute.

1. What is the probability that exactly 10 customers arrive in the next 3 minutes?
   Let X_1 be the number of customers that arrive in the next 3 minutes. By definition of the Poisson process, X_1 follows a Poisson distribution with rate $3\lambda = 6$. Then, we have
   $$\Pr(X_1 = 10) = e^{-6}\frac{6^10}{10!} = 0.041$$
2. What is the probability that at least 2 customers arrive in the next 30 seconds?
   The number of customers in the next 30 seconds, X_2, follows a Poisson distribution with rate $0.5\lambda = 1$, and so
   $$\begin{aligned}\Pr(X_2 \geq 2) &= 1 - \Pr(X_2 < 2)\\ &= 1 - \Pr(X_2 = 0) - \Pr(X_2 = 1)\\ &= 1 - e^{-1}\frac{1^0}{0!} - e^{-1}\frac{1^1}{1!}\\ &= 0.264\end{aligned}$$
3. What is the probability that the next customer will arrive within 15 seconds?
   This is equivalent to the probability the number of customers arriving in the next 15 seconds, X_3, is at least one. We know that X_3 follows a Poisson distribution with rate $0.25\lambda = 0.5$, so
   $$\begin{aligned}\Pr(X_3 \geq 1) &= 1 - \Pr(X_2 = 0)\\ &= 1 - e^{-0.5}\frac{0.5^0}{0!}\\ &= 0.393\end{aligned}$$
4. What is the probability that exactly 3 customers arrive in the first minute and exactly 10 customers arrive in the first three minutes?

Let X_4 and X_5 be the numbers of customers arriving in the first minute and in the first two minutes, respectively. These two random variables are not independent, but we can rephrase the question in terms of X_4 and $X_6 = X_5 - X_4$ (the number of customers arriving in the second and third minutes), which are independent. We want the probability that $X_4 = 3$ and $X_6 = 7$, and we know that $X_4 \sim \text{Poisson}(2)$ and $X_6 \sim \text{Poisson}(4)$, so

$$\begin{aligned}\Pr(X_4 = 3, X_6 = 7) &= \Pr(X_4 = 3) \times \Pr(X_6 = 7) \\ &= e^{-2}\frac{2^3}{3!} \times e^{-4}\frac{4^7}{7!} \\ &= 0.011\end{aligned}$$

In many contexts, the variable of interest is not the number of events N_t, but the times at which events occur, and the lengths of time intervals between events (e.g., expected time between two high-magnitude earthquakes in a region). For example, we might be interested in the times between events, denoted as $T_1, T_2, \ldots$ in Figure 3.3, or in the times of events, denoted as $S_1, S_2, \ldots$ in Figure 3.3. These random variables are called "interarrival times" and "arrival times", respectively.

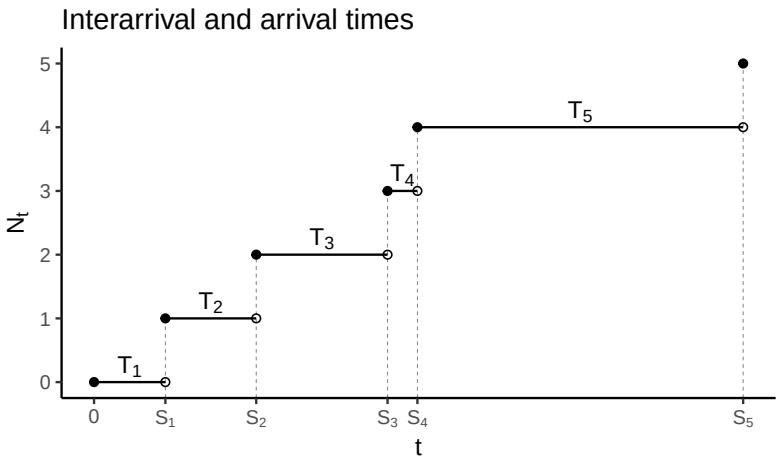


Figure 3.3: *Illustration of interarrival times $T_1, T_2, \ldots$ and arrival times $S_1, S_2, \ldots$ of a Poisson process.*

Definition 3.4

The **interarrival times** of a Poisson process are the random variables measuring the lengths of time intervals between successive events. We denote as T_n the inter-arrival time between the $(n-1)^{\text{th}}$ and the n^{th} events (where T_1 is the time of the first event).

Definition 3.5

The **arrival times** of a Poisson process are the times at which events occur. We denote as S_n the arrival time for the n^{th} event, i.e., $S_n = T_1 + T_2 + \cdots + T_n$.

The terminology of "arrivals" comes from queueing theory, where the focus is on modelling the

arrivals and departures of customers from a queue. The interarrival times are also sometimes called "waiting times", and the arrival times are sometimes called "event times".

3.1.2 Infinitesimal definition

An alternative definition of the Poisson process describes the distribution of points over infinitesimal time intervals. It requires the little-o notation, and we first define this. We write $f(h) = o(h)$ (read as "little-o of h"), if

$$\lim_{h \to \infty} \frac{f(h)}{h} = 0.$$

That is, we use $o(h)$ to denote any terms that are small relative to h, in the technical sense described above.

Definition 3.6 (alternative definition of Poisson process)

The counting process $(N_t)_{t \geq 0}$ is a Poisson process with rate $\lambda > 0$ if

1. $N_0 = 0$;
2. $N(t)$ has independent increments;
3. $\Pr(N_{t+h} - N_t = 1) = \lambda h + o(h)$;
4. $\Pr(N_{t+h} - N_t > 1) = o(h)$.

We will not prove the equivalence of the two definitions, but it turns out that these conditions imply that the count of events over a time interval follows a Poisson distribution.

Some books present all continuous-time stochastic processes using the little-o notation, so it is good to understand the intuition behind it. Essentially, condition 4 ensures that there cannot be more than one event in a very short time interval (or at least, the probability of that happening is negligible). Condition 3 states that the probability of an event occuring over a short time interval is approximately proportional to the rate parameter of the process, λ.

3.2 Distribution of interarrival times

Proposition 3.1

Consider a Poisson process (N_t) with rate parameter $\lambda > 0$. The interarrival times of (N_t) $(T_1, T_2, ...)$ are independent exponentially distributed random variables with rate parameter λ. That is, they have the probability distribution function $f(t) = \lambda e^{-\lambda t}$, for $t \geq 0$

Proof

We first prove the result for T_1, the time of the first event. We have $\Pr(T_1 > t) = \Pr(N_t = 0) = e^{-\lambda t}$. We recognise the cumulative distribution function of the exponential distribution, so $T_1 \sim \text{Exp}(\lambda)$.

Then, we turn to T_n, the interarrival time between the $(n-1)^{\text{th}}$ and the n^{th} events, for $n \geq 2$:

$$\begin{aligned}
\Pr(T_n > t) &= \int_0^\infty \Pr(T_n > t \mid S_n = s) f_{S_n}(s)\, ds && \text{(a)} \\
&= \int_0^\infty \Pr(N_{s+t} - N_s = 0 \mid N_s = n) f_{S_n}(s)\, ds && \text{(b)} \\
&= \int_0^\infty \Pr(N_t = 0) f_{S_n}(s)\, ds && \text{(c)} \\
&= e^{-\lambda t} \int_0^\infty f_{S_n}(s)\, ds && \text{(d)} \\
&= e^{-\lambda t}, && \text{(e)}
\end{aligned}$$

showing that T_n follows an exponential distribution with rate λ. In the derivation above, (a) is the continuous version of the law of total probability, (b) translates a statement about interarrival times (T_n) to be in terms of counts (N_t), (c) uses the stationarity and independence of increments in a Poisson process, (d) substitutes the probability mass function of the Poisson distribution, and (e) follows from the fact that probability density functions integrate to 1 (including f_{S_n}).

The exponential distribution has a mode of zero, and a mean of $1/\lambda$. This is consistent with our intuition of the Poisson process: the higher the rate, the shorter the interarrival times. The exponential distribution is illustrated in Figure 3.4.

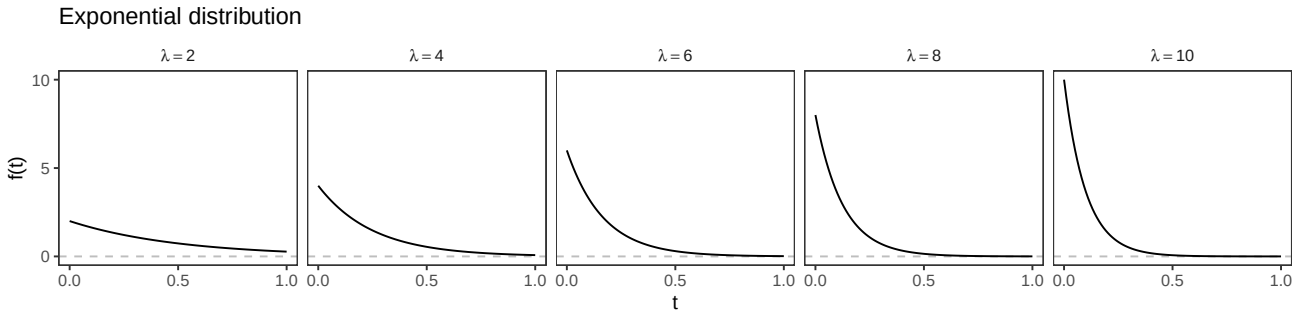


Figure 3.4: *Probability density function of exponential distribution for different values of the rate parameter λ.*

3.2.1 Memorylessness

The above result on the distribution of interarrival times is important because the exponential distribution is the only continuous distribution with the memorylessness property.

Definition 3.7

A random variable X is **memoryless** if, for all $s, t > 0$,

$$\Pr(X > t + s \mid X > t) = \Pr(X > s).$$

Because the interarrival times $(T_1, T_2, ...)$ of a Poisson process follow an exponential distribution, they have the memoryless property, and so we have

$$\Pr(T_n > t + s \mid T_n > t) = \Pr(T_n > s),$$

for all $s, t > 0$. In words, this means that, regardless of how long we have been waiting since the last event (which occurred at time t), the distribution of the time we still need to wait until the next event is the same as the distribution of the original waiting time.

Example 3.2

1. Assume that a bakery has on average 6 customers per hour, and that the interarrival times between customers follow an exponential distribution (with rate parameter $\lambda = 6$). The mean interarrival time is 1/6 hour = 10 minutes. Let's say that the last customer came 30 minutes ago. This is an unusually long waiting time, so we might expect that the next customer is likely to arrive in the next few minutes, but this would be incorrect. In fact, regardless of the 30-minute wait, the distribution of the waiting time until the next customer arrives is still an exponential distribution with mean 10 minutes. This is because different customers arrive at the bakery independently, and the arrival of the next customer does not depend on the last customer.

2. A physical example of memorylessness is radioactive decay. Carbon-14 atoms decay into Nitrogen-14 over time, which takes 8267 years on average (this is called the "mean-life" of Carbon-14). For a given atom, time until decay follows an exponential distribution with rate parameter $\lambda = 1/8267$. The probability that the atom decays in the next year does not depend on its age.

3.2.2 Simulating from a Poisson process

The exponential distribution of interarrival times give us a convenient way to simulate from a Poisson process, based on the following algorithm. We initialise $S_0 = 0$ and, for $n = 1, 2, ...,$

1. generate an interarrival time T_n from the exponential distribution;
2. compute the arrival time $S_n = S_{n-1} + T_n$;
3. let $N_t = n - 1$ for $S_{n-1} \leq t < S_n$.

In practice, we simulate the interarrival and the arrival times (e.g., using `rexp()` in R), and store those. Then, for a given time t, we find the corresponding value of the Poisson process (n) by looking for the two successive arrival times S_n and S_{n+1} such that $S_n \leq t < S_{n+1}$. Figure 3.5 shows four example realisations of a Poisson process with rate parameter $\lambda = 0.8$ over $0 \leq t \leq 10$, simulated using this algorithm.

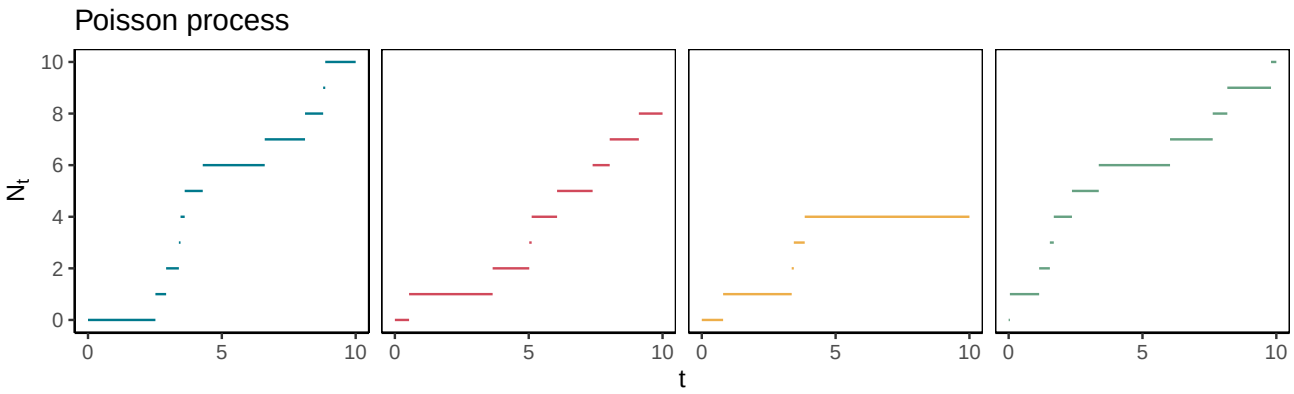


Figure 3.5: *Four realisations of a Poisson process with rate* $\lambda = 0.8$.

The following R chunk shows some example code that could be used to simulate from that Poisson process, and to find the count N_t based on the sequence of simulated arrival times. This code could easily be modified to change the rate, the time window, or the number of realisations.

```
# Set random seed for reproducibility
set.seed(652)

# Set a few parameters
tmax <- 10
rate <- 0.8

# Loop until reaching tmax
times <- 0
while(times[length(times)] < tmax) {
    interarrival_time <- rexp(n = 1, rate = rate)
    arrival_time <- times[length(times)] + interarrival_time
    times <- c(times, arrival_time)
}
times
```

```
 [1]  0.000000  2.518512  2.917831  3.395475  3.457126  3.612453  4.286739
 [8]  6.602555  8.107993  8.785203  8.862875 10.029713
```

```r
# Get event count in [0, 5] from arrival times
count <- length(which(times < 5))
count
```

```
[1] 7
```

3.3 Distribution of arrival times

There are two different results about the distribution of arrival times in a Poisson process: the distribution of one arrival time S_n, and the joint distribution of all arrival times conditional on the number of events.

3.3.1 Marginal distribution of S_n

We first present a basic result about the distribution of the sum of independent random variables, which uses an operation called the convolution of two functions.

Proposition 1.2

Let X and Y be two independent continuous random variables, and $Z = X + Y$. The probability density function of Z is

$$f(Z = z) = \int_{-\infty}^{\infty} f(X = z - y) f(Y = y)\, dy.$$

The intuition behind this formula is that we consider every possible combination of values for X and Y that would yield $Z = z$. Because Z is defined as the sum of X and Y, we know that, if $Z = z$ and $Y = y$, then we must have $X = z - y$. So, to go through all possible combinations, we integrate ("sum") over all possible values for Y, and this also determines the value of X. This is illustrated visually in Grant Sanderson's Youtube video "Convolutions | Why X+Y in probability is a beautiful mess". A similar result applies in the case of two discrete random variables, with a sum instead of an integral.

Proposition 3.3

Consider a Poisson process with parameter λ, and let S_n be the n^{th} arrival time for that process. Then, $S_n \sim \text{gamma}(n, \lambda)$, where n is called the shape parameter, and λ the rate parameter of the gamma distribution. That is, the probability density function of S_n is

$$f_{S_n}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t}, \quad \text{for } t > 0.$$

Proof

By definition, $S_n = T_1 + T_2 + \cdots + T_n$ is the sum of n independent random variables that all follow an exponential distribution with rate parameter λ. We can use a proof by induction to show that a random variable defined by this sum follows a gamma distribution with shape n and rate λ.

1. **Base case:** Show that this holds for $n = 1$.
   We have $S_1 = T_1$, and we know that $T_1 \sim Exp(\lambda)$, so

   $$f_{S_1}(s) = f_{T_1}(s) = \lambda e^{-\lambda s},$$

   and this can be under the form of a gamma density function as

   $$f_{S_1}(s) = \frac{\lambda^n e^{-\lambda s} s^{n-1}}{(n-1)!}, \quad \text{where } n = 1.$$

   That is, $S_1 \sim \text{gamma}(1, \lambda)$, so the hypothesis stands for $n = 1$.

2. **Induction step:** Show that, if $S_{n-1} \sim \text{gamma}(n-1, \lambda)$, then $S_n = S_{n-1} + T_n \sim \text{gamma}(n, \lambda)$.
   By assumption, we have

   $$f_{S_{n-1}}(s) = \frac{\lambda^{n-1} e^{\lambda s} s^{n-2}}{(n-2)!}, \quad \text{and } f_{T_n}(t) = \lambda e^{-\lambda t}.$$

   If we define $S_n = S_{n-1} + T_n$, we can derive its density function by convolution,

   $$\begin{aligned}
   f_{S_n}(s) &= \int_0^s f_{S_{n-1}}(z) f_{T_n}(s-z) dz \\
   &= \int_0^s \frac{\lambda^{n-1} e^{-\lambda z} z^{n-2}}{(n-2)!} \lambda e^{-\lambda(s-z)} dz \\
   &= \frac{\lambda^n e^{-\lambda s}}{(n-2)!} \int_0^s z^{n-2} dz \\
   &= \frac{\lambda^n e^{-\lambda s}}{(n-2)!} \left[\frac{z^{n-1}}{n-1} \right]_0^s \\
   &= \frac{\lambda^n e^{-\lambda s}}{(n-2)!} \left(\frac{s^{n-1}}{n-1} - 0 \right) \\
   &= \frac{\lambda^n e^{-\lambda s} s^{n-1}}{(n-1)!}
   \end{aligned}$$

   which is the density function of a gamma distribution with shape n and rate λ, as required.

The mean and variance of the gamma distribution can be conveniently expressed in terms of the shape and rate parameters, so we also have that

$$E[S_n] = \frac{n}{\lambda} \quad \text{and} \quad Var[S_n] = \frac{n}{\lambda^2}.$$

That is, the expected value for the time n^{th} event is proportional to n and inversely proportional to λ.

Figure 3.6 shows the probability density function of the gamma distribution for several combinations of the shape n and rate λ.

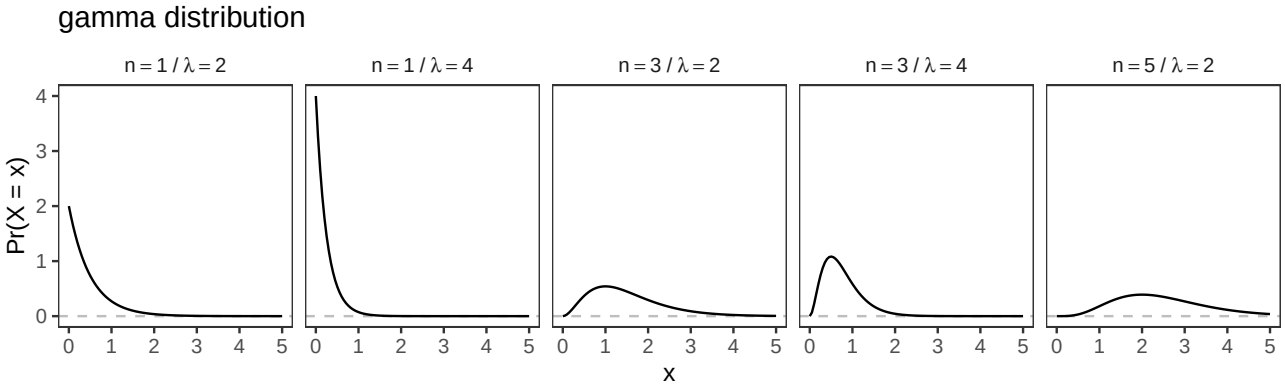


Figure 3.6: *Probability density function of gamma distribution for different parameter values.*

In this context, the shape parameter must be a positive integer ($n \in \mathbb{R}_{>0}$); when this is the case, the gamma distribution is sometimes called the Erlang distribution. Note that, generally, the shape of the gamma distribution can be any positive real number, and the $(n-1)!$ in the denominator of the probability density function is replaced by the Gamma function $\Gamma(n)$ (which generalises the factorial to non-integers), but this is not needed in the present context.

3.3.2 Conditional joint distribution of $S_1, \ldots, S_n$

We then turn to the joint distribution of arrival times over $(0, t)$ conditional on the number of events n in that interval. We must first define the concept of order statistic. If $X_1, X_2, \ldots, X_n$ are n random variables, we call $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ the corresponding order statistics if $X_{(k)}$ is the k^{th} smallest value among the $\{X_k\}$. That is, the order statistics are the random variables in increasing order. If the $\{X_k\}$ are independent and identically distributed random variables with probability density function f_X, then the joint density of the order statistics $\{X_{(k)}\}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f_X(x_i), \quad \text{where } x_1 < x_2 < \cdots < x_n.$$

In particular, if the $\{X_k\}$ are independent and uniformly distributed over the interval $(0, t)$, the joint density of the corresponding order statistics is

$$f_{X_{(1)},\dots,X_{(n)}}(x_1, \dots, x_n) = \frac{n!}{t^n}, \quad \text{where } 0 < x_1 < x_2 < \cdots < x_n < t.$$

The reason that the $n!$ factor appears is that there are $n!$ combinations of the unordered variables that give rise to a given ordered sequence $x_1, x_2, \dots, x_n$ (the $n!$ permutations of $x_1, x_2, \dots, x_n$). It might seem odd that the joint probability density function of n uniform variables can be multiplied by $n!$ and still give rise to a valid probability density function. This is because, when the variables are ordered, this decreases the domain of the distribution (over which the density is non-zero) by a factor $n!$. This is illustrated in two dimensions in Figure 3.7, where the domain of the distribution is divided by two, and in three dimensions in the interactive plot below (which will unfortunately only show in the html version of the notes, not the PDF file).

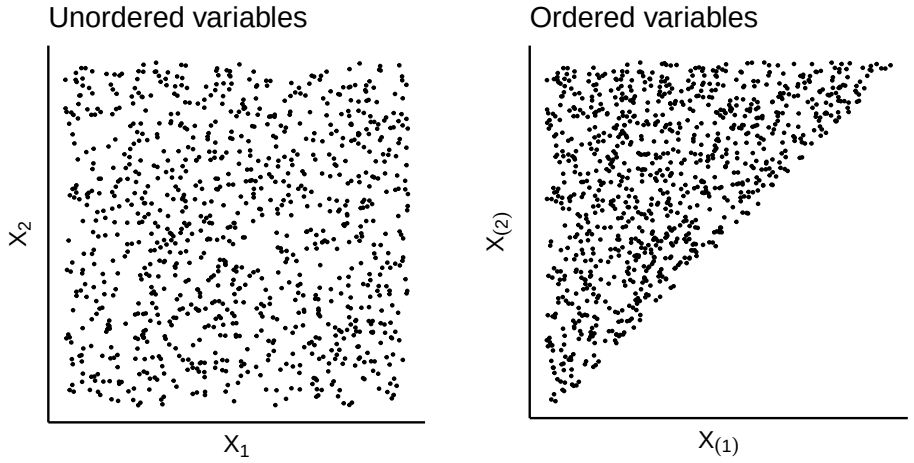


Figure 3.7: *Simulation of order statistics corresponding to two independent uniform variables.*

../../../../../../tmp/RtmpolYbBY/file8886501f0fcd.png

We can use the concept of order statistic to desribe the joint conditional distribution of the arrival times of a Poisson process.

Proposition 3.4

Consider a Poisson process (N_t), and let S_n be the n^{th} arrival time for that process. Given that $N_t = n$, the n arrival times $S_1, S_2, \dots, S_n$ are distributed as the order statistics of a uniform distribution on $(0, t)$.

Proof

The proof requires the derivation of the joint probability density of $\{S_1, \dots, S_n, N_t\}$, which we can rewrite in terms of the distributions of interarrival times $T_1, \dots, T_n$. If the arrival times are $S_1 = s_1, \dots, S_n = s_n$ and the number of events is $N_t = n$, then we know that

the n first interarrival times are $T_1 = s_1, T_2 = s_2 - s_1, \dots, T_n = s_n - s_{n-1}$, and that the $(n+1)^{\text{th}}$ interarrival time must satisfy $T_{n+1} > t - s_n$.
Then, we have

$$f_{S_1,\dots,S_n|N_t}(s_1,\dots,s_n \mid N_t = n) = \frac{f_{S_1,\dots,S_n,N_t}(s_1,\dots,s_n,n)}{\Pr(N_t = n)} \quad \text{(a)}$$

$$= \frac{f_{T_1}(s_1) f_{T_2}(s_2 - s_1) \dots f_{T_n}(s_n - s_{n-1}) \Pr(T_{n+1} > t - s_n)}{\Pr(N_t = n)}, \quad \text{(b)}$$

where (a) comes from the definition of conditional probability, and (b) translates the statement in terms of interarrival times.

The f_{T_i} are given by the probability density of the exponential distribution, $\Pr(T_{n+1} > t - s_n)$ is obtained from the cumulative distribution function of the exponential distribution, and $\Pr(N_t = n)$ is the probability mass function of a Poisson distribution. Making these substitutions, we find

$$\begin{aligned} f_{S_1,\dots,S_n}(s_1,\dots,s_n \mid N_t = n) &= \frac{\lambda e^{-\lambda s_1} \lambda e^{-\lambda(s_2 - s_1)} \dots \lambda e^{-\lambda(s_n - s_{n-1})} e^{-\lambda(t - s_n)}}{(\lambda t)^n e^{-\lambda t}/n!} \\ &= \frac{\lambda^n \exp(\lambda(s_1 - s_1 + s_2 - s_2 + \dots + s_n - s_n - t))}{\lambda^n t^n \exp(-\lambda t)/n!} \\ &= \frac{\lambda^n e^{-\lambda t}}{\lambda^n e^{-\lambda t} t^n/n!} \\ &= \frac{n!}{t^n}, \end{aligned}$$

as required.

Another way to state the proposition above is that the arrival times, considered as unordered random variables, are uniformly distributed conditionally on the number of events in $(0, t)$. Note that we cannot state that the distribution of event times is therefore uniform over $[0, \infty)$, because the uniform distribution is only well defined over finite intervals.

This result suggests an alternative method to simulate from a Poisson process:

1. set t, and simulate $N_t \sim \text{Poisson}(\lambda t)$;
2. simulate $U_1, U_2, \dots, U_{N_t} \sim \text{Unif}(0, t)$;
3. define the arrival times as $S_i = U_{(i)}$, where $U_{(i)}$ is the i^{th} smallest value in $\{U_1, \dots, U_n\}$;
4. let $N_s = n - 1$ over $S_{n-1} \leq s < S_n$.

Example 3.3

Assume that the occurence of major earthquakes in Canada since January 1, 2000 can be described by a Poisson process with rate $\lambda = 1$ per year.

1. Find the probability that the 20th major earthquake since January 1, 2000 occured in 2022.

   The arrival time of the 20th earthquake, S_{20}, follows a gamma distribution with shape 20 and with rate 1 (where January 1, 2000 is treated as $t = 0$). We want the probability

   $$\begin{aligned}\Pr(22 < S_{20} < 23) &= \Pr(S_{20} < 23) - \Pr(S_{20} < 22) \\ &= 0.762 - 0.694 \\ &= 0.068\end{aligned}$$

   using the cumulative distribution function of the gamma distribution (e.g., `pgamma()` in R).

2. Given that 20 major earthquakes took place in Canada between January 1, 2000 and December 31, 2022, what is the probability that the 20th earthquake occured in 2022?

   We want the probability that the last earthquake took place in 2022, i.e., that the maximum of 20 random variables from Unif$(0, 22)$ is greater than 22. Let $U_1, U_2, \dots, U_{20}$ be Unif$(0, 22)$ random variables, and let M be their maximum. We can get the conditional probability as follows,

   $$\begin{aligned}\Pr(22 < S_{20} < 23 \mid N_{23} = 20) &= \Pr(22 < M < 23) \\ &= 1 - \Pr(M \leq 22) \\ &= 1 - \Pr(U_1 \leq 22, U_2 \leq 22, \dots, U_{20} \leq 22) \\ &= 1 - \Pr(U_1 \leq 22) \Pr(U_2 \leq 22) \cdots \Pr(U_{20} \leq 22) \\ &= 1 - \Pr(U_1 \leq 22)^{20} \\ &= 1 - (22/23)^{20} \\ &= 1 - 0.411 \\ &= 0.589\end{aligned}$$

Notably, the answers to questions 1 and 2 are different.

3.4 Statistical inference

Given a sequence of event times $(S_1, \dots, S_N)$, we want to estimate the rate parameter λ of the Poisson process. The simplest way to do this is to use the distribution of interarrival times to derive the likelihood, and optimise it with respect to λ.

For $n = 1, \dots, N$, the interarrival times $T_n = S_n - S_{n-1}$ arise from an exponential distribution with rate parameter λ (where we define $S_0 = 0$). The likelihood function is therefore

$$L(\lambda \mid S_1, \dots, S_N) = \prod_{n=1}^{N} \lambda e^{-\lambda T_n},$$

and the log-likelihood is

$$\ell(\lambda \mid S_1, \dots, S_N) = \sum_{n=1}^{N} (\log(\lambda) - \lambda T_n) = N \log(\lambda) - \lambda \sum_{n=1}^{N} T_n.$$

To find the maximum likelihood estimator, we differentiate the log-likelihood with respect to λ and we set to zero:

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda}(\hat{\lambda} \mid S_1, \dots, S_N) = 0 \Rightarrow \frac{N}{\hat{\lambda}} - \sum_{n=1}^{N} T_n = 0 \\ \Rightarrow \hat{\lambda} = \frac{N}{\sum_{n=1}^{N} T_n} \\ \Rightarrow \hat{\lambda} = \frac{N}{S_N}. \end{aligned}$$

This result is intuitive: our best guess for the rate of the process (i.e., expected number of events per unit time) is the number of observed events over the length of the period of observation.

Example 3.4

The Old Faithful is a geyser in the Yellowstone National Park, Wyoming, USA, which erupts at very predictable intervals. The data set of eruption times and inter-eruption intervals is a classic example used to illustrate time series and point process models. It is automatically loaded in R as the `faithful` data object; it has one column for durations or eruptions, and one column for inter-eruption waiting times (i.e., interarrival times). Here, we are interested in the latter.

1. Assuming that the eruptions of the Old Faithful geyser can be described as a Poisson process, what is maximum likelihood estimate of the eruption rate?

```
# Load data
data("faithful")
head(faithful)
```

```
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
```

```
# Compute MLE of rate
lambda <- nrow(faithful) / sum(faithful$waiting)
lambda
```

```
[1] 0.01410496
```

The estimated rate is $\hat{\lambda} = 0.014$ eruptions per minutes. That is, the estimated mean waiting time between eruptions is $1/0.014 = 70.9$ min (which is just the average of the `waiting` column of the data frame).

2. What is the probability that more than 30 eruptions take place on a given day?

   The number of eruptions during a day follows a Poisson distribution with rate $\lambda(24 \times 60)$, i.e., the rate parameter multiplied by the number of minutes in a day. The probability is

$$\begin{aligned} \Pr(N > 30) &= 1 - \Pr(N \leq 30) \\ &= 1 - 0.984 \\ &= 0.016, \end{aligned}$$

   using the cumulative distribution function of the Poisson distribution.

3.5 Non-homogeneous Poisson process

In many applications, the rate of events is not constant through time. For example, hurricanes on the East coast of North America are more common in the summer months, and arrivals of customers at a restaurant are more frequent around lunch and dinner. This phenomenon can be modelled with the non-homogeneous Poisson process, an extension of the Poisson process where the rate depends on time. We call the time-varying rate $\lambda(t)$ the intensity function of the process.

Definition 3.8

A counting process $(N_t)_{t \geq 0}$ is a **non-homogeneous Poisson process** with intensity function $\lambda(t)$ if

1. $N_0 = 0$

2. for all $t > 0$, N_t has a Poisson distribution with mean

$$E[N_t] = \int_0^t \lambda(s)\, ds;$$

3. (N_t) has independent increments.

The Poisson process is a special case where $\lambda(t) = \lambda$ is constant through time. Note that the non-homogeneous Poisson process does not generally have stationary increments, because the distribution of an increment depends on $\lambda(t)$. For this reason, it is also sometimes called the non-stationary Poisson process.

Proposition 3.5

If (N_t) is a non-homogeneous Poisson process with intensity function $\lambda(t)$, then, for $0 < s < t$, we have

$$N_t - N_s \sim \text{Poisson}\left(\int_s^t \lambda(x)\, dx\right).$$

That is, the number of events that occur between s and t follows a Poisson distribution, with rate the integral of the intensity function over that interval.

The intuition is that more events will take place over time intervals where the intensity function is high. This is illustrated with an example in Figure 3.8.

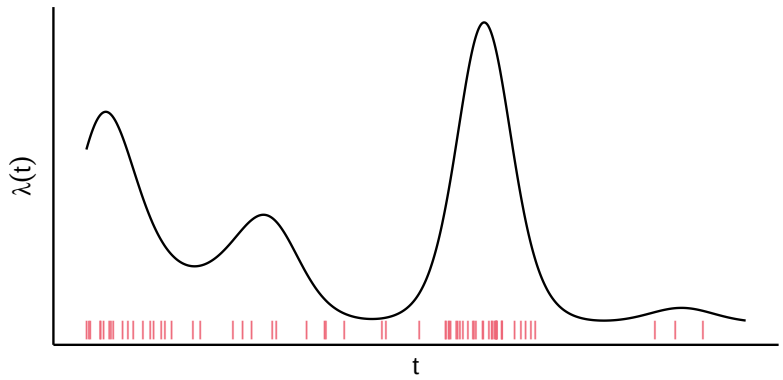


Figure 3.8: *Example of intensity function $\lambda(t)$ (black line), and simulated event times from the corresponding non-homogeneous Poisson process (red vertical ticks).*

3.6 Merging and splitting Poisson processes

3.6.1 Merging Poisson processes

Proposition 3.6

Let $(N_t^{(1)}), (N_t^{(2)}), \ldots, (N_t^{(m)})$ be m independent Poisson processes with rates $\lambda_1, \ldots, \lambda_m$, respectively. The process (N_t) defined by

$$N_t = N_t^{(1)} + N_t^{(2)} + \cdots + N_t^{(m)}, \quad \text{for all } t \geq 0$$

is a Poisson process with rate $\lambda_1 + \lambda_2 + \cdots + \lambda_m$.

We omit the proof here; it first requires showing that the sum of independent Poisson random variables is also a Poisson random variable, and applying this to the number of events over a time interval.

Figure 3.9 shows an example of a process constructed by merging two Poisson processes.

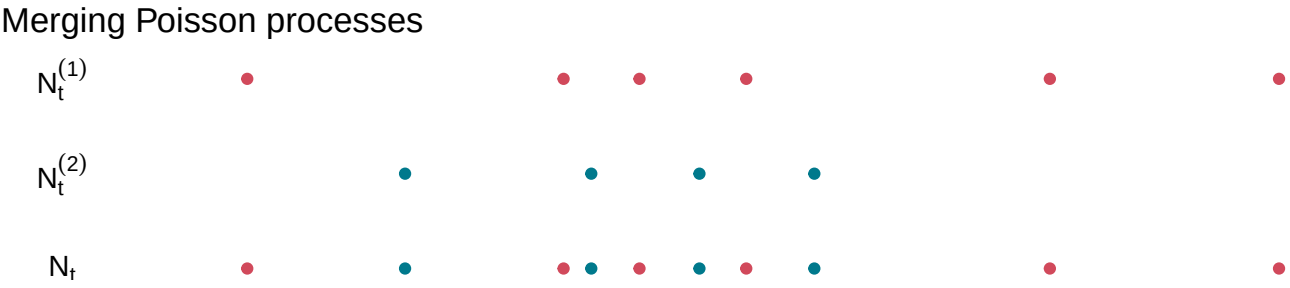


Figure 3.9: *Event times for three Poisson processes, where (N_t) is the process defined by $N_t = N_t^{(1)} + N_t^{(2)}$.*

Example 3.5

Consider a soccer game between teams A and B, and assume that goals scored by the teams can be modelled with two Poisson process, with rates $\lambda_A = 1.1$ and $\lambda_B = 1.4$ (goals per hour), respectively.

1. What is the probability that no goals have been scored by the end of the game, i.e., after 90 minutes?

   The total number of goals scored by both teams follows a Poisson process with rate $\lambda = 1.1 + 1.4 = 2.5$. In particular, the number of goals scored after 90 minutes (= 1.5 hour) follows a Poisson distribution with rate $1.5 \times 2.5 = 3.75$, so the required probability is

   $$\Pr(X = 0) = \frac{3.75^0 \times e^{-3.75}}{0!} = e^{-3.75} = 0.024,$$

so there is a 2.4% probability that zero goals will be scored during the game.

2. What is the probability that team B wins?

$$\begin{aligned}
\Pr(N_A(T) < N_B(T)) &= \sum_{n=0}^{\infty}\sum_{k=1}^{\infty} \Pr(N_A(T) = n, N_B(T) = n+k) \\
&= \sum_{n=0}^{\infty}\sum_{k=1}^{\infty} \Pr(N_A(T) = n)\Pr(N_B(T) = n+k) \\
&= \sum_{n=0}^{\infty}\sum_{k=1}^{\infty} \frac{(\lambda_A T)^n}{(n)!} e^{-\lambda_A T} \frac{(\lambda_B T)^{n+k}}{(n+k)!} e^{-\lambda_B T} \\
&= e^{-(\lambda_A+\lambda_B)T} \sum_{n=0}^{\infty} \left[\frac{(\lambda_A \lambda_B T^2)^n}{n!} \sum_{k=1}^{\infty} \frac{(\lambda_B T)^k}{(n+k)!} \right].
\end{aligned}$$

Substituting the values of λ_A, λ_B and T, this becomes

$$\begin{aligned}
\Pr(N_A(1.5) < N_B(1.5)) &= e^{-1.5\times(1.1+1.4)} \sum_{n=0}^{\infty} \left[\frac{(1.1 \times 1.4 \times 1.5^2)^n}{n!} \sum_{k=1}^{\infty} \frac{(1.4 \times 1.5)^k}{(n+k)!} \right] \\
&= e^{-3.75} \sum_{n=0}^{\infty} \left[\frac{3.465^n}{n!} \sum_{k=1}^{\infty} \frac{2.1^k}{(n+k)!} \right] \\
&\approx 0.316.
\end{aligned}$$

You should think about how you would use R to compute an approximation of the series above.

3.6.2 Splitting a Poisson process

Proposition 3.7

Let (N_t) be a Poisson process with rate λ. Assume that each event is marked as a "type k" event with probability p_k for $k = 1, \ldots, K$, where $p_1 + \cdots + p_K = 1$. Let $N_t^{(k)}$ be the number of events of type k in $[0, t]$. Then, the processes $(N_t^{(1)}), \ldots, (N_t^{(K)})$ are independent Poisson processes with rates $\lambda p_1, \ldots, \lambda p_K$, respectively.

Each component process $(N_t^{(k)})$ is called a **thinned Poisson process**, because it represents a "thinned" sequence of events, where events are included and excluded with some probability. Thinned Poisson processes are often used in contexts where the process of interest is only observed partially, with some detection probability associated with each event.

Example 3.6

We are interested in the frequency of Northern lights in Churchill, Manitoba. We set up an instrument to record them, but it only works during the night and when the sky is clear. Over the course of a year, it is on average dark enough for the detector to work 40% of the time, and the sky is clear around 60% of the time.

1. How can we model the sequence of recorded Northern lights with a Poisson process?

   We assume that the number of Northern lights follows a Poisson process with rate λ (in events per year). Each event is either observed or not, with some probability, so the recorded Northern lights follow a thinned Poisson process. The detection probability is $0.4 \times 0.6 = 0.24$, so the rate of the thinned process is $\lambda_{\text{thin}} = 0.24\lambda$. That is, there is a 24% probability of detecting a given Northern light with the instrument.

2. The detector recorded 27 Northern lights last year. Estimate the rate of Northern lights in Churchill.

   The number of *recorded* events gives us an estimate of the rate of the thinned process, $\hat{\lambda}_{\text{thin}} = 27/1 = 27$. We are interested in the rate of the non-thinned process, though, and we can calculate the estimate as $\hat{\lambda} = \hat{\lambda}_{\text{thin}}/0.24 = 112.5$. We estimate that the rate of Northern lights in Churchill is 112.5 per year.

4 Continuous-time Markov processes

We will now describe continuous-time Markov processes, which include Poisson processes as a special case. We will continue using the letter t to represent time as a continuous variable (as opposed to the discrete index n in Chapter 2).

4.1 Introduction

4.1.1 Definition

The discrete-time Markov property is written in terms of a sequence of random variables defined over a regular time grid. It needs to be slightly modified in the continuous-time setting.

Definition 4.1

The process $(X_t)_{t\geq 0}$ is a **continuous-time Markov process** with countable state space $\mathcal{S}$ if, for all $0 \leq r \leq s \leq t$ and $i, j \in \mathcal{S}$,

$$\Pr(X_t = j \mid X_s, X_r) = \Pr(X_t = j \mid X_s).$$

This is the continuous-time version of the Markov property.

This is very similar to the definition of a discrete-time Markov chain. In the discrete-time case, we said that the process at time $n+1$ was independent of its values at times $\{0, \dots, n-1\}$ conditionally on its last value, i.e., at time n. In continuous time, there is no canonical time interval, so the property is instead defined for three arbitrary times $r \leq s \leq t$ on $[0, \infty)$. Given several past values of the process (X_s and X_r), only the most recent (X_s) is informative to write the distribution of the current value of the process (X_t).

Continuous-time Markov processes can be defined over countable (discrete) or uncountable (continuous) state spaces, and we will mostly focus on the countable case, where they are sometimes called "Markov jump processes" (because the process jumps between discrete values). We will talk about "jumps", "switches", and "transitions" interchangeably.

Definition 4.2

A continuous-time Markov chain (X_t) is **time-homogeneous** if, for any $s, t \geq 0$,

$$\Pr(X_{s+t} = j \mid X_s = i) = \Pr(X_t = j \mid X_0 = i), \quad \text{for } i, j \in \mathcal{S}.$$

That is, the probability of a transition over some time interval does not depend on the start time of the interval.

In this chapter, we will only consider time-homogeneous Markov chains, i.e., whose dynamics are constant through time.

4.1.2 Holding times

Just like in the discrete-time case, thinking about the distribution of holding times (i.e., times between state transitions) is useful to understand what realisations from a continuous-time Markov process look like. Let D_i be the holding time in state i, i.e., the amount of time the process stays in state i before switching to another state. Unlike in the discrete-time case, D_i is a continuous variable here, defined over $[0, \infty)$. It turns out that the Markov property fully determines the distribution of holding times.

Proposition 4.1

The dwell time of a continuous-time Markov process follows an exponential distribution.

Proof

We first prove that that the dwell time is memoryless, i.e., $\Pr(D_i > s + t \mid D_i > s) = \Pr(D_i > t)$.

$$\begin{aligned}
\Pr(D_i > s + t \mid D_i > s) &= \Pr(X_u = i \text{ for } u \in [0, s + t] \mid X_u = i \text{ for } u \in [0, s]) \\
&= \Pr(X_u = i \text{ for } u \in [s, s + t] \mid X_u = i \text{ for } u \in [0, s]) \\
&= \Pr(X_u = i \text{ for } u \in [s, s + t] \mid X_s = i) && \text{(a)} \\
&= \Pr(X_u = i \text{ for } u \in [0, t] \mid X_0 = i) && \text{(b)} \\
&= \Pr(D_i > t),
\end{aligned}$$

where (a) follows from the Markov property, and (b) follows from the time-homogeneity of the chain.

The exponential distribution is the only memoryless probability distribution with continuous support, and so the dwell time must be exponentially distributed. A proof of this special feature of the exponential distribution is for example presented in Section 5.2.2 of

Ross (2019).

This property does not tell us what the rate of the exponential distribution is, or how we determine what state to jump to at the end of the holding time, so we don't yet have enough information to simulate from a continuous-time Markov chain. We will answer those questions in the next section.

Figure 4.1 shows an example of a 2-state continuous-time Markov process with $\mathcal{S} = \{0, 1\}$. The times between state transitions do not occur on a predefined grid like in the discrete-time case; instead, they can occur at any continuous time.

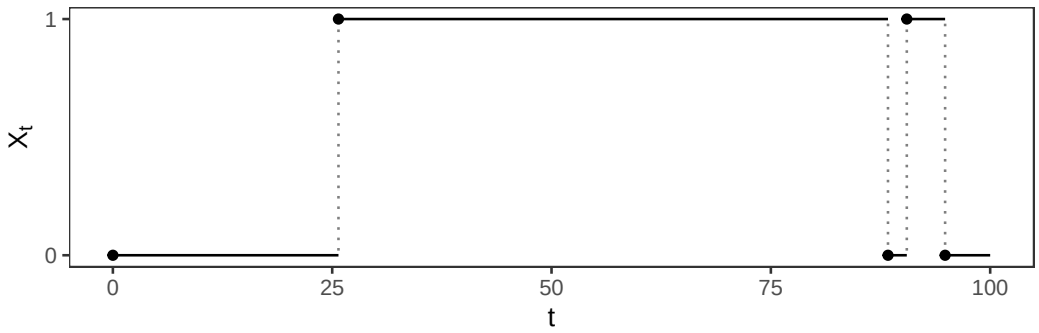


Figure 4.1: *Example realisation from a continuous-time Markov process with state space* $\mathcal{S} = \{0, 1\}$, *for* $t \in [0, 100]$.

4.2 Model specification

4.2.1 Transition rates

Because time is now continuous, there is a no particular time grid of interest over which to define transition probabilities. However, because we know that the holding times follow an exponential distribution, the model can instead be specified in terms of two sets of parameters:

1. the rate parameter of the holding time distribution in each state i;
2. the probabilities of jumping from any state i to any other state $j \neq i$.

With this in mind, a continuous-time Markov process can be described as follows. When the process enters some state i, a waiting time is generated from an exponential distribution with rate parameter $q_{ij} > 0$ for each other state $j \neq i$, say D_{ij}. Then, the process jumps to the state with the shortest waiting time (out of all the $j \neq i$). The holding time before a transition is therefore $D_i = \min\{D_{ij}\}_{j \neq i}$, and it can be shown that it follows an exponential distribution (as required by the Markov property), with rate $q_i = \sum_{j \neq i} q_{ij}$. By property of the exponential distribution, the expected holding time in state i is $1/q_i$.

Once we have generated a holding time $D_i \sim \text{Exp}(q_i)$, then, how do we know which state to jump to after D_i? When it leaves state i, the process switches to the state with the minimum

waiting time; for each state $j \neq i$, this occurs with probability

$$\widetilde{P}_{ij} = \frac{q_{ij}}{\sum_{k \neq i} q_{ik}} = \frac{q_{ij}}{q_i}.$$

The discrete-time Markov chain with transition probabilities $\widetilde{P}_{ij}$ is called the embedded chain, or sometimes the skeleton of the continuous-time process.

The dynamics of the process are therefore fully defined by the transition rates $\{q_{ij}\}_{i \neq j}$. A Markov process with finite state space of size $|\mathcal{S}| = N$ has $N \times (N-1)$ such transition rates. It is convenient to write the transition rates in the form of a matrix where, by convention, the i^{th} diagonal entry is set to $q_{ii} = -q_i$,

$$Q = \begin{pmatrix} -q_0 & q_{01} & q_{02} & \cdots \\ q_{10} & -q_1 & q_{12} & \cdots \\ q_{20} & q_{21} & -q_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

such that the rows sum to zero: $\sum_j q_{ij} = 0$ for all $i \in \mathcal{S}$. Q is called the **transition rate matrix**, or **infinitesimal generator matrix** of the process.

We can represent a continuous-time Markov chain as a directed weighted graph, where the edges are weighted by the transition rates. Unlike the transition graphs of Chapter 2, a transition rate graph never has arrows from one state to itself.

Example: Figure 4.2 shows the transition rate graph of the continuous-time Markov chain with transition rate matrix

$$Q = \begin{pmatrix} -3 & 1 & 2 \\ 0 & -1 & 1 \\ 3 & 1 & -4 \end{pmatrix}$$

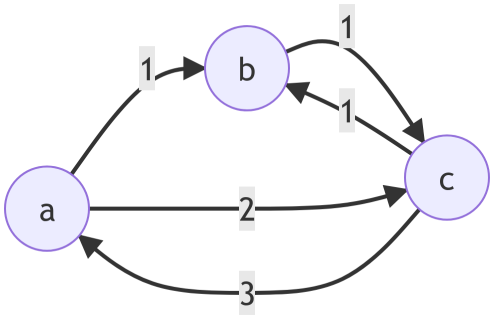


Figure 4.2: *Example transition rate graph of 3-state continuous-time Markov chain*

The label of each edge gives the transition rate, and can be interpreted as the frequency of the given transition in the long run. In this model, transitions from "a" to "c" are twice as frequent

as transitions from "a" to "b", transitions from "b" to "a" are prohibited, and transitions from "c" to "a" are three times as frequent as from "c" to "b".

Example 4.1

1. The general 2-state continuous-time Markov chain has transition rate matrix

$$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

with $\alpha, \beta \geq 0$. The process will switch between the two states, with holding times from $\text{Exp}(\alpha)$ in state 0, and from $\text{Exp}(\beta)$ in state 2.

The embedded discrete-time Markov chain has transition probability matrix

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

2. Consider the 3-state continuous-time Markov chain with transition rate matrix

$$Q = \begin{pmatrix} -3 & 1 & 2 \\ 1 & -2 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Holding times in state 0 are from the Exp(3) distribution, and holding times in state 1 are from Exp(2). Once the process transitions to state 3, it never transitions to states 1 and 2: state 3 is an absorbing state.

The transition probability matrix of the embedded discrete-time Markov chain is

$$\widetilde{P} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}$$

4.2.2 Simulating from a continuous-time Markov process

We now know how to simulate from a continuous-time Markov chain, given a transition rate matrix. At each iteration, we generate waiting times for each possible transition from exponential distributions, and the next state is the one with the shortest waiting time.

The following code simulates from a continuous-time Markov process with state space $\{0, 1, 2\}$

and transition rate matrix

$$Q = \begin{pmatrix} -3 & 1 & 2 \\ 0.5 & -1 & 0.5 \\ 0.5 & 1 & -1.5 \end{pmatrix}$$

We also need a way to choose the initial state, and here we specify the initial distribution as $(0, 1, 0)$, i.e., such that $X_0 = 1$. The algorithm runs until it reaches $t = 10$.

```
# Random seed for reproducibility
set.seed(46012)

# Setup parameters
tmax <- 10
u <- c(0, 1, 0)
Q <- matrix(c(-3, 1, 2,
              0.5, -1, 0.5,
              0.5, 1, -1.5),
            nrow = 3, byrow = TRUE)

# Initialise process
X <- sample(0:2, size = 1, prob = u)
times <- 0

# Loop until reaching tmax
while(times[length(times)] < tmax) {
    # Get relevant row of Q
    current_state <- X[length(X)]
    rates <- Q[current_state + 1, -(current_state + 1)]

    # Simulate waiting times and choose shorterst one
    waiting_times <- rexp(2, rate = rates)
    choice <- which.min(waiting_times)

    # Save time of next jump
    new_time <- times[length(times)] + waiting_times[choice]
    times <- c(times, new_time)

    # Save next state
    new_state <- (0:2)[-(current_state + 1)][choice]
    X <- c(X, new_state)
}
```

```
times
```

```
 [1]  0.0000  0.0656  0.5156  1.6927  3.3044  3.3405  4.1567  4.1841  5.1422
[10]  6.9326  7.7817  8.3152  8.3434  8.7406  8.8888  9.1658  9.6722  9.9540
[19] 10.3034
```

```
X
```

```
 [1] 1 0 2 1 0 1 0 2 1 2 0 1 2 0 2 1 0 2 0
```

The output is a sequence of transition times, and the states to which the process jumps. This is all we need to know the value of the process at any time $t \in [0, 10]$.

4.2.3 Explosive Markov chains

Definition 4.3

A continuous-time Markov chain is called **explosive** if an infinite number of transitions can happen in a finite amount of time.

To illustrate this concept, consider the process defined over $\mathcal{S} = \mathbb{N}$, with initial distribution $(1, 0, 0, \dots)$, and transition rate matrix

$$Q = \begin{pmatrix} -1 & 1 & \cdot & \cdot & \cdot \\ \cdot & -2 & 2 & \cdot & \cdot \\ \cdot & \cdot & -4 & 4 & \cdot \\ \cdot & \cdot & \cdot & \ddots & \ddots \end{pmatrix}$$

The chain starts in state 0, and then:

- it switches to 1 after a holding time from Exp(1);
- it switches to 2 after a holding time from Exp(2);
- it switches to 3 after a holding time from Exp(4);

and so on.

The holding times will be shorter and shorter, in such a way that an infinite number of transitions can occur in a finite amount of time. To see this mathematically, denote as T_n the n^{th}

holding time, and S_n the time of the n^{th} transition (i.e., $S_n = T_1 + T_2 + \cdots + T_n$). Then,

$$\begin{aligned} E[S_n] &= E\left[\sum_{k=1}^{n} T_k\right] \\ &= \sum_{k=1}^{n} E[T_k] \\ &= \sum_{k=0}^{n-1} \frac{1}{2^k} \end{aligned}$$

using the property that the expectation of an exponential random variable is the inverse of its rate. But $\lim_{n\to\infty} E[S_n] = 2$ is finite, so an arbitrarily large number of transitions are expected to have happened by $t = 2$.

In the following, we assume that the Markov processes are non-explosive. This is always the case for processes with finite state spaces; in the infinite case, we can ensure that $\sup_i\{q_i\} < \infty$, i.e., the transition rates are bounded by a finite number.

4.3 Transient behaviour

4.3.1 Transition probabilities

Definition 4.4

Let (X_t) be a homogeneous continuous-time Markov process. The **transition probability** from state i to state j over a time interval of length $t \geq 0$ is

$$P_{ij}(t) = \Pr(X_{s+t} = j \mid X_s = i),$$

i.e., it is the probability that the process will be in state j after t time units, given that it started in state i.

For a given time interval t, the transition probability matrix is

$$P(t) = \begin{pmatrix} P_{00}(t) & P_{01}(t) & P_{02}(t) & \cdots \\ P_{10}(t) & P_{11}(t) & P_{12}(t) & \cdots \\ P_{20}(t) & P_{21}(t) & P_{22}(t) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

For a continuous-time Markov process, the transition probabilities can only be defined with respect to some chosen time interval, as there is no predefined time grid. Each transition probability is therefore a function of time. These are *not* the same as the transition probabilities of the embedded discrete-time Markov chain, and we might use the phrase "transition probabil-

ity function" to make the distinction. We will see that transition probability function can be evaluated for any time t from the transition rates.

Remark: for any $i \neq j$, we have $P_{ij}(0) = 0$ and $P_{ii}(0) = 1$, so $P(0) = I$.

Example 4.2

Consider the 3-state continuous-time Markov chain with transition rate matrix

$$Q = \begin{pmatrix} -1 & 0.5 & 0.5 \\ 1 & -3 & 2 \\ 0.5 & 1.5 & -2 \end{pmatrix}$$

The transition probability matrix of the embedded Markov chain is

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{4} & \frac{3}{4} & 0 \end{pmatrix}$$

Let's think about the interpretation of a particular transition probability function over some given time interval, say $P_{01}(8.3)$. It measures the probability that the process starts in state 0 and ends up in state 1 (after 8.3 time units), accounting for all possible combinations and timings of transitions in between. Maybe the process jumped directly from 0 to 1, or maybe it also spent some time in state 2 during the interval. Maybe the last transition to state 1 occurred at $t = 8$, or maybe it occurred at $t = 6.8$. As you can see (and in contrast with the discrete-time setting), the probability $P_{01}(8.3)$ has to account for an infinite number of possible sequence of events.

Proposition 4.2 (Chapman-Kolmogorov equation)

If $P(t)$ denotes the transition probability matrix of a continuous-time Markov process over a time interval of length t, then we have

$$P(s + t) = P(s)P(t)$$

Proof

The proof is identical to the discrete-time case:

$$\begin{aligned}
P_{ij}(s+t) &= \Pr(X_{s+t} = j \mid X_0 = i) \\
&= \sum_{k \in \mathcal{S}} \Pr(X_{s+t} = j, X_s = k \mid X_0 = i) && \text{(a)} \\
&= \sum_{k \in \mathcal{S}} \Pr(X_{s+t} = j \mid X_s = k, X_0 = i) \Pr(X_s = k \mid X_0 = i) && \text{(b)} \\
&= \sum_{k \in \mathcal{S}} \Pr(X_{s+t} = j \mid X_s = k) \Pr(X_s = k \mid X_0 = i) && \text{(c)} \\
&= \sum_{k \in \mathcal{S}} P_{kj}(t) P_{ik}(s) \\
&= [P(s)P(t)]_{ij}
\end{aligned}$$

where (a) is the law of total probability, (b) is the definition of conditional probability, and (c) is the Markov property.

Because the transition probabilities are continuous functions of time, we can study them using tools from analysis. In fact, we can derive differential equations to describe the dynamics of the distribution of the chain through time. But first, we need the following result, which links each transition probability function to a transition rate.

Proposition 4.3

The transition probability functions and transition rates satisfy

$$\lim_{h \to 0} \frac{P_{ij}(h)}{h} = q_{ij} \quad \text{and} \quad \lim_{h \to 0} \frac{1 - P_{ii}(h)}{h} = q_i.$$

Proof

We will only prove the first part of the proposition, but the second part follows a similar reasoning.

We consider the probability of jumping from i to j over a short time interval of length h. This requires two independent events: (1) that the holding time in state i is shorter than h, and (2) that, once the process jumps, it jumps to state j. Things are a little more complex because there could be more than one jump, but the probability of that is $o(h)$ (i.e., very small for short time intervals).

Let D_i be the holding time in state i, and $\widetilde{P}_{ij}$ the transition probability from i to j conditional on a jump (as defined in Section 4.2.1). We can rewrite the transition probability

function as

$$\begin{aligned} P_{ij}(h) &= \Pr(X_h = j \mid X_0 = i) \\ &= \Pr(D_i \leq h)\widetilde{P}_{ij} + o(h) \\ &= (1 - e^{-q_i h})\frac{q_{ij}}{q_i} + o(h) && \text{(a)} \\ &= (1 - [1 - q_i h + o(h)])\frac{q_{ij}}{q_i} + o(h) && \text{(b)} \\ &= q_{ij}h + o(h), && \text{(c)} \end{aligned}$$

where (a) uses the cumulative distribution function of the exponential distribution, (b) uses the Taylor expansion of the exponential function, and (c) uses the fact that $k \times o(h) = o(h)$ for any constant k.

Because $o(h)/h$ tends to zero as $h \to 0$, we can now compute the required limit as

$$\begin{aligned} \lim_{h\to 0} \frac{P_{ij}(h)}{h} &= \lim_{h\to 0} \left\{ \frac{q_{ij}h + o(h)}{h} \right\} \\ &= q_{ij}. \end{aligned}$$

Note that this proof provides an alternative definition of the transition rates, through the relationship

$$P_{ij}(h) = q_{ij}h + o(h).$$

Some books present continuous-time Markov chains first using the transition probability functions, and then define the rates using this formula. This is similar to the second ("little-o") definition of Poisson processes that we saw in Chapter 3, and it says that, over a short time interval h, the probability of jumping from state i to state j is approximately proportional to the transition rate q_{ij}.

Proposition 4.4 (Kolmogorov equations)

If we denote as P' the matrix with elements $P'_{ij} = dP_{ij}/dt$, we have the two following relationships.

Forward equation:

$$P'(t) = P(t)Q$$

Backward equation:

$$P'(t) = QP(t)$$

Proof

To prove the forward equation, we look at the range of change of the transition probability function over a short time interval.

$$\begin{aligned}
\frac{P_{ij}(t+h) - P_{ij}(t)}{h} &= \frac{1}{h}\left\{\sum_{k\in\mathcal{S}} P_{ik}(t)P_{kj}(h) - P_{ij}(t)\right\} && \text{(a)}\\
&= \frac{1}{h}\left\{P_{ij}(t)P_{jj}(h) + \sum_{k\neq j} P_{ik}(t)P_{kj}(h) - P_{ij}(t)\right\} && \text{(b)}\\
&= \frac{1}{h}\left\{P_{ij}(t)[P_{jj}(h) - 1] + \sum_{k\neq j} P_{ik}(t)P_{kj}(h)\right\} && \text{(c)}\\
&= P_{ij}(t)\frac{P_{jj}(h) - 1}{h} + \sum_{k\neq j} P_{ik}(t)\frac{P_{kj}(h)}{h},
\end{aligned}$$

where (a) is the Chapman-Kolmogorov equation, (b) takes the $k = j$ term out of the sum, and (c) factorises the P_{ij} terms.

Taking the limit as $h \to 0$ on both sides, and using the definition $q_{jj} = -q_j$, we find

$$\begin{aligned}
P'_{ij}(t) &= -q_j P_{jj}(t) + \sum_{k\neq j} q_{kj}P_{ik}(t)\\
&= \sum_{k\in\mathcal{S}} q_{kj}P_{ik}(t)\\
&= [P(t)Q]_{ij}.
\end{aligned}$$

The proof of the backward equation is almost identical, except it starts from $P_{ij}(t+h) = \sum_{k\in\mathcal{S}} P_{ik}(h)P_{kj}(t)$.

The Kolmogorov equations are differential equations with a familiar form; the scalar analogue is $f'(t) = qf(t)$. Like in the scalar case, only one function satisfies this equation: the exponential. This gives us a convenient relationship between the transition probability functions and the generator matrix of a continuous-time Markov chain.

Proposition 4.5

Consider a continuous-time Markov chain with transition function $P(t)$ and generator matrix Q. We have

$$P(t) = \exp(tQ).$$

In other words, by definition of the matrix exponential,

$$P(t) = \sum_{n=0}^{\infty} \frac{1}{n!}(tQ)^n = I + tQ + \frac{t^2}{2}Q^2 + \frac{t^3}{6}Q^3 + \cdots,$$

which is, in general, *not* the same as taking the exponential of each element in tQ.

This is a very important result, as it gives a direct way to compute the transition probabilities of a continuous-time process over any time interval, in terms of the transition rate matrix. As required, this formula accounts for all possible sequence of events during that interval.

Computing matrix exponentials with high accuracy is a difficult general problem in numerical analysis, but we won't worry about it here. There are many efficient algorithms, e.g., implemented in the R function `expm()` (from the eponymous package), and those work fine for our purposes.

Example 4.3

We return to the 3-state continuous-time Markov chain from a previous example, with transition rate matrix

$$Q = \begin{pmatrix} -1 & 0.5 & 0.5 \\ 1 & -3 & 2 \\ 0.5 & 1.5 & -2 \end{pmatrix}$$

Using R, compute $P(1)$, $P(2)$, and $P(5)$.

```
# Load library for matrix exponential
library(expm)

# Define transition rate matrix
Q <- matrix(c(-1, 0.5, 0.5,
              1, -3, 2,
              0.5, 1.5, -2),
            nrow = 3, byrow = TRUE)

# Compute transition probability matrices
expm(1 * Q)
```

```
      [,1]  [,2]  [,3]
[1,] 0.522 0.203 0.275
[2,] 0.357 0.270 0.374
[3,] 0.324 0.268 0.408
```

```
expm(2 * Q)
```

```
      [,1]  [,2]  [,3]
[1,] 0.434 0.234 0.332
[2,] 0.403 0.245 0.351
[3,] 0.397 0.247 0.356
```

```
expm(5 * Q)
```

```
      [,1]  [,2]  [,3]
[1,] 0.414 0.241 0.345
[2,] 0.414 0.241 0.345
[3,] 0.414 0.241 0.345
```

We can make a few observations:

- as expected, the rows of the transition probability matrices sum to 1;
- similarly to the discrete-time case, the transition probabilities over long time intervals seem to converge to some distribution.

4.3.2 Marginal distribution

The marginal distribution of X_t is the probability distribution $u(t) = (u_1(t), u_2(t), \dots)$ defined by

$$u_j(t) = \Pr(X_t = j), \quad \text{for all } j \in \mathcal{S}.$$

Like in discrete time, we can use the law of total probability to rewrite this as

$$\begin{aligned} u_j(t) &= \sum_{i \in \mathcal{S}} \Pr(X_t = j \mid X_0 = i) \Pr(X_0 = i) \\ &= \sum_{i \in \mathcal{S}} P_{ij}(t) u_i(0) \\ &= [u(0)P(t)]_j \end{aligned}$$

Finally, given the initial distribution $u(0)$ and the transition rate matrix Q, the distribution of X_t can then be computed as

$$\begin{aligned} u(t) &= u(0)P(t) \\ &= u(0) \exp(tQ). \end{aligned}$$

4.4 Long-term behaviour

Like for their discrete-time counterparts, we are often interested in the long-term properties of continuous-time Markov chains, and in particular the convergence of the distribution of the process to some limit. Many discrete-time results have a continuous-time version, and we go over them more briefly in this chapter.

Definition 4.5

Consider a continuous-time Markov process (X_t) with transition probability function $P(t)$. The probability distribution π is a **stationary distribution** of (X_t) if, for all $t \geq 0$,

$$\pi = \pi P(t)$$

It is usually more useful to rewrite this definition in terms of the transition rates, rather than the transition probabilities.

Proposition 4.6

The distribution π is a stationary distribution of the Markov chain with generator matrix if and only if

$$\pi Q = 0.$$

Proof

To prove the equivalence, we prove each implication separately: (1) $\pi Q = 0 \Rightarrow \pi P(t) = \pi$, and (2) $\pi P(t) = \pi \Rightarrow \pi Q = 0$.

1. Assume that there exists a distribution π such that $\pi Q = 0$. If we take Kolmogorov's backward equation, and multiply each side by π, we get

$$\begin{aligned} & P'(t) = QP(t) \\ \Rightarrow\ & \pi P'(t) = \pi QP(t) \\ \Rightarrow\ & \pi P'(t) = 0, \end{aligned}$$

   where the last step uses the assumption that $\pi Q = 0$. So, we have

$$\frac{d}{dt}\pi P(t) = 0,$$

   i.e., $\pi P(t)$ is constant with respect to t. But we also know that $P(0) = I$, so $\pi P(t) = \pi P(0) = \pi$ for all $t \geq 0$, as required.

2. Assume that there exists a distribution π such that $\pi P(t) = \pi$ for all $t \geq 0$. So, for any $h > 0$, we have

$$\begin{aligned} & \pi P(h) = \pi \\ \Rightarrow\ & \pi(P(h) - I) = 0 \\ \Rightarrow\ & \pi\left(\frac{P(h) - I}{h}\right) = 0 \\ \Rightarrow\ & \pi\left(\frac{P(h) - P(0)}{h}\right) = 0 \end{aligned}$$

Taking the limit as $h \to 0$, we find

$$\begin{aligned} & \pi P'(0) = 0 \\ \Rightarrow\ & \pi P(0)Q = 0 \quad \text{(a)} \\ \Rightarrow\ & \pi Q = 0, \qquad \text{(b)} \end{aligned}$$

where (a) uses the Kolmogorov forward equation, and (b) follows from the assumption that $\pi P(t) = \pi$.

Finally, we have shown the equivalence.

This last result gives us a practical method to find the stationary distribution of a continuous-time Markov process based on its transition rate matrix, by solving a system of linear equations.

Now that we have defined what a stationary distribution distribution is, we turn to its connection to the long-term behaviour of the chain in the following two theorems. The properties of communication, irreducibility, transience and recurrence are defined in the same way as for discrete-time chains; and note that periodicity does not exist in continuous time.

Theorem 4.1

Let (X_t) be a finite, irreducible continuous-time Markov chain with transition function $P(t)$. Then, there exists a unique stationary distribution π, which is the limiting distribution. That is, for all $i, j \in \mathcal{S}$,

$$\lim_{t\to\infty} P_{ij}(t) = \pi_j.$$

This limit theorem explains the phenomenon observed in a previous example that each row of $P(t)$ seems to converge to the same distribution as $t \to \infty$. It offers an alternative, pragmatic method to compute the stationary/limiting distribution of an irreducible continuous-time Markov process: compute $\exp(tQ)$ for some large t.

Theorem 4.2

Let (X_t) be an irreducible, positive recurrent continuous-time Markov chain with unique stationary distribution π. Then, for any $i \in \mathcal{S}$, the long-run proportions are

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{I}_{\{X_s = i\}}\, ds = \pi_i.$$

So, like in the discrete-time case, the stationary distribution gives the long-run proportion of time spent in each state.

Example 4.4

We use continuous-time Markov chains to model the behavioural states of an animal (e.g., "eating", "resting"). Let's say that pandas have three behavioural states, 0 = "resting", 1 = "eating" , and 2 = "travelling", and that they jump between them at the following rates:

$$Q = \begin{pmatrix} -3 & 2 & 1 \\ 5 & -20 & 15 \\ 6 & 2 & -8 \end{pmatrix}$$

where time is measured in days. Looking at the diagonal elements, we see that, on average, pandas rest for $24/3 = 8$ h, eat for $24/20 = 1.2$ h, and travel for $24/8 = 3$ h before jumping to another activity.

Compute the proportion of time that pandas spend in each behavioural state in the long term. (This is often called the "activity budget" by biologists.)

We can rewrite πQ as the system of equations

$$\begin{cases} -3\pi_0 + 5\pi_1 + 6\pi_2 = 0 & \text{(A)} \\ 2\pi_0 - 20\pi_1 + 2\pi_2 = 0 & \text{(B)} \\ \pi_0 + 15\pi_1 - 8\pi_2 = 0 & \text{(C)} \end{cases}$$

Note that the equations are not linearly independent, because (A) $= -($(B) + (C)$)$, but we can replace one of them by the constraint $\pi_0 + \pi_1 + \pi_2 = 1$. Solving these equations, either by hand or using a computer (e.g., `solve()` in R), we find

$$\begin{cases} \pi_0 = \frac{65}{99} \approx 0.657 \\ \pi_1 = \frac{1}{11} \approx 0.091 \\ \pi_2 = \frac{25}{99} \approx 0.253 \end{cases}$$

That is, pandas spend approximately 66% of their time sleeping, 9% of their time eating, and 25% of their time travelling.

4.5 Some special cases

4.5.1 Birth-death process

A birth-death process is a continuous-time Markov chain with state space $\mathcal{S} = \{0, 1, 2, \dots\}$, which models the number of individuals in a population. As the name suggests, the dynamics of the process is defined by two phenomena, whose rates can depend on the population size: a

birth increases the population by 1, and a death decreases it by 1. We can view this model as a continuous-time Markov chain with generator matrix

$$Q = \begin{pmatrix} -q_0 & q_{01} & 0 & 0 & \cdots \\ q_{10} & -q_1 & q_{12} & 0 & \cdots \\ 0 & q_{21} & -q_2 & q_{23} & \cdots \\ 0 & 0 & q_{32} & -q_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where, for a population of size $n > 0$,

- $q_{n,n+1}$ is the birth rate, i.e., the rate of transitions from n to $n+1$;
- $q_{n,n-1}$ is the death rate, i.e., the rate of transitions from n to $n-1$;
- $q_n = q_{n,n-1} + q_{n,n-1}$ is the rate of transitions out of state n.

The embedded discrete-time Markov chain has transition probabilities

$$\widetilde{P} = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots \\ \frac{q_{10}}{q_{10}+q_{12}} & 0 & \frac{q_{12}}{q_{10}+q_{12}} & 0 & \cdots \\ 0 & \frac{q_{21}}{q_{21}+q_{23}} & 0 & \frac{q_{23}}{q_{21}+q_{23}} & \cdots \\ 0 & 0 & \frac{q_{32}}{q_{32}+q_{34}} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

and the waiting time in state n follows an exponential distribution with rate $q_{n,n-1} + q_{n,n-1}$.

Remark: A Poisson process is a birth-death process where the death rate is zero, and where the birth rate is not dependent on the population size ($q_{01} = q_{12} = \cdots = \lambda$). That is, the population can only increase through time, never decrease.

Example 4.5: linear birth-death process

A seemingly reasonable assumption would be that the birth and death rates are both proportional to population size. This gives rise to the birth-death process with transition rates

$$q_{n,n+1} = \lambda n$$
$$q_{n,n-1} = \mu n$$

where $\lambda > 0$ and $\mu > 0$ are the per-individual birth and death rates, respectively. This model is called a linear birth-death model, because the rates are linear in the population size.

How do we expect the population to evolve over time, under this model? When the population is n, the rate of increase is λn and the rate of decrease is μn, so we can think of the overall rate of population change as $(\lambda - \mu)n$. Using the Kolmogorov forward equations, we can show that the expected change is exponential under this model; the

population increases exponentially if $\mu < \lambda$, and it decreases exponentially if $\lambda < \mu$. Note that zero is an absorbing state in this process: if the population size is zero, then it will remain there. Figure 4.3 shows example realisations from linear birth-death processes with different values of the birth and death rates.

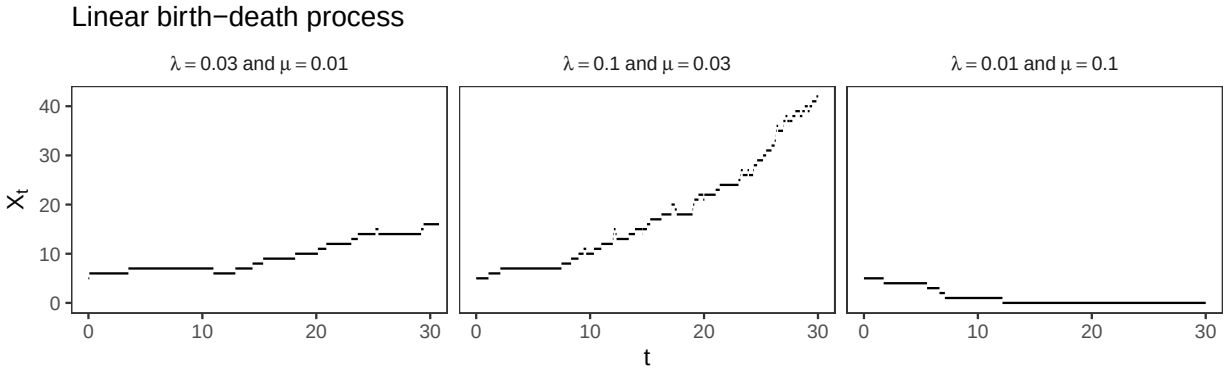


Figure 4.3: *Examples of linear birth-death process.*

4.5.2 Queueing process

The study of queues has many applications, such as customers at a store or a bank, patients in the emergency service of a hospital, or people waiting on hold on the phone. A common model, called the $M/M/k$ model (where the "M"s stand for "Markov" or "memoryless"), assumes that:

1. there are k servers (e.g., doctors at the ER, or tills at the store);
2. arrivals in the queue follow a Poisson process with rate λ;
3. the serving time by each operator follows an exponential distribution with rate μ.

From this model, one can derive the expected number of people waiting in line in the long run (when it exists), or the expected waiting time spent in line, as functions of k, λ and μ. An ER service could for example use this to determine how many doctors they need to ensure that patients never have to wait longer than 3 hours before being treated.

4.6 Continuous state space: Brownian motion

Many well-studied continuous-time Markov processes are defined over an uncountable state space, e.g., $\mathcal{S} = \mathbb{R}$ or $\mathcal{S} = [0, \infty)$. This formulation is useful in situations where the phenomenon of interest is continuous, such as the value of a stock price, or the position of a particle in space. An important class of continuous-time continuous-space Markov processes is diffusion processes, which have been widely used in fields such as physics, biology, and finance. Brownian motion is the building block of all diffusion processes, and we introduce it briefly in this section.

The motivation for Brownian motion was the observation (by 19th century biologist Robert

Brown) that particles of pollen in water follow erratic and seemingly random trajectories. This phenomenon is caused by collisions with water molecules, and it was first described mathematically in the early 20th century by Albert Einstein. The properties of the resulting process were further explored by Norbert Wiener, and Brownian motion is also called the Wiener process.

Definition 4.7

A continuous-time stochastic process (B_t) is a **standard Brownian motion** if it satisfies the following properties.

1. For all $s, t > 0$, $B_{s+t} - B_s$ has a normal distribution with mean 0 and variance t.
2. For any $0 \leq q < r \leq s < t$, the increments $B_t - B_s$ and $B_r - B_q$ are independent random variables.
3. The function $t \mapsto B_t$ is continuous, with probability 1. (*Also sometimes stated as: the sample paths of* (B_t) *are continuous.*)

The additional condition that $B_0 = 0$ is also sometimes included in the definition of Brownian motion. More generally, we can assume that B_0 is specified as part of the model formulation as an initial condition.

The properties of Brownian motion suggest the following method to generate sample paths from the process over some time grid $t_0, t_1, \ldots, t_n$. We start from some intial condition $B_{t_0} = b_0$, and, for $i = 0, \ldots, n-1$,

1. generate a normally distributed increment $\varepsilon_i \sim N(0, t_{i+1} - t_i)$;
2. compute the next value of the process as $B_{t_{i+1}} = B_{t_i} + \varepsilon_i$.

This algorithm can for example be implemented in R using the random number generator `rnorm()`. This is very similar to the procedure used in Chapter 2 to simulate from a (discrete-time) Gaussian random walk, and Brownian motion can be viewed as the continuous-time analogue. Note that, here, we can sample the path over an arbitrarily fine time grid. Figure 4.4 shows three simulated realisations from a standard Brownian motion with initial condition $B_0 = 0$, over $t \in [0, 10]$. The three paths start from 0, and they spread more and more as they fluctuate randomly through time.

Example 4.6: Brown's pollen

The motivation for developing the theory of Brownian motion was the movements of pollen in water. The definition of Brownian motion describes a one-dimensional process, so how can this be used for the movement of pollen in two dimensions? The simplest approach is to assume that the two coordinates follow two independent Brownian motions (this is called "isotropy").

Figure 4.5 shows a path simulated from a two-dimensional isotropic Brownian motion,

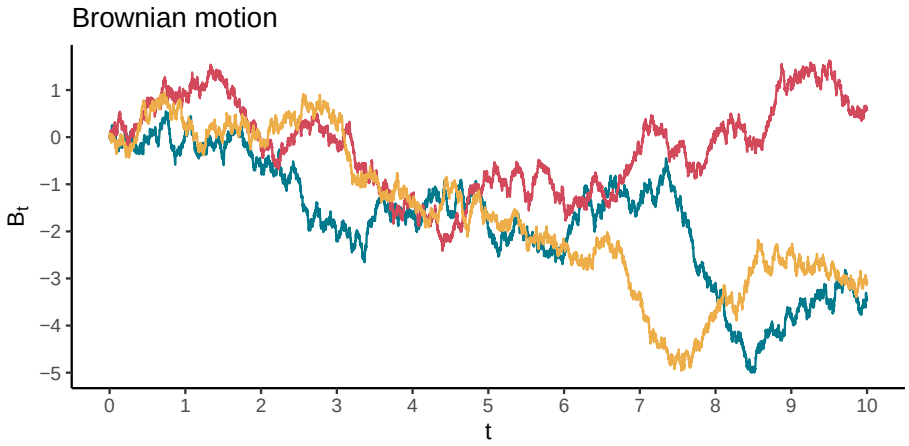


Figure 4.4: *Example realisations from a standard Brownian motion process over the interval* $0 \leq t \leq 10$.

perhaps resembling the pollen trajectories observed by Brown. A more sophisticated model could assume that the x and y coordinates are correlated, which could for example favour movement along a particular direction.

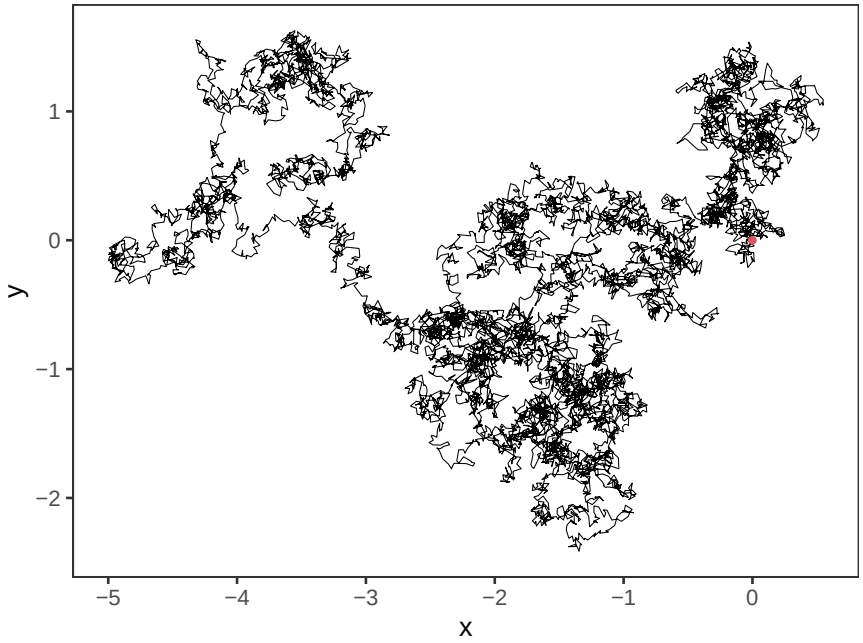


Figure 4.5: *Example realisation from a standard isotropic Brownian motion process in two dimensions. The initial position is shown as a red dot.*

One key feature of Brownian motion is its scaling property: no matter how much we "zoom in", the process is still a Brownian motion.

Proposition 4.7

Let (B_t) be a standard Brownian motion process with initial condition $B_0 = 0$. For any $a > 0$, B_{at} and $\sqrt{a}B_t$ have the same distribution.

Proof

By definition of Brownian motion, $B_{at} \sim N(0, at)$. But, using the properties $E[aX] = aE[X]$ and $Var[aX] = a^2 Var[X]$, we also have

$$B_t \sim N(0, t)$$
$$\Rightarrow \sqrt{a} B_t \sim N(0, at).$$

The rescaling property of Brownian motion is illustrated in Figure 4.6. No matter how much we zoom into a Brownian motion path, the behaviour of the process is the same, in the sense that it has independent, normally distributed increments with variance proportional to the length of the time interval. Brownian motion can be viewed as part of the general mathematical family of fractals.

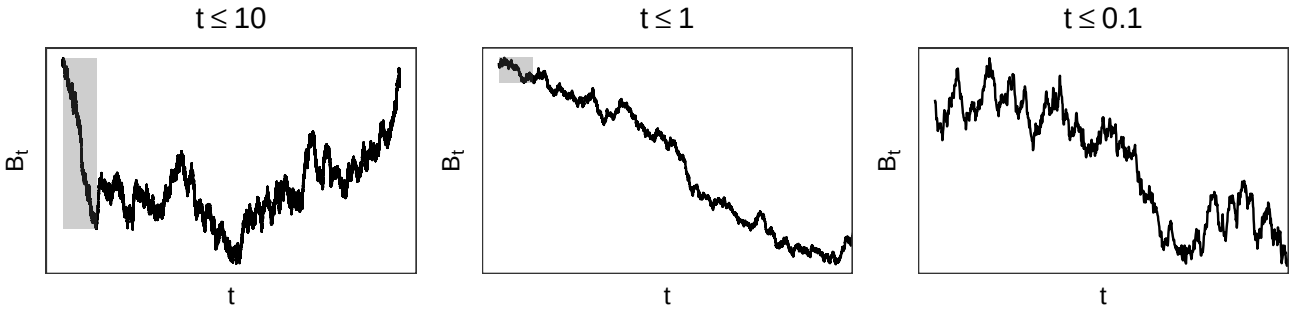


Figure 4.6: *Simulated Brownian motion path over three different time intervals. In the first two panels, the grey boxes show the zoomed area of the next panel.*

It turns out that the rescaling property of Brownian motion implies that its paths are nowhere differentiable, even if it is continuous everywhere. We will not prove this result, but it relies on the following intuition. The derivative of the process can be defined as

$$\frac{d}{dt} B_t = \lim_{h \to 0} \frac{B_{t+h} - B_t}{h}.$$

By definition of Brownian motion, $B_{t+h} - B_t$ has a normal distribution with mean 0 and variance h. So, $(B_{t+h} - B_t)/h$ is also normally distributed with mean 0 and, using $Var[aX] = a^2 Var[X]$, we have

$$Var\left[\frac{B_{t+h} - B_t}{h}\right] = \frac{1}{h^2} Var[B_{t+h} - B_t] = \frac{1}{h}.$$

As $h \to 0$, the variance tends to ∞, so the limit is not well defined and the derivative does not exist.

5 Hidden Markov models

Markov processes are a convenient approach to model temporal dependence while retaining (some) mathematical simplicity but, to estimate the Markov process from data, we need to observe that process directly. However, there are many situations where the process is observed only indirectly, i.e., our observations depend on something that can be described by a Markov process. There is a vast literature on models for such situations; they are called state-space models when the state space of the Markov process is continuous, and hidden Markov models when it is discrete.

Hidden Markov models were developed more recently than other models covered in this course, as illustrated in Figure 5.1, but they are now widely-used in various areas of applications (finance, medicine, ecology, etc.).

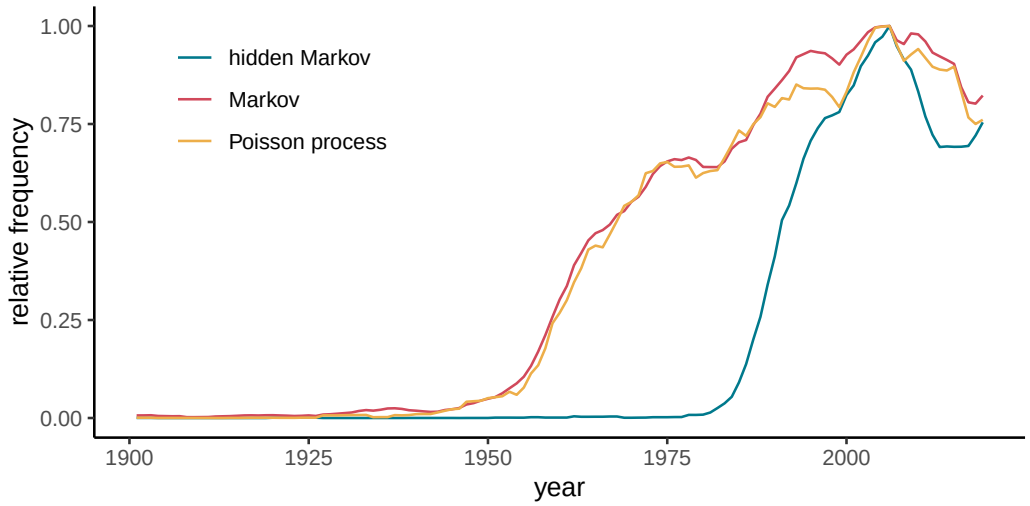


Figure 5.1: *Relative frequency of model names according to the Google Ngram viewer.*

5.1 Mixture models

We start with the description of mixture models, on which hidden Markov models build. A mixture model describes a random variable which can come from several different distributions, each with some probability.

Consider the random variable Z, which follows one of K distributions, with respective probability density (or mass) functions $b_1, b_2, \ldots, b_K$. For any $k \in \{1, \ldots, K\}$, we further assume that Z follows the k^{th} distribution (b_k) with probability π_k, where $\sum_{k=1}^{K} \pi_k = 1$.

Notation

Throughout this chapter, we will consider random variables that can be either discrete or continuous. Rather than writing every equation twice, we will use generic notation that applies in both cases.

Specifically, we will use $f(Z = z)$ to represent either a probability (if Z is discrete) or a probability density (if Z is continuous).

The probability mass/density function of Z under this model is a linear combination of the component functions, each weighted by the probability of the component:

$$\begin{aligned} f(Z = z) &= \sum_{k=1}^{K} f(Z = z \mid C = k) \times \Pr(C = k) \\ &= \sum_{k=1}^{K} \pi_k b_k(z) \end{aligned}$$

Examples of mixture model with three components is shown in Figure 5.2. For one of them, the b_k are normal probability distribution functions; for the other, they are Poisson probability mass functions. In both cases, the mixture model has much more flexibility than a single distribution from that family.

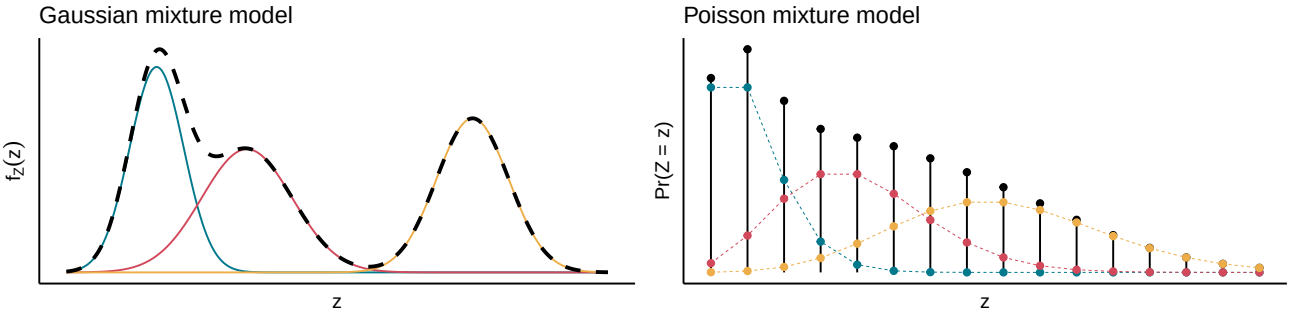


Figure 5.2: *Example mixture models with three components p_1, p_2, p_3, resulting in the probability distribution f_Z shown in black. The component distributions are weighted by the probabilities π_1, π_2, π_3. On the left, each component is a normal distribution; on the right, each component is a Poisson distribution.*

Mixture models have been popular for model-based clustering. Consider n observations $z_1, \ldots, z_n$, assumed to be realisations from n independent random variables described by some mixture model. Various approaches have been developed to estimate parameters of the component distributions and the weight π_i of each component, and to group the observations by "most likely component".

Example 5.1

Consider the distribution of flipper length from 344 penguins shown in Figure 5.3 (from the R package `palmerpenguins`). The distribution is clearly bimodal, and we suspect that data from two different species have been mixed up.

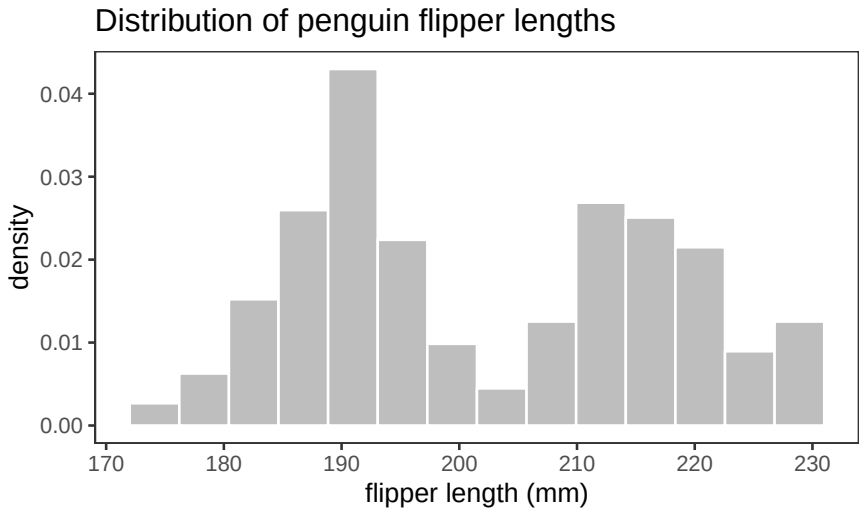


Figure 5.3: *Histogram of flipper lengths in penguin data set.*

A Gaussian mixture model could be used to answer questions such as:

1. Which species does each data point belong to?
2. What is the distribution of flipper lengths for each species?
3. What is the overall distribution of flipper lengths for both species?

Figure 5.4 shows the two distributions that we would obtain from a Gaussian mixture model fitted to these data.

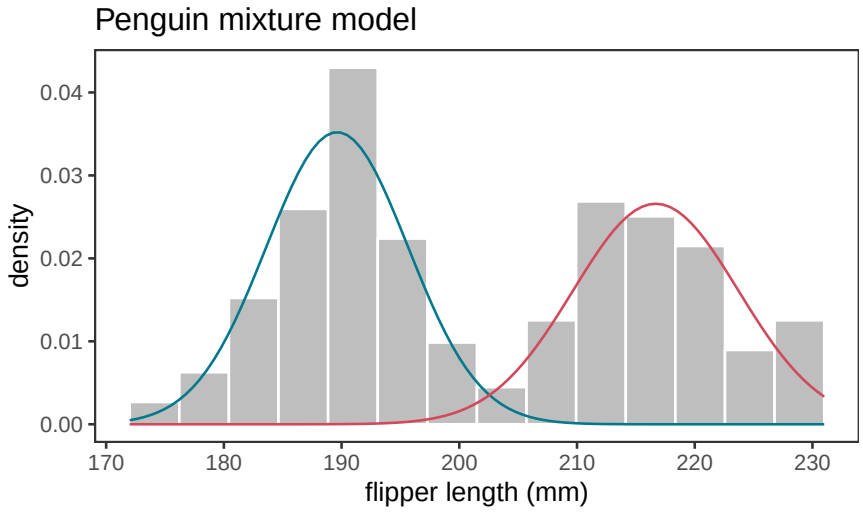


Figure 5.4: *Estimated mixture distributions for penguin flipper data.*

Hidden Markov models can be viewed as dependent mixture models, i.e., where successive observations are not independent.

5.2 Hidden Markov models

5.2.1 Definition

A hidden Markov model (HMM) consists of two stochastic processes, a state process (X_n), and a state-dependent observation process (Z_n). Both processes can in principle be continuous-valued, but we will focus on the case where X_n is defined over a countable set $\mathcal{S} \equiv \{0, 1, 2, \dots\}$. Time can also be discrete or continuous, and in this chapter we focus on discrete-time HMMs, with the index m or n. (We sometimes prefer m because n is traditionally used for the size of the data set, as we do when discussing the likelihood derivation below.)

HMMs are characterised by the following dependence assumptions:

1. The state process (X_n) is a Markov chain, such that

$$\Pr(X_{n+1} \mid X_n, X_{n-1}, \dots, X_0) = \Pr(X_{n+1} \mid X_n).$$

2. The observation Z_n is independent of past values of the process, conditional on the current state X_n. That is,

$$f(Z_n \mid Z_{n-1}, \dots, Z_0, X_n, \dots, X_0) = f(Z_n \mid X_n)$$

That is, Z_n comes from a mixture model, where the mixture component active at time n is given by X_n. In most situations, the Markov chain is such that there is persistence in the state (i.e., the process tends to remain in the same state for several time steps), and this creates correlation between successive values of Z_n.

A simulated realisation from an HMM where $X_n \in \{0, 1, 2\}$ and $Z_n \mid X_n = j \sim \text{Pois}(\lambda_j)$ is shown in Figure 5.5. It is clear that the dependence on X_n induces autocorrelation in the observation process (Z_n).

> Terminology: What exactly is "hidden"?
>
> This is called a *hidden* Markov model because, in practice, we usually only have observations from the state-dependent process (Z_n), but no direct observations from the state process (X_n). The state process is therefore "hidden", or "unobserved", or "latent". The problem is then to try to infer the dynamics of the hidden state process, based on the observations. This type of inference is very common, because it is often the case that we cannot directly observe the phenomenon of interest, e.g., because of measurement error.

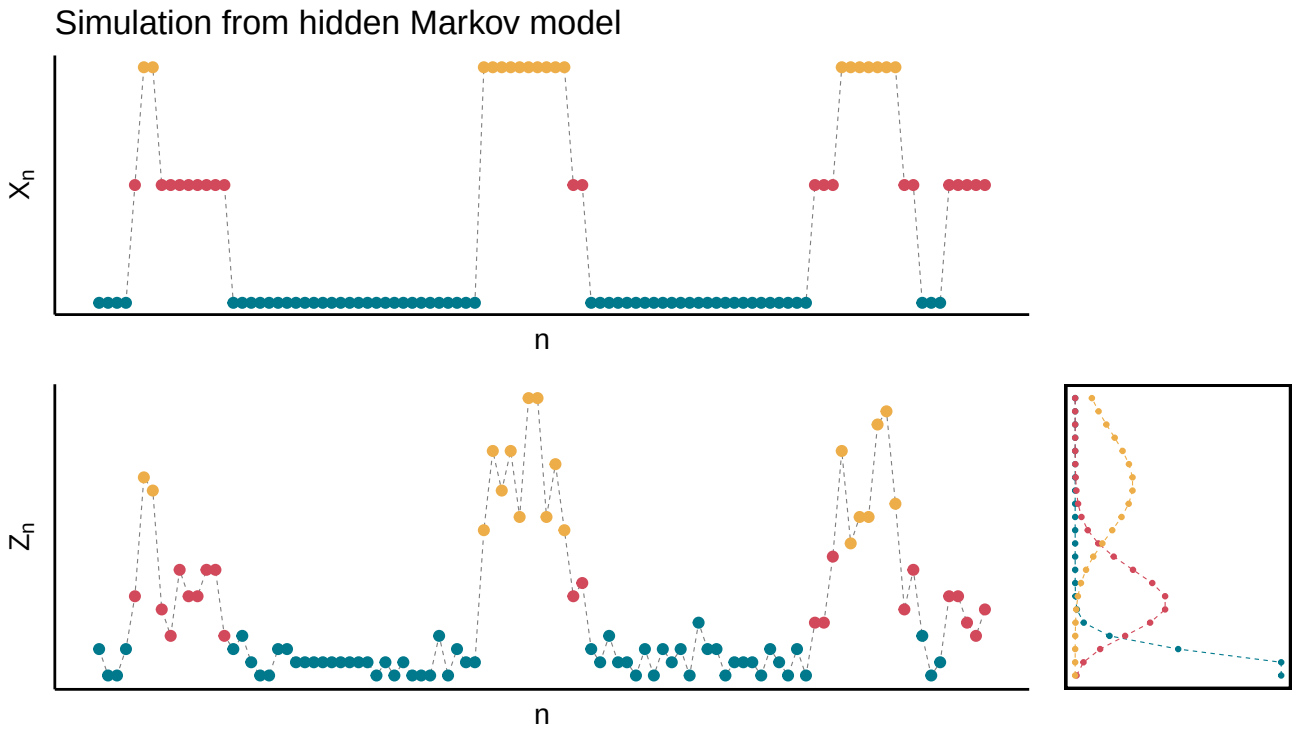


Figure 5.5: *Simulation from a 3-state hidden Markov model with Poisson state-dependent distributions. The top panel shows the simulated state sequence, the bottom-left panel shows the simulated observation sequence, and the bottom-right panel shows the three Poisson distributions.*

5.2.2 Marginal distribution

The model for the observation process (Z_t) is defined through the conditional distribution of Z_t given the state X_t. We denote as b_k the probability density/mass function of Z_t in state k, i.e., $b_k(z) = f(Z_n = k \mid X_n = k)$. Let $B(z)$ be the diagonal matrix with i^{th} diagonal element $b_i(z)$. Like in Chapter 2, let $u^{(n)} = (u_0^{(n)}, u_1^{(n)}, \dots)$ be the probability distribution of X_n, i.e., $u_i^{(n)} = \Pr(X_n = i)$.

We are sometimes interested in the *marginal* distribution of the observation in a hidden Markov model, i.e., $f(Z_n = z)$ (not conditional on the state X_n). By the law of total probability, we have

$$\begin{aligned} f(Z_n = z) &= \sum_{i \in \mathcal{S}} \Pr(X_n = i) f(Z_n = z \mid X_n = i) \\ &= \sum_{i \in \mathcal{S}} u_i^{(n)} b_i(z) \\ &= u^{(n)} B(z) 1^\top, \end{aligned}$$

where 1 is a (row) vector of ones. From Chapter 2, we know that the distribution of the state variable X_n can be written in terms of its initial distribution $u^{(0)}$ and transition probability matrix P, as $u^{(n)} = u^{(0)} P^n$. (This was a consequence of the Chapman-Kolmogorov equations.) Using this result, we find

$$f(Z_n = z) = u^{(0)} P^n B(z) 1^\top.$$

5.2.3 Simulating from a hidden Markov model

We use the dependence structure to define a simulation procedure. The hidden state process is simply a discrete-time Markov chain, so we can simulate it as described in Chapter 2. Once we have a simulated state sequence, we can simulate the observation process at each time step conditionally on the state.

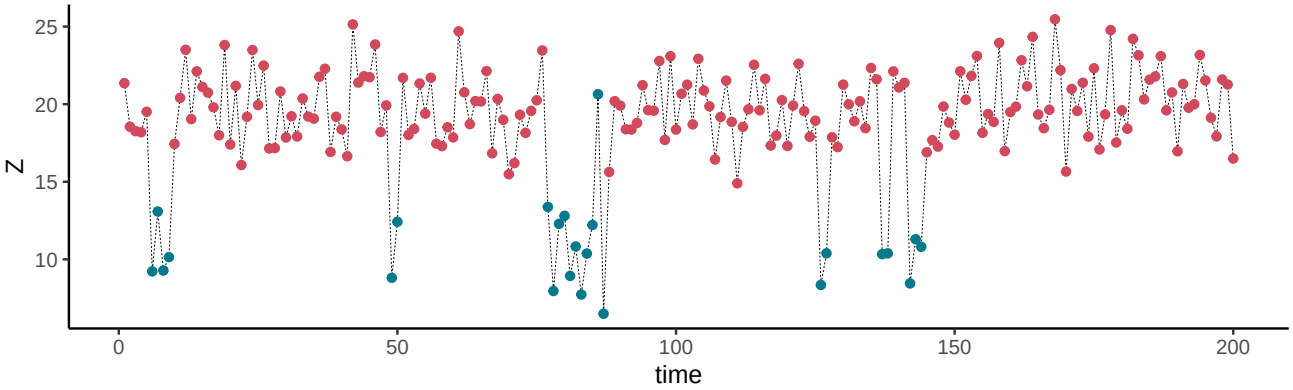
1. Initialise X_0 based on the initial distribution $u^{(0)}$.
2. For $n = 1, 2, ...$, simulate X_n conditionally on $X_{n+1} = i$ using the transition probabilities $\{\gamma_{ij}\}_{j \in \mathcal{S}}$ (i.e., the i^{th} row of the transition probability matrix).
3. For $n = 0, 1, ...$, simulate Z_n conditionally on $X_n = i$ using the observation distribution b_i.

The following code shows an example over 100 time steps, for a 2-state hidden Markov model with normal state-dependent distributions, with model parameters

$$
\begin{aligned}
u^{(0)} &= (0.5, 0.5) \\
P &= \begin{pmatrix} 0.9 & 0.1 \\ 0.05 & 0.95 \end{pmatrix} \\
Z_n \mid X_n = 0 &\sim N(10, 3^2) \\
Z_n \mid X_n = 1 &\sim N(20, 2^2)
\end{aligned}
$$

```
# Set random seed for reproducibility
set.seed(294)

# Define parameters
n <- 200
u <- c(0.5, 0.5)
P <- matrix(c(0.9, 0.1,
              0.05 , 0.95),
            nrow = 2, byrow = TRUE)
mu <- c(10, 20)
sigma <- c(3, 2)

# Simulate state process
X <- rep(NA, length = n)
X[1] <- sample(0:1, size = 1, prob = u)
for(i in 2:n) {
    P_row <- P[X[i-1] + 1,]
    X[i] <- sample(0:1, size = 1, prob = P_row)
```

```r
}

# Simulate observation process
Z <- rep(NA, length = n)
for(i in 1:n) {
    Z[i] <- rnorm(1, mean = mu[X[i] + 1], sd = sigma[X[i] + 1])
}

ggplot(data.frame(time = 1:n, X = X, Z = Z), aes(time, Z)) +
    geom_line(lty = 2, linewidth = 0.1) +
    geom_point(aes(col = factor(X))) +
    scale_color_manual(values = c("#00798c", "#d1495b"), guide = "none")
```



5.3 Likelihood

Hidden Markov models are mostly interesting to applied statisticians and scientists (rather than probabilists), and so most of the related work has been on statistical inference. Given a sequence of observations, the main questions are usually:

- Can we estimate the state-dependent distributions b_k?
- Can we estimate the transition probabilities of the underlying Markov chain?
- Can we infer the most likely value for the state process at each time step?

The starting point is then to look at the likelihood function for this model. This section presents two methods to compute the likelihood. The mathematical derivations might seem a little tedious, but you will notice that we only use basic probability rules, and we take advantage of the dependence structure of the hidden Markov model to find the likelihood.

5.3.1 First attempt

We want to find the joint probability density/mass function of the random variables $Z_0, Z_1, \dots, Z_n$, which we will denote as $L = f(Z_0 = z_0, \dots, Z_n = z_n)$ for convenience. When

viewed as a function of the unknown model parameters, L is the likelihood, and can be maximised numerically for estimation. We will also write $Z_{0:n} = \{Z_0, \dots, Z_n\}$ and similar notation for brevity.

By repeatedly applying the law of total probability, and leveraging the Markov property of (X_n), we see that

$$\begin{aligned}
L &= \sum_{X_n \in \mathcal{S}} f(Z_{0:n} \mid X_n) \Pr(X_n) \\
&= \sum_{X_n \in \mathcal{S}} \sum_{X_{n-1} \in \mathcal{S}} f(Z_{0:n} \mid X_{n-1}, X_n) \Pr(X_n \mid X_{n-1}) \Pr(X_{n-1}) \\
&= \sum_{X_n \in \mathcal{S}} \sum_{X_{n-1} \in \mathcal{S}} \sum_{X_{n-2} \in \mathcal{S}} f(Z_{:n} \mid X_{n-2}, X_{n-1}, X_n) \\
&\qquad \times \Pr(X_n \mid X_{n-1}, X_{n-2}) \Pr(X_{n-1} \mid X_{n-2}) \Pr(X_{n-2}) \\
&= \sum_{X_n \in \mathcal{S}} \sum_{X_{n-1} \in \mathcal{S}} \sum_{X_{n-2} \in \mathcal{S}} f(Z_{0:n} \mid X_{n-2}, X_{n-1}, X_n) \\
&\qquad \times \Pr(X_n \mid X_{n-1}) \Pr(X_{n-1} \mid X_{n-2}) \Pr(X_{n-2}) \\
&= \dots \\
&= \sum_{X_0 \in \mathcal{S}} \dots \sum_{X_n \in \mathcal{S}} \left\{ f(Z_{0:n}, \mid X_{0:n}) \times \Pr(X_0) \times \prod_{m=1}^{n} \Pr(X_m \mid X_{m-1}) \right\}
\end{aligned}$$

(Above, we are using a slight abuse of notation, so that the formula fits on a page. Take some time to think about it, and make sure you understand what the probabilities and f refer to.)

Now, remember that the observations are conditionally independent given the states, so

$$f(Z_{0:n}, \mid X_{0:n}) = \prod_{m=0}^{n} f(Z_m \mid X_m)$$

We now recognise that the likelihood can be written in terms of the state-dependent distributions, the initial distribution of the state process, and the transition probabilities. As before, we use the notation

- $b_k(z) = f(Z_m = z \mid X_m = k)$,
- $P_{ij} = \Pr(X_{m+1} = j \mid X_m = i)$,
- $u_i^{(m)} = \Pr(X_m = i)$.

Then, the joint density of the observations is

$$L = \sum_{x_0 \in \mathcal{S}} \dots \sum_{x_n \in \mathcal{S}} \left\{ u_{x_0}^{(0)} \prod_{m=0}^{n} b_{x_m}(z_m) \prod_{m=1}^{n} P_{x_{m-1}, x_m} \right\}$$

This is a relatively simple expression, which would in principle be straightforward to implement

with a computer (for maximum likelihood estimation, for example). However, it is extremely computationally expensive, with $|\mathcal{S}|^{n+1}$ terms to sum, and often not a practical option. The challenge is to sum over all unobserved state sequences, because there are so many possible combinations. In the next section, we present an alternative approach to evaluating the likelihood, which uses matrix operations and offers an elegant solution to this problem.

5.3.2 Second attempt: forward algorithm

Definition 5.1

The **forward probability** $\alpha_k^{(m)}$ is defined as

$$\alpha_k^{(m)} = f(Z_0 = z_0, \dots, Z_m = z_m, X_m = k),$$

for state $k \in \mathcal{S}$ and time $m \in \{0, 1, 2, \dots\}$.

When the observation variables are continuous, $\alpha_k^{(m)}$ represents a probability density rather than a probability, but the term "forward probability" tends to be used loosely in both cases. In what follows, we denote as $\alpha^{(m)} = (\alpha_0^{(m)}, \alpha_1^{(m)}, \dots)$ the vector of forward probabilities.

There is a close link between the forward probabilities and the likelihood. Once again, we use the law of total probability to notice that

$$\begin{aligned} L &= \sum_{k \in \mathcal{S}} f(Z_0 = z_0, \dots, Z_n = z_n, X_n = k) \\ &= \sum_{k \in \mathcal{S}} \alpha_k^{(n)} \\ &= \alpha^{(n)} 1^\top \end{aligned}$$

Proposition 5.1

The likelihood of a hidden Markov model is given by

$$L = u^{(0)} B(z_0) P B(z_1) \dots P B(z_n) 1^\top$$

Proof

Our aim is to derive an iterative procedure to compute the forward probabilities $\alpha^{(n)}$. First, note that

$$f(Z_{0:m}, X_{0:m}) = f(X_0) \prod_{l=0}^{m} f(Z_l \mid X_l) \prod_{l=1}^{m} \Pr(X_l \mid X_{l-1})$$

and

$$f(Z_{0:m+1}, X_{0:m+1}) = f(X_0) \prod_{l=0}^{m+1} f(Z_l \mid X_l) \prod_{l=1}^{m+1} \Pr(X_l \mid X_{l-1})$$
$$= f(Z_{0:m}, X_{0:m}) f(Z_{m+1} \mid X_{m+1}) \Pr(X_{m+1} \mid Z_{m+1}).$$

Summing over all possible values of $X_0, X_1, \ldots, X_{m-1}$ (as we do in the law of total probability), this becomes

$$f(Z_{0:m+1}, X_m, X_{m+1}) = f(Z_{0:m}, X_m) \Pr(X_{m+1} \mid X_m) f(Z_{m+1} \mid X_{m+1}),$$

and, summing over all possible values of X_m, we get

$$f(Z_{0:m+1}, X_{m+1}) = \sum_{k \in \mathcal{S}} f(Z_{0:m}, X_m = k) \Pr(X_{m+1} \mid X_m) f(Z_{m+1} \mid X_{m+1}).$$

Recognising the forward probabilities, this can be written as the matrix product

$$\alpha^{(m+1)} = \alpha^{(m)} PB(z_{m+1}).$$

We now have a method to iteratively compute the forward probabilities; this procedure is called the forward algorithm. This initially requires calculating $\alpha^{(0)}$, which is simply $u^{(0)} B(z_1)$.
Applying this iteration from $m = 0$ to $m = n - 1$ yields

$$\alpha^{(n)} = u^{(0)} B(z_0) PB(z_1) \ldots PB(z_n)$$

Finally, we get the required result,

$$L = \alpha^{(n)} 1^\top$$
$$= u^{(0)} B(z_0) PB(z_1) \ldots PB(z_n) 1^\top$$

Remarkably, the number of operations required to evaluate the matrix product is much smaller than for the original nested sums. It is of the order of $n|\mathcal{S}|^2$, and makes it possible to compute the likelihood in many situations. This efficient algorithm has greatly contributed to the popularity of hidden Markov models.

A similar algorithm (called the "Viterbi algorithm", after its inventor) can be used to derive the most likely state sequence, given the observations and a set of estimated parameters. In many applications, this is the main object of inference.

5.4 Some examples

5.4.1 Animal telemetry

In the last couple of decades, it has become increasingly possible to track wild animals using telemetry devices. This could be a GPS collar on a polar bear, a depth sensor on a beaked whale, or an accelerometer on an albatross. The resulting data contain an incredible amount of information about the behaviour of those animals, which would be very difficult to obtain otherwise.

Consider the acceleration data shown in Figure 5.6, which comes from an albatross tagged in South Georgia (a small island in the South Atlantic) and was analysed by Conners et al. (2021). The variable shown here is a derived metric of "heave acceleration", i.e., acceleration along the bird's up-down axis, measured every 30 seconds. It is clear that the distribution of acceleration is multimodal, and it looks from the time series plot that there is strong autocorrelation: high acceleration is likely to be followed by high acceleration. The multimodality suggests that a mixture model might be adequate, and the autocorrelation suggests that some dependence is required, making hidden Markov models a natural choice.

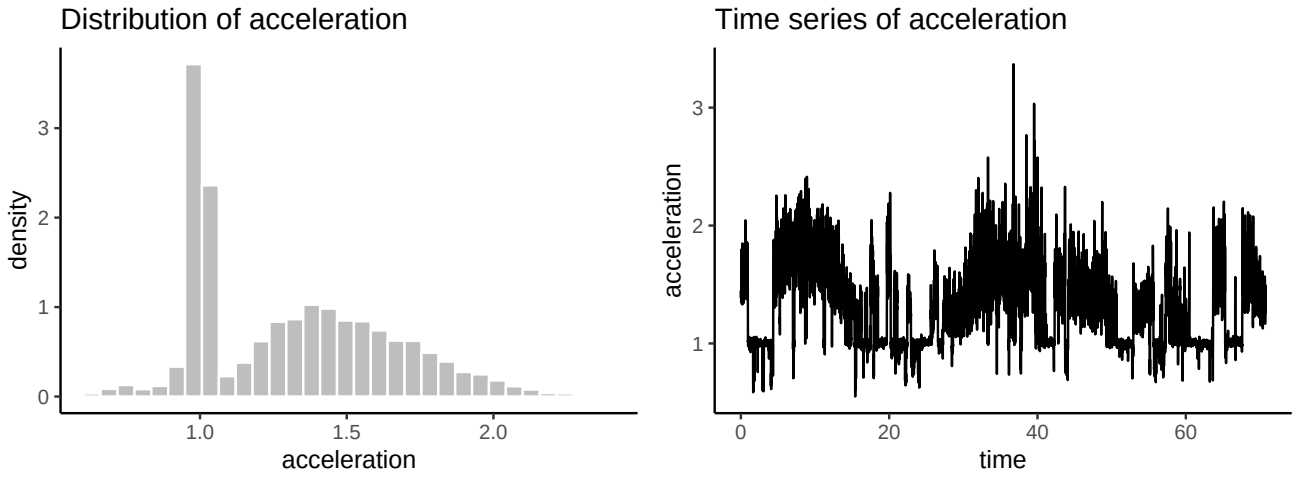


Figure 5.6: *Albatross accelerometer data.*

Now, say that we use a hidden Markov model with three states, i.e., $X_m \in \mathcal{S} = \{0, 1, 2\}$, to identify three mixture components. Within each state, we choose to model the acceleration with a normal distribution, i.e., b_k is the Gaussian probability density function (with parameters dependent on k). Maximum likelihood estimation based on the forward algorithm can be used to estimate all transition probabilities of (X_t), as well as the state-dependent parameters of the normal distribution of acceleration.

Figure 5.7 shows the estimated state-dependent distribution b_k, and the most likely sequence of unobserved states. In this model, the three states ($X_m = 0, 1, 2$) correspond to very low, low, and high heave acceleration, respectively. A biologist could propose a tentative interpretation in terms of albatross behaviour, and the model could then be used to distinguish phases where the animal is resting on water and flying, for example.

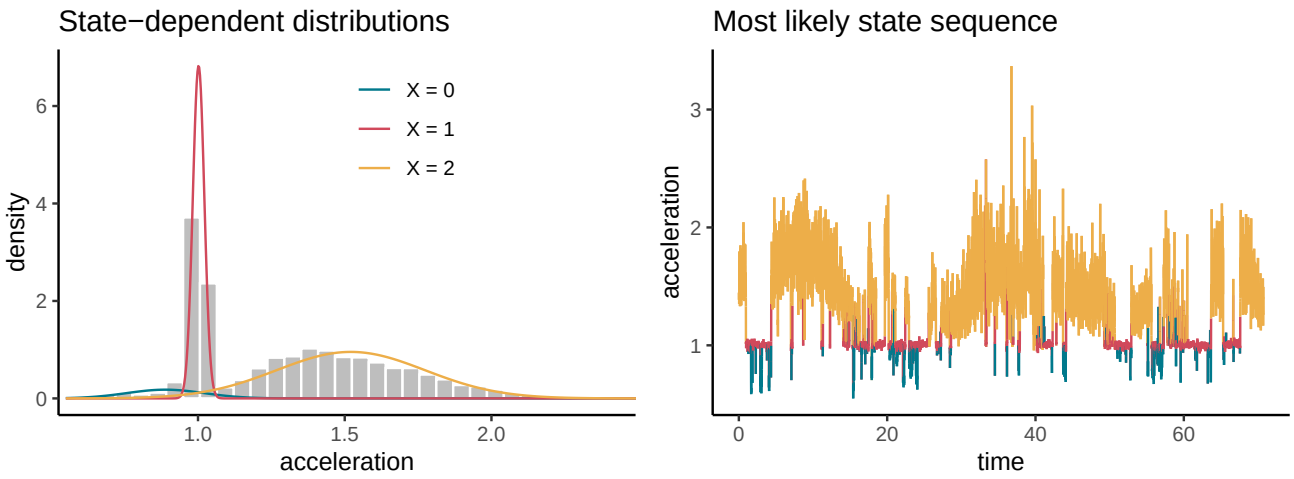


Figure 5.7: *Results of albatross analysis.*

Another output of the model is the estimated transition probability matrix of the state process,

$$\widehat{P} = \begin{pmatrix} 0.783 & 0.21 & 0.007 \\ 0.042 & 0.935 & 0.023 \\ 0 & 0.013 & 0.987 \end{pmatrix}$$

The large diagonal transition probabilities reflect a strong tendency to persist in each state. We can use the results from Chapter 2 to get insights into the behaviour of albatross. For example, from $\widehat{P}$, we can see that the expected holding times in the three states are

$$\frac{1}{1-0.783} = 4.6, \quad \frac{1}{1-0.935} = 15, \quad \text{and } \frac{1}{1-0.987} = 77.$$

We can compute the stationary distribution of the Markov chain using any of the methods from Chapter 2, which gives us an estimate of the long-run proportion of time that the bird spends in each behavioural state. Here, we find

$$\pi = (0.06, 0.33, 0.61).$$

These results all suggest that the albatross spends most of its time in the state $X = 2$.

5.4.2 Oil price

We now turn to the problem of understanding the dynamics of oil prices through time. Figure 5.8 shows the daily changes in oil prices in the USA between 1986 and 2023, obtained from the US Energy Information Administration. The histogram does not display multimodality this time, but the distribution looks heavy-tailed and would not be modelled well with something like a normal distribution. This is another reason to use a mixture model. The time series plot shows an interesting pattern of alternance between long periods of high and low variability. We might interpret high variability as a sign of financial instability and, indeed, there are

large price changes after the 2008 financial crisis, as well as during the first few months of the Covid-19 pandemic.

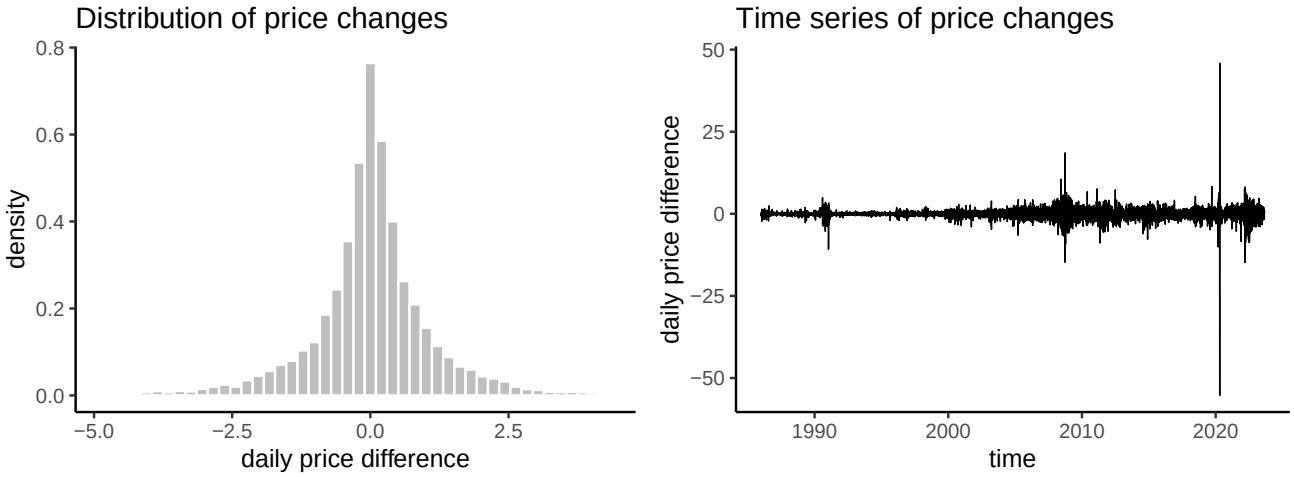


Figure 5.8: *Oil price data.*

Just like in the previous example, we analyse the data with a hidden Markov model with three states and normal state-dependent distributions. Using maximum likelihood estimation, we get estimates of the transition probabilities and of the parameters of the normal distribution in each state.

Figure 5.9 shows the estimated state-dependent distributions, and the most likely state sequence for the fitted hidden Markov model. In contrast with the albatross example, what distinguishes the three states this time is not the mean of the distributions, but their variances. They roughly represent low ($X_m = 0$), intermediate ($X_m = 1$), and high ($X_m = 2$) variances, corresponding to different levels of financial instability.

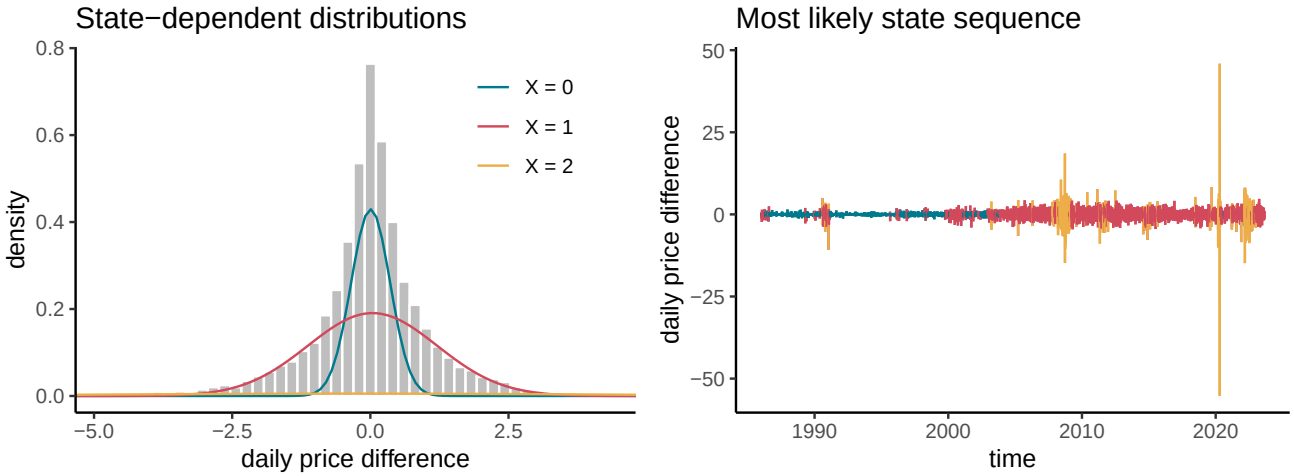


Figure 5.9: *Results of oil price analysis.*

The estimated transition probability matrix is

$$\widehat{P} = \begin{pmatrix} 0.989 & 0.011 & 0 \\ 0.007 & 0.98 & 0.012 \\ 0 & 0.1 & 0.9 \end{pmatrix}$$

indicating that there is strong autocorrelation in the state process. The expected holding times (measured in days) are

$$\frac{1}{1-0.989} = 91, \quad \frac{1}{1-0.98} = 50, \quad \text{and } \frac{1}{1-0.9} = 10,$$

and the stationary distribution is

$$\pi = (0.38, 0.56, 0.07).$$

References

Brin, Sergey, and Lawrence Page. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30 (1-7): 107–17.

Conners, Melinda G, Théo Michelot, Eleanor I Heywood, Rachael A Orben, Richard A Phillips, Alexei L Vyssotski, Scott A Shaffer, and Lesley H Thorne. 2021. "Hidden Markov Models Identify Major Movement Modes in Accelerometer and Magnetometer Data from Four Albatross Species." *Movement Ecology* 9 (1): 1–16.

Dobrow, Robert P. 2016. *Introduction to Stochastic Processes with R*. John Wiley & Sons.

Grimmett, Geoffrey, and David Stirzaker. 2020. *Probability and Random Processes, Fourth Edition*. Oxford University Press.

Korosteleva, Olga. 2022. *Stochastic Processes with R: An Introduction*. CRC Press.

Norris, James R. 1998. *Markov Chains*. 2. Cambridge University Press.

Ross, Sheldon M. 2019. *Introduction to Probability Models, 12th Edition*. Academic Press.

Zucchini, Walter, Iain L MacDonald, and Roland Langrock. 2017. *Hidden Markov Models for Time Series: An Introduction Using R, 2nd Edition*. CRC Press.