CHAPTER 1

Preliminaries: mixtures and Markov chains

1.1 Introduction

Hidden Markov models (HMMs) are models in which the distribution that generates an observation depends on the state of an underlying and unobserved Markov process. They provide flexible general-purpose models for univariate and multivariate time series, especially for discrete-valued series, including categorical series and series of counts.

The purposes of this chapter are to provide a brief and informal introduction to HMMs, and to their many potential uses, and then to discuss two topics that will be fundamental in understanding the structure of such models. In Section 1.2 we give an account of (finite) mixture distributions, because the marginal distribution of a hidden Markov model is a mixture distribution. Then, in Section 1.3, we introduce Markov chains, which provide the underlying 'parameter process' of a hidden Markov model.

Consider, as an example, the series of annual counts of major earthquakes (i.e. magnitude 7 and above) for the years 1900–2006, both inclusive, displayed in Table 1.1 and Figure 1.1.* For this series, the application of standard models such as autoregressive moving-average (ARMA) models would be inappropriate, because such models are based on the normal distribution. Instead, the usual model for unbounded counts is the Poisson distribution, but, as will be demonstrated later, the series displays considerable overdispersion relative to the Poisson distribution, and strong positive serial dependence. A model consisting of independent Poisson random variables would therefore for two reasons also be inappropriate. An examination of Figure 1.1 suggests that there may be some periods with a low rate of earthquakes, and some with a relatively high rate. HMMs, which allow the probability distribution of each observation to depend on the unobserved (or 'hidden') state of a Markov chain, can accommodate both overdispersion and serial dependence. We

* These data were downloaded from `http://neic.usgs.gov/neis/eqlists` on 25 July 2007. Note, however, that the US Geological Survey undertook a systematic review, and there may be minor differences between the information now available and the data we present here.

Table 1.1 *Number of major earthquakes (magnitude 7 or greater) in the world, 1900–2006; to be read across rows.*

13	14	8	10	16	26	32	27	18	32	36	24	22	23	22	18	25	21	21	14
8	11	14	23	18	17	19	20	22	19	13	26	13	14	22	24	21	22	26	21
23	24	27	41	31	27	35	26	28	36	39	21	17	22	17	19	15	34	10	15
22	18	15	20	15	22	19	16	30	27	29	23	20	16	21	21	25	16	18	15
18	14	10	15	8	15	6	11	8	7	18	16	13	12	13	20	15	16	12	18
15	16	13	15	16	11	11													

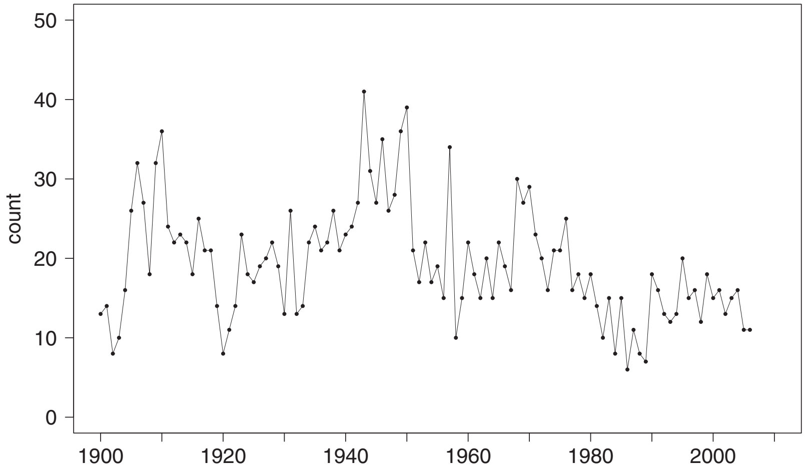


Figure 1.1 *Number of major earthquakes (magnitude 7 or greater) in the world, 1900–2006.*

shall use this series of earthquake counts as a running example in Part I of the book, in order to illustrate the fitting of a Poisson–HMM and many other aspects of that model.

HMMs have been used for at least three decades in signal-processing applications, especially in the context of automatic speech recognition, but interest in their theory and application has expanded to other fields, for example:

- all kinds of recognition – face, gesture, handwriting, signature;
- bioinformatics – biological sequence analysis;
- environment – rainfall, earthquakes, wind direction;

- finance – series of daily returns;
- biophysics – ion channel modelling;
- ecology – animal behaviour.

Attractive features of HMMs include their simplicity, their general mathematical tractability, and specifically the fact that the likelihood is relatively straightforward to compute. The main aim of this book is to illustrate how HMMs can be used as general-purpose models for time series.

Following this preliminary chapter, the book introduces what we shall call the **basic HMM**: basic in the sense that it is univariate, is based on a homogeneous Markov chain, and has neither trend nor seasonal variation. The observations may be either discrete- or continuous-valued, but we initially ignore information that may be available on covariates. We focus on the following issues:

- parameter estimation (Chapters 3 and 4);
- point and interval forecasting (Chapter 5);
- decoding, i.e. estimating the sequence of hidden states (Chapter 5);
- model selection, model checking and outlier detection (Chapter 6).

In Chapter 7 we give one example of the Bayesian approach to inference. In Chapter 8 we give examples of how several **R** packages can be used to fit basic HMMs to data and to decode.

In Part II we discuss the many possible extensions of the basic HMM to a wider range of models. These include HMMs for series with trend and seasonal variation, methods to include covariate information from other time series, multivariate models of various types, HMM approximations to hidden semi-Markov models and to models with continuous-valued state process, and HMMs for longitudinal data.

Part III of the book offers fairly detailed applications of HMMs to time series arising in a variety of subject areas. These are intended to illustrate the theory covered in Parts I and II, and also to demonstrate the versatility of HMMs. Indeed, so great is the variety of HMMs that it is hard to imagine this diversity being exhaustively covered by any single software package. In some applications the model needs to accommodate some special features of the time series, which makes it necessary to write one's own code. We have found the computing environment **R** (Ihaka and Gentleman, 1996; R Core Team, 2015) to be particularly convenient for this purpose.

Many of the chapters contain exercises, some theoretical and some practical. Because one always learns more about models by applying them in practice, and because some aspects of the theory of HMMs are covered only in these exercises, we regard these as an important part of

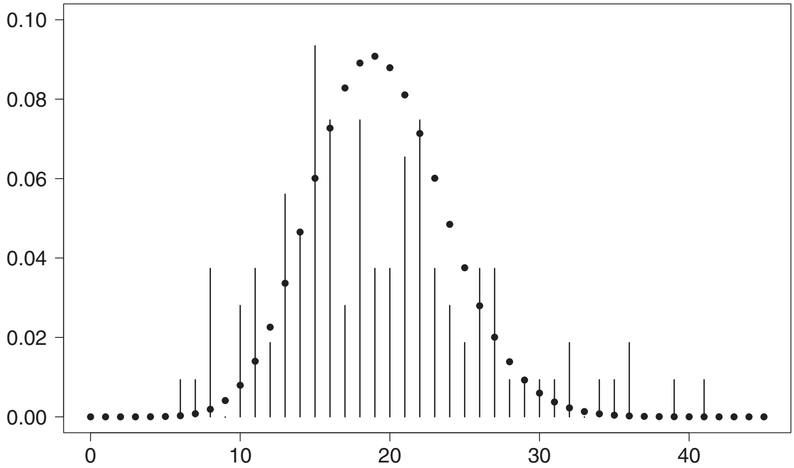


Figure 1.2 *Major earthquakes, 1900–2006: bar plot of relative frequencies of counts, and fitted Poisson distribution.*

the book. As regards the practical exercises, our strategy has been to give examples of **R** functions for some important but simple cases, and to encourage readers to learn to write their own code, initially just by modifying the functions given in Appendix A.

1.2 Independent mixture models

1.2.1 Definition and properties

Consider again the series of earthquake counts displayed in Figure 1.1. A standard model for unbounded counts is the Poisson distribution, with its probability function $p(x) = e^{-\lambda}\lambda^x/x!$ and the property that the variance equals the mean. However, for the earthquakes series the sample variance, $s^2 \approx 52$, is much larger than the sample mean, $\bar{x} \approx 19$, which indicates strong overdispersion relative to the Poisson distribution. The lack of fit is confirmed by Figure 1.2, which displays the fitted Poisson distribution and a bar plot of the relative frequencies of the counts.

One method of dealing with overdispersed observations with a bimodal or (more generally) multimodal distribution is to use a mixture model. Mixture models are designed to accommodate unobserved heterogeneity in the population; that is, the population may consist of unobserved groups, each having a distinct distribution for the observed variable.

Consider, for example, the distribution of the number, X, of packets of cigarettes bought by the customers of a supermarket. The customers can be divided into groups, for example, non-smokers, occasional smokers, and regular smokers. Now even if the number of packets bought by customers within each group were Poisson-distributed, the distribution of X would not be Poisson; it would be overdispersed relative to the Poisson, and maybe even multimodal.

Analogously, suppose that each count in the earthquakes series is generated by one of two Poisson distributions, with means λ_1 and λ_2, where the choice of mean is determined by some other random mechanism which we call the **parameter process**. Suppose also that λ_1 is selected with probability δ_1 and λ_2 with probability $\delta_2 = 1 - \delta_1$. We shall see later in this chapter that the variance of the resulting distribution exceeds the mean by $\delta_1\delta_2(\lambda_1 - \lambda_2)^2$. If the parameter process is a series of independent random variables, the counts are also independent, hence the term 'independent mixture'.

In general, an independent mixture distribution consists of a finite number, say m, of component distributions and a 'mixing distribution' which selects from these components. The component distributions may be either discrete or continuous. In the case of two components, the mixture distribution depends on two probability or density functions:

component	1	2
probability or density function	$p_1(x)$	$p_2(x)$.

To specify the component, one needs a discrete random variable C which performs the mixing:

$$C = \begin{cases} 1 & \text{with probability } \delta_1 \\ 2 & \text{with probability } \delta_2 = 1 - \delta_1. \end{cases}$$

The structure of that process for the case of two continuous component distributions is illustrated in Figure 1.3. In that example one can think of C as the outcome of tossing a coin with probability 0.75 of 'heads': if the outcome is 'heads', then $C = 1$ and an observation is drawn from p_1; if it is 'tails', then $C = 2$ and an observation is drawn from p_2. We suppose that we do not know the value C, that is, which of p_1 or p_2 was active when the observation was generated.

The extension to m components is straightforward. Let $\delta_1, \ldots, \delta_m$ denote the probabilities assigned to the different components, and let $p_1, \ldots, p_m$ denote their probability or density functions. Let X denote the random variable which has the mixture distribution. In the discrete

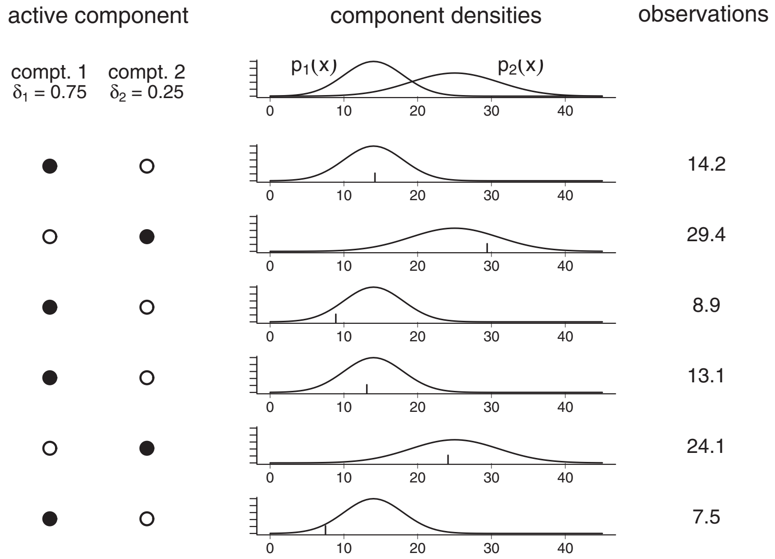


Figure 1.3 *Process structure of a two-component mixture distribution. From top to bottom, the states are* $1, 2, 1, 1, 2, 1$. *The corresponding component distributions are shown in the middle. The observations are generated from the active component density.*

case the probability function of X is given by

$$\begin{aligned} p(x) &= \sum_{i=1}^{m} \Pr(X = x \mid C = i) \Pr(C = i) \\ &= \sum_{i=1}^{m} \delta_i p_i(x). \end{aligned}$$

The continuous case is analogous. The expectation of the mixture can be given in terms of the expectations of the component distributions. Letting Y_i denote the random variable with probability function p_i, we have

$$\mathrm{E}(X) = \sum_{i=1}^{m} \Pr(C = i)\, \mathrm{E}(X \mid C = i) = \sum_{i=1}^{m} \delta_i\, \mathrm{E}(Y_i).$$

The same result holds for a mixture of continuous distributions.

More generally, for a mixture the kth moment about the origin is simply a linear combination of the kth moments of its components Y_i:

$$\mathrm{E}(X^k) = \sum_{i=1}^{m} \delta_i \,\mathrm{E}(Y_i^k), \quad k = 1, 2, \ldots .$$

Note that the analogous result does not hold for central moments. In particular, the variance of X is not a linear combination of the variances of its components Y_i. Exercise 1 asks the reader to prove that, in the two-component case, the variance of the mixture is given by

$$\mathrm{Var}(X) = \delta_1 \mathrm{Var}(Y_1) + \delta_2 \mathrm{Var}(Y_2) + \delta_1\delta_2 \big(\mathrm{E}(Y_1) - \mathrm{E}(Y_2)\big)^2 .$$

1.2.2 Parameter estimation

The estimation of the parameters of a mixture distribution is often performed by maximum likelihood (ML). The likelihood of a mixture model with m components is given, for both discrete and continuous cases, by

$$L(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, \delta_1, \ldots, \delta_m \mid x_1, \ldots, x_n) = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i p_i(x_j, \boldsymbol{\theta}_i). \qquad (1.1)$$

Here $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ are the parameter vectors of the component distributions, $\delta_1, \ldots, \delta_m$ are the mixing parameters, totalling 1, and $x_1, \ldots, x_n$ are the n observations. Thus, in the case of component distributions each specified by one parameter, $2m - 1$ independent parameters have to be estimated. Except perhaps in special cases, analytic maximization of such a likelihood is not possible, but it is in general straightforward to evaluate it fast; see Exercise 3. Numerical maximization will be illustrated here by considering the case of a mixture of Poisson distributions.

Suppose that $m = 2$ and the two components are Poisson-distributed with means λ_1 and λ_2. Let δ_1 and δ_2 be the mixing parameters (with $\delta_1 + \delta_2 = 1$). The mixture distribution p is then given by

$$p(x) = \delta_1 \frac{\lambda_1^x e^{-\lambda_1}}{x!} + \delta_2 \frac{\lambda_2^x e^{-\lambda_2}}{x!} .$$

Since $\delta_2 = 1 - \delta_1$, there are only three parameters to be estimated: λ_1, λ_2 and δ_1. The likelihood is

$$L(\lambda_1, \lambda_2, \delta_1 \mid x_1, \ldots, x_n) \quad = \quad \prod_{i=1}^{n} \left(\delta_1 \frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} + (1 - \delta_1) \frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!} \right) .$$

The analytic maximization of L with respect to λ_1, λ_2 and δ_1 would be awkward, as L is the product of n factors, each of which is a sum. First taking the logarithm and then differentiating does not greatly simplify matters either. Therefore parameter estimation is more conveniently carried out by direct numerical maximization of the likelihood (or its logarithm), although the EM algorithm is a commonly used alternative; see, for example, McLachlan and Peel (2000) or Frühwirth-Schnatter (2006). (We shall in Chapter 4 discuss the EM algorithm more fully in the context of the estimation of HMMs.) A useful **R** package for estimation in mixture models is `flexmix` (Leisch, 2004). However, it is straightforward to write one's own **R** code to evaluate, and then maximize, mixture likelihoods in simple cases.

This log-likelihood can then be maximized by using (for example) the **R** function `nlm`. However, the parameters $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are constrained by $\sum_{i=1}^{m} \delta_i = 1$ and (for $i = 1, \ldots, m$) $\delta_i > 0$ and $\lambda_i > 0$. It is therefore necessary to reparametrize if one wishes to use an unconstrained optimizer such as `nlm`. One possibility is to maximize the likelihood with respect to the $2m - 1$ unconstrained 'working parameters'

$$\eta_i = \log \lambda_i \quad (i = 1, \ldots, m)$$

and

$$\tau_i = \log\left(\frac{\delta_i}{1 - \sum_{j=2}^{m} \delta_j}\right) \quad (i = 2, \ldots, m).$$

One recovers the original 'natural parameters' via

$$\lambda_i = e^{\eta_i} \quad (i = 1, \ldots, m),$$

$$\delta_i = \frac{e^{\tau_i}}{1 + \sum_{j=2}^{m} e^{\tau_j}} \quad (i = 2, \ldots, m),$$

and $\delta_1 = 1 - \sum_{j=2}^{m} \delta_i$. The following code implements the above ideas in order to fit a mixture of four Poisson distributions to the earthquake counts. The results are given for $m = 1, 2, 3, 4$ in Table 1.2.

```
# Function to compute -log(likelihood)
mllk <- function(wpar,x){ zzz <- w2n(wpar)
        -sum(log(outer(x,zzz$lambda,dpois)%*%zzz$delta)) }

# Function to transform natural to working parameters
n2w  <- function(lambda,delta)log(c(lambda,delta[-1]/(1-sum(delta[-1]))))

# Function to transform working to natural parameters
w2n  <- function(wpar){m <- (length(wpar)+1)/2
        lambda <- exp(wpar[1:m])
        delta  <- exp(c(0,wpar[(m+1):(2*m-1)]))
return(list(lambda=lambda,delta=delta/sum(delta))) }

# Read data, specify starting values, minimize -log(likelihood),
# and transform to natural parameters
x        <- read.table("earthquakes.txt")[,2] # Set your own path.
wpar     <- n2w(c(10,20,25,30),c(1,1,1,1)/4)
w2n(nlm(mllk,wpar,x)$estimate)
```

Notice how, in this code, the use of the function `outer` makes it possible to evaluate a Poisson mixture log-likelihood in a single compact expression rather than a loop. But if the distributions being mixed were distributions with more than one parameter (e.g. normal), a slightly different approach would be needed.

1.2.3 Unbounded likelihood in mixtures

There is one aspect of mixtures of continuous distributions that differs from the discrete case and is worth highlighting. It is this: it can happen that, in the vicinity of certain parameter combinations, the likelihood is unbounded. For instance, in the case of a mixture of normal distributions, the likelihood becomes arbitrarily large if one sets a component mean equal to one of the observations and allows the corresponding variance to tend to zero. The problem has been extensively discussed in the literature on mixture models, and there are those who would say that, if the likelihood is thus unbounded, the ML estimates simply 'do not exist'; see, for instance, Scholz (2006, p. 4630).

The source of the problem, however, is just the use of densities rather than probabilities in the likelihood; it would not arise if one were to replace each density value in a likelihood by the probability of the interval corresponding to the recorded value. (For example, an observation recorded as '12.4' is associated with the interval $[12.35, 12.45)$.) In the context of independent mixtures one replaces the expression

$$\prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i p_i(x_j, \boldsymbol{\theta}_i)$$

for the likelihood (see equation (1.1)) by the **discrete likelihood**

$$L = \prod_{j=1}^{n} \sum_{i=1}^{m} \delta_i \int_{a_j}^{b_j} p_i(x, \boldsymbol{\theta}_i)\, \mathrm{d}x, \tag{1.2}$$

where the interval (a_j, b_j) consists of those values which, if observed, would be recorded as x_j. This simply amounts to acknowledging explicitly the interval nature of all supposedly continuous observations. More generally, the discrete likelihood of observations on a set of random variables $X_1, X_2, \ldots, X_n$ is a probability of the form $\Pr(a_t < X_t < b_t, \text{ for all } t)$. We use the term **continuous likelihood** for the joint density evaluated at the observations.

Another way of avoiding the problem is to impose a lower bound on the variances and search for the best local maximum subject to that bound. It can happen, though, that one is fortunate enough to avoid the likelihood 'spikes' when searching for a local maximum; in this respect

Table 1.2 *Poisson independent mixture models fitted to the earthquakes series. The number of components is m, the mixing probabilities are denoted by δ_i, and the component means by λ_i. The maximized likelihood is L.*

Model	i	δ_i	λ_i	$-\log L$	Mean	Variance
$m = 1$	1	1.000	19.364	391.9189	19.364	19.364
$m = 2$	1	0.676	15.777	360.3690	19.364	46.182
	2	0.324	26.840			
$m = 3$	1	0.278	12.736	356.8489	19.364	51.170
	2	0.593	19.785			
	3	0.130	31.629			
$m = 4$	1	0.093	10.584	356.7337	19.364	51.638
	2	0.354	15.528			
	3	0.437	20.969			
	4	0.116	32.079			
observations					19.364	51.573

good starting values can help. The phenomenon of unbounded likelihood does not arise for discrete-valued observations because the likelihood is in that case a probability and thereby bounded by 0 and 1.

For a thorough account of the unbounded likelihood 'problem', see Liu, Wu and Meeker (2015). Liu *et al.* use the terms 'density-approximation likelihood' and 'correct likelihood' for what we call the continuous likelihood and discrete likelihood, respectively.

1.2.4 Examples of fitted mixture models

Mixtures of Poisson distributions

If one uses `nlm` to fit a mixture of m Poisson distributions ($m = 1, 2, 3, 4$) to the earthquakes data, one obtains the results displayed in Table 1.2. Notice that there is a very clear improvement in likelihood resulting from the addition of a second component, and very little improvement from addition of a fourth – apparently insufficient to justify the additional two parameters. Section 6.1 will discuss the model selection problem in more detail. Figure 1.4 presents a histogram of the observed counts and the

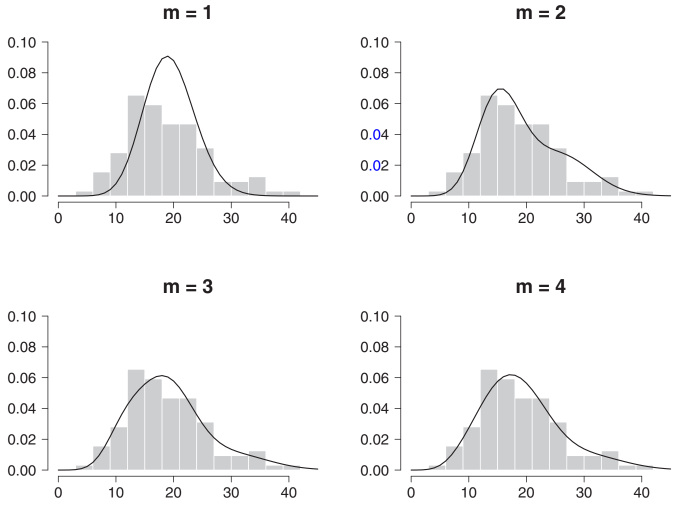


Figure 1.4 *Earthquakes data: histogram of counts, compared to mixtures of one, two, three and four Poisson distributions.*

four models fitted. It is clear that the mixtures fit the observations much better than does a single Poisson distribution, and visually the three- and four-state models seem adequate. The better fit of the mixtures is also evident from the variances of the four models as presented in Table 1.2. In computing the means and variances of the models we have used $\mathrm{E}(X) = \sum_i \delta_i \lambda_i$ and $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$, with $\mathrm{E}(X^2) = \sum_i \delta_i (\lambda_i + \lambda_i^2)$. For comparison we also used the **R** package `flexmix` to fit the same four models. The results corresponded closely except in the case of the four-component model, where the highest likelihood value that we found by `flexmix` was 356.7759 and the component means differed somewhat.

Note, however, that the above discussion ignores the possibility of serial dependence in the earthquakes data, a point we shall take up in Chapter 2.

A mixture of normal distributions

As a very simple example of the fitting of an independent mixture of normal distributions, consider the data presented in Table 8.1 of Hastie, Tibshirani and Friedman (2009, p. 273); see our Table 1.3. Hastie *et al.* use the EM algorithm to fit a mixture model with two normal components.

Table 1.3 *Data of Hastie* et al. *(2009), plus two mixture models. The first model was fitted by direct numerical maximization in* **R***, the second is the model fitted by EM by Hastie* et al.

−0.39	0.12	0.94	1.67	1.76	2.44	3.72	4.28	4.92	5.53
0.06	0.48	1.01	1.68	1.80	3.25	4.12	4.60	5.28	6.22

i	δ_i	μ_i	σ_i^2	$-\log L$
1	0.4454	4.656	0.8188	38.9134
2	0.5546	1.083	0.8114	
1	0.454	4.62	0.87	
2	0.546	1.06	0.77	

Our two-component model, fitted by direct numerical maximization of the log-likelihood in **R**, has log-likelihood −38.9134, and is also displayed in Table 1.3. (Here we used the continuous likelihood, i.e. the joint density of the observations, not the discrete likelihood.) The parameter estimates are close to those given by Hastie *et al.*, but not identical.

1.3 Markov chains

We now introduce Markov chains, a second building-block of hidden Markov models. Our treatment is restricted to those few aspects of discrete-time Markov chains that we need. Thus, although we shall make passing reference to properties such as irreducibility and aperiodicity, we shall not dwell on such technical issues. For a general account of the topic, see Grimmett and Stirzaker (2001, Chapter 6), or Feller's classic text (Feller, 1968).

1.3.1 Definitions and example

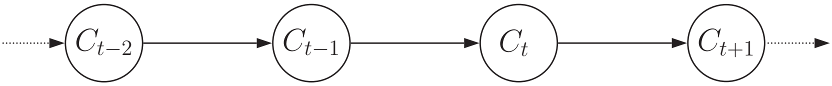
A sequence of discrete random variables $\{C_t : t \in \mathbb{N}\}$ is said to be a (discrete-time) **Markov chain** (MC) if, for all $t \in \mathbb{N}$, it satisfies the **Markov property**

$$\Pr(C_{t+1} \mid C_t, \ldots, C_1) = \Pr(C_{t+1} \mid C_t).$$

That is, conditioning on the 'history' of the process up to time t is equivalent to conditioning only on the most recent value C_t. For compactness we define $\mathbf{C}^{(t)}$ as the history $(C_1, C_2, \ldots, C_t)$, in which case the Markov property can be written as

$$\Pr(C_{t+1} \mid \mathbf{C}^{(t)}) = \Pr(C_{t+1} \mid C_t).$$

The Markov property can be regarded as a first relaxation of the assumption of independence. The random variables $\{C_t\}$ are dependent in a specific way that is mathematically convenient, as displayed in the following directed graph in which the past and the future are dependent only through the present.



Important quantities associated with a Markov chain are the conditional probabilities called **transition probabilities**:

$$\Pr(C_{s+t} = j \mid C_s = i).$$

If these probabilities do not depend on s, the Markov chain is called **homogeneous**, otherwise non-homogeneous. Unless there is an explicit indication to the contrary, we shall assume that the Markov chain under discussion is homogeneous, in which case the transition probabilities will be denoted by

$$\gamma_{ij}(t) = \Pr(C_{s+t} = j \mid C_s = i).$$

Notice that the notation $\gamma_{ij}(t)$ does not involve s. The matrix $\mathbf{\Gamma}(t)$ is defined as the matrix with (i,j) element $\gamma_{ij}(t)$.

An important property of all finite state-space homogeneous Markov chains is that they satisfy the **Chapman–Kolmogorov equations**:

$$\mathbf{\Gamma}(t+u) = \mathbf{\Gamma}(t)\,\mathbf{\Gamma}(u).$$

The proof requires only the definition of conditional probability and the application of the Markov property: this is Exercise 10. The Chapman–Kolmogorov equations imply that, for all $t \in \mathbb{N}$,

$$\mathbf{\Gamma}(t) = \mathbf{\Gamma}(1)^t;$$

that is, the matrix of t-step transition probabilities is the tth power of $\mathbf{\Gamma}(1)$, the matrix of one-step transition probabilities. The matrix $\mathbf{\Gamma}(1)$, which will be abbreviated as $\mathbf{\Gamma}$, is a square matrix of probabilities with row sums equal to 1:

$$\mathbf{\Gamma} \quad = \quad \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{pmatrix},$$

where (throughout this text) m denotes the number of states of the Markov chain. The statement that the row sums are equal to 1 can be written as $\mathbf{\Gamma 1}' = \mathbf{1}'$; that is, the column vector $\mathbf{1}'$ is a right eigenvector of $\mathbf{\Gamma}$ and corresponds to eigenvalue 1. We shall refer to $\mathbf{\Gamma}$ as the (one-step) **transition probability matrix** (t.p.m.). Many authors use instead the term 'transition matrix'; we avoid that term because of possible confusion with a matrix of transition counts, or a matrix of transition intensities.

The **unconditional probabilities** $\Pr(C_t = j)$ of a Markov chain being in a given state at a given time t are often of interest. We denote these by the row vector

$$\mathbf{u}(t) = (\Pr(C_t = 1), \dots, \Pr(C_t = m)), \quad t \in \mathbb{N}.$$

We refer to $u(1)$ as the **initial distribution** of the Markov chain. To deduce the distribution at time $t + 1$ from that at t we postmultiply by the transition probability matrix $\mathbf{\Gamma}$:

$$\mathbf{u}(t+1) = \mathbf{u}(t)\mathbf{\Gamma}. \tag{1.3}$$

The proof of this statement is left as an exercise.

Example. Imagine that the sequence of rainy and sunny days is such that each day's weather depends only on the previous day's, and the transition probabilities are given by the following table.

	day $t + 1$	
day t	rainy	sunny
rainy	0.9	0.1
sunny	0.6	0.4

That is, if today is rainy, the probability that tomorrow will be rainy is 0.9; if today is sunny, that probability is 0.6. The weather is then a two-state homogeneous Markov chain, with t.p.m. $\mathbf{\Gamma}$ given by

$$\mathbf{\Gamma} \;=\; \begin{pmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{pmatrix}.$$

Now suppose that today (time 1) is a sunny day. This means that the distribution of today's weather is

$$\mathbf{u}(1) = \big(\Pr(C_1 = 1), \Pr(C_1 = 2)\big) = \big(0, 1\big).$$

The distribution of the weather of tomorrow, the day after tomorrow, and so on, can be calculated by repeatedly postmultiplying $\mathbf{u}(1)$ by $\mathbf{\Gamma}$, the t.p.m.:

$$\begin{aligned}\mathbf{u}(2) &= \big(\Pr(C_2=1), \Pr(C_2=2)\big) = \mathbf{u}(1)\boldsymbol{\Gamma} = (0.6, 0.4),\\ \mathbf{u}(3) &= \big(\Pr(C_3=1), \Pr(C_3=2)\big) = \mathbf{u}(2)\boldsymbol{\Gamma} = (0.78, 0.22), \text{ etc.}\end{aligned}$$

1.3.2 Stationary distributions

A Markov chain with transition probability matrix $\boldsymbol{\Gamma}$ is said to have **stationary distribution** $\boldsymbol{\delta}$ (a row vector with non-negative elements) if $\boldsymbol{\delta\Gamma} = \boldsymbol{\delta}$ and $\boldsymbol{\delta}\mathbf{1}' = 1$. The first of these requirements expresses the stationarity, the second is the requirement that $\boldsymbol{\delta}$ is indeed a probability distribution. For instance, the Markov chain with t.p.m. given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

has as stationary distribution $\boldsymbol{\delta} = \frac{1}{32}(15, 9, 8)$.

Since $\mathbf{u}(t+1) = \mathbf{u}(t)\boldsymbol{\Gamma}$, a Markov chain started from its stationary distribution will continue to have that distribution at all subsequent time points, and we shall refer to such a process as a **stationary Markov chain**. It is perhaps worth stating that this assumes more than merely homogeneity. Homogeneity alone would not be sufficient to render the Markov chain a stationary process, and we prefer to reserve the adjective 'stationary' for homogeneous Markov chains that have the additional property that the initial distribution $\mathbf{u}(1)$ is the stationary distribution and are therefore stationary processes. Not all authors use this terminology, however; see, for example, McLachlan and Peel (2000, p. 328), who use the word 'stationary' of a Markov chain where we would say 'homogeneous'.

An irreducible (homogeneous, discrete-time, finite state-space) Markov chain has a unique, strictly positive, stationary distribution. Note that although the technical assumption of irreducibility is needed for this conclusion, aperiodicity is not; see Grimmett and Stirzaker (2001, Lemma 6.3.5 on p. 225 and Theorem 6.4.3 on p. 227).

If, however, one does add the assumption of aperiodicity, it follows that a unique limiting distribution exists, and is precisely the stationary distribution; see Feller (1968, p. 394). Since we shall always assume aperiodicity and irreducibility, the terms 'limiting distribution' and 'stationary distribution' are for our purposes synonymous.

A general result that can conveniently be used to compute a stationary distribution (see Exercise 9(a)) is as follows. The vector $\boldsymbol{\delta}$ with non-negative elements is a stationary distribution of the Markov chain with t.p.m. $\mathbf{\Gamma}$ if and only if

$$\boldsymbol{\delta}(\mathbf{I}_m - \mathbf{\Gamma} + \mathbf{U}) = \mathbf{1},$$

where $\mathbf{1}$ is a row vector of ones, $\mathbf{I}_m$ is the $m \times m$ identity matrix, and $\mathbf{U}$ is the $m \times m$ matrix of ones. Alternatively, a stationary distribution can be found by deleting one of the equations in the system $\boldsymbol{\delta}\mathbf{\Gamma} = \boldsymbol{\delta}$ and replacing it by $\sum_i \delta_i = 1$.

1.3.3 Autocorrelation function

We shall have occasion, for example in Section 2.2.3 and in Exercise 4(f) in Chapter 2, to compare the autocorrelation function (ACF) of a hidden Markov model with that of its underlying Markov chain $\{C_t\}$, on the states 1, 2, ..., m. We assume that these states are quantitative and not merely categorical. The ACF of $\{C_t\}$, assumed stationary and irreducible, may be obtained as follows.

Firstly, defining $\mathbf{v} = (1, 2, \ldots, m)$ and $\mathbf{V} = \text{diag}(1, 2, \ldots, m)$, we have, for all non-negative integers k,

$$\text{Cov}(C_t, C_{t+k}) = \boldsymbol{\delta}\mathbf{V}\mathbf{\Gamma}^k\mathbf{v}' - (\boldsymbol{\delta}\mathbf{v}')^2; \qquad (1.4)$$

the proof is Exercise 11. Secondly, if $\mathbf{\Gamma}$ is diagonalizable, and its eigenvalues (other than 1) are denoted by $\omega_2, \omega_3, \ldots, \omega_m$, then $\mathbf{\Gamma}$ can be written as

$$\mathbf{\Gamma} = \mathbf{U}\mathbf{\Omega}\mathbf{U}^{-1},$$

where $\mathbf{\Omega}$ is $\text{diag}(1, \omega_2, \omega_3, \ldots, \omega_m)$ and the columns of $\mathbf{U}$ are corresponding right eigenvectors of $\mathbf{\Gamma}$. We then have, for non-negative integers k,

$$\begin{aligned} \text{Cov}(C_t, C_{t+k}) &= \boldsymbol{\delta}\mathbf{V}\mathbf{U}\mathbf{\Omega}^k\mathbf{U}^{-1}\mathbf{v}' - (\boldsymbol{\delta}\mathbf{v}')^2 \\ &= \mathbf{a}\mathbf{\Omega}^k\mathbf{b}' - a_1 b_1 \\ &= \sum_{i=2}^{m} a_i b_i \omega_i^k, \end{aligned}$$

where $\mathbf{a} = \boldsymbol{\delta}\mathbf{V}\mathbf{U}$ and $\mathbf{b}' = \mathbf{U}^{-1}\mathbf{v}'$. Hence $\text{Var}(C_t) = \sum_{i=2}^m a_i b_i$ and, for non-negative integers k,

$$\rho(k) \equiv \text{Corr}(C_t, C_{t+k}) = \sum_{i=2}^{m} a_i b_i \omega_i^k \Big/ \sum_{i=2}^{m} a_i b_i. \qquad (1.5)$$

This is a weighted average of the kth powers of the eigenvalues ω_2, ω_3, ..., ω_m, and somewhat similar to the ACF of a Gaussian autoregressive process of order $m-1$. Note that equation (1.5) implies in the case $m = 2$ that $\rho(k) = \rho(1)^k$ for all non-negative integers k, and that $\rho(1)$ is the eigenvalue other than 1 of $\mathbf{\Gamma}$.

1.3.4 Estimating transition probabilities

If we are given a realization of a Markov chain, and wish to estimate the transition probabilities, one approach – but not the only one – is to find the transition counts and estimate the transition probabilities as relative frequencies. For instance, if the MC has three states and the observed sequence is

2332111112 3132332122 3232332222 3132332212 3232132232
3132332223 3232331232 3232331222 3232132123 3132332121,

then the matrix of transition counts is

$$(f_{ij}) = \begin{pmatrix} 4 & 7 & 6 \\ 8 & 10 & 24 \\ 6 & 24 & 10 \end{pmatrix},$$

where f_{ij} denotes the number of transitions observed from state i to state j. Since the number of transitions from state 2 to state 3 is 24, and the total number of transitions from state 2 is 8+10+24, a relative frequency estimate of γ_{23} is 24/42. The t.p.m. $\mathbf{\Gamma}$ is therefore plausibly estimated by

$$\begin{pmatrix} 4/17 & 7/17 & 6/17 \\ 8/42 & 10/42 & 24/42 \\ 6/40 & 24/40 & 10/40 \end{pmatrix}.$$

We shall now show that this is in fact the conditional ML estimate of $\mathbf{\Gamma}$, conditioned on the first observation.

Suppose, then, that we wish to estimate the m^2-m parameters γ_{ij} $(i \neq j)$ of an m-state Markov chain $\{C_t\}$ from a realization $c_1, c_2, \ldots, c_T$. The likelihood conditioned on the first observation is

$$L = \prod_{i=1}^{m} \prod_{j=1}^{m} \gamma_{ij}^{f_{ij}}.$$

The log-likelihood is

$$l = \sum_{i=1}^{m} \left(\sum_{j=1}^{m} f_{ij} \log \gamma_{ij} \right) = \sum_{i=1}^{m} l_i \text{ (say)},$$

and we can maximize l by maximizing each l_i separately. Substituting $1 - \sum_{k \neq i} \gamma_{ik}$ for γ_{ii}, differentiating l_i with respect to an off-diagonal transition probability γ_{ij}, and equating the derivative to zero yields

$$0 = \frac{-f_{ii}}{1 - \sum_{k \neq i} \gamma_{ik}} + \frac{f_{ij}}{\gamma_{ij}} = -\frac{f_{ii}}{\gamma_{ii}} + \frac{f_{ij}}{\gamma_{ij}}.$$

Hence, unless a denominator is zero in the above equation, $f_{ij}\gamma_{ii} = f_{ii}\gamma_{ij}$, and so $\gamma_{ii} \sum_{j=1}^{m} f_{ij} = f_{ii}$. This implies that, at a maximum of the likelihood,

$$\gamma_{ii} = f_{ii} \Big/ \sum_{j=1}^{m} f_{ij} \quad \text{and} \quad \gamma_{ij} = f_{ij}\gamma_{ii}/f_{ii} = f_{ij} \Big/ \sum_{j=1}^{m} f_{ij}.$$

(We could instead use Lagrange multipliers to express the constraints $\sum_{j=1}^{m} \gamma_{ij} = 1$ subject to which we seek to maximize the terms l_i and therefore the likelihood; see Exercise 12.)

The estimator $\widehat{\gamma}_{ij} = f_{ij} / \sum_{k=1}^{m} f_{ik}$ $(i, j = 1, \dots, m)$ – which is just the empirical transition probability – is thereby seen to be a conditional ML estimator of γ_{ij}. This estimator of $\mathbf{\Gamma}$ satisfies the requirement that the row sums should equal 1.

The assumption of stationarity of the Markov chain was not used in the above derivation. If we wish to assume stationarity, we may use the unconditional likelihood. This is the conditional likelihood as above, multiplied by the stationary probability δ_{c_1}. The unconditional likelihood or its logarithm may then be maximized numerically, subject to nonnegativity and row-sum constraints, in order to estimate the transition probabilities γ_{ij}. Bisgaard and Travis (1991) show in the case of a two-state Markov chain that, barring some extreme cases, the unconditional likelihood equations have a unique solution. For some non-trivial special cases of the two-state chain, they also derive explicit expressions for the unconditional maximum likelihood estimates (MLEs) of the transition probabilities. Since we use one such result later (in Section 17.3.1), we state it here.

Suppose the Markov chain $\{C_t\}$ takes the values 0 and 1, and that we wish to estimate the transition probabilities γ_{ij} from a sequence of observations in which there are f_{ij} transitions from state i to state j $(i, j = 0, 1)$, and $f_{11} > 0$ but $f_{00} = 0$. So in the observations a zero is always followed by a one. Define $c = f_{10} + (1 - c_1)$ and $d = f_{11}$. Then the unconditional MLEs of the transition probabilities are given by

$$\widehat{\gamma}_{01} = 1 \quad \text{and} \quad \widehat{\gamma}_{10} = \frac{-(1+d) + \big((1+d)^2 + 4c(c+d-1)\big)^{\frac{1}{2}}}{2(c+d-1)}. \qquad (1.6)$$

1.3.5 Higher-order Markov chains

This section is somewhat specialized, and the material is used only in Section 10.3 and parts of Sections 17.3.2 and 19.2.2. It will therefore not interrupt the continuity greatly if the reader should initially omit this section.

In cases where observations on a process with finite state space appear not to satisfy the Markov property, one possibility that suggests itself is to use a higher-order Markov chain, that is, a model $\{C_t\}$ satisfying the following generalization of the Markov property for some $l \geq 2$:

$$\Pr(C_t \mid C_{t-1}, C_{t-2}, \ldots) = \Pr(C_t \mid C_{t-1}, \ldots, C_{t-l}).$$

An account of such higher-order Markov chains may be found, for instance, in Lloyd (1980, Section 19.9). Although such a model is not in the usual sense a Markov chain (i.e. not a 'first-order' Markov chain), we can redefine the model in such a way as to produce an equivalent process which is. If we let $\mathbf{Y}_t = (C_{t-l+1}, C_{t-l+2}, \ldots, C_t)$, then $\{\mathbf{Y}_t\}$ is a first-order Markov chain on M^l, where M is the state space of $\{C_t\}$. Although some properties may be more awkward to establish, no essentially new theory is involved in analysing a higher-order Markov chain rather than a first-order one.

A *second-order* Markov chain, if stationary, is characterized by the transition probabilities

$$\gamma(i,j,k) = \Pr(C_t = k \mid C_{t-1} = j, C_{t-2} = i),$$

and has stationary bivariate distribution $u(j,k) = \Pr(C_{t-1} = j, C_t = k)$ satisfying

$$u(j,k) = \sum_{i=1}^{m} u(i,j)\gamma(i,j,k) \quad \text{and} \quad \sum_{j=1}^{m}\sum_{k=1}^{m} u(j,k) = 1.$$

For example, the most general stationary second-order Markov chain $\{C_t\}$ on the two states 1 and 2 is characterized by the following four transition probabilities:

$$\begin{aligned}
a &= \Pr(C_t{=}2 \mid C_{t-1}{=}1, C_{t-2}{=}1),\\
b &= \Pr(C_t{=}1 \mid C_{t-1}{=}2, C_{t-2}{=}2),\\
c &= \Pr(C_t{=}1 \mid C_{t-1}{=}2, C_{t-2}{=}1),\\
d &= \Pr(C_t{=}2 \mid C_{t-1}{=}1, C_{t-2}{=}2).
\end{aligned}$$

The process $\{\mathbf{Y}_t\} = \{(C_{t-1}, C_t)\}$ is then a first-order Markov chain, on the four states (1,1), (1,2), (2,1), (2,2), with transition probability matrix

$$\begin{pmatrix} 1-a & a & 0 & 0 \\ 0 & 0 & c & 1-c \\ 1-d & d & 0 & 0 \\ 0 & 0 & b & 1-b \end{pmatrix}. \tag{1.7}$$

Notice the structural zeros appearing in this matrix. It is not possible, for instance, to make a transition directly from $(2,1)$ to $(2,2)$; hence the zero in row 3 and column 4 in the t.p.m. (1.7). The parameters a, b, c and d are bounded by 0 and 1 but are otherwise unconstrained. The stationary distribution of $\{\mathbf{Y}_t\}$ is proportional to the vector

$$\big(b(1-d), ab, ab, a(1-c)\big),$$

from which it follows that the matrix $(u(j,k))$ of stationary bivariate probabilities for $\{C_t\}$ is

$$\frac{1}{b(1-d)+2ab+a(1-c)}\begin{pmatrix} b(1-d) & ab \\ ab & a(1-c) \end{pmatrix}.$$

The use of a general higher-order Markov chain (instead of a first-order one) increases the number of parameters of the model; a general Markov chain of order l on m states has $m^l(m-1)$ independent transition probabilities. Pegram (1980) and Raftery (1985a,b) have therefore proposed certain classes of parsimonious models for higher-order chains. Pegram's models have $m+l-1$ parameters, and those of Raftery $m(m-1)+l-1$. For $m=2$ the models are equivalent, but for $m>2$ those of Raftery are more general and can represent a wider range of dependence patterns and autocorrelation structures. In both cases an increase of one in the order of the Markov chain requires only one additional parameter.

Raftery's models, which he terms 'mixture transition distribution' (MTD) models, are defined as follows. The process $\{C_t\}$ takes values in $M=\{1,2,\dots,m\}$ and satisfies

$$\Pr(C_t=j_0 \mid C_{t-1}=j_1,\dots,C_{t-l}=j_l) = \sum_{i=1}^{l} \lambda_i\, q(j_i,j_0), \qquad (1.8)$$

where $\sum_{i=1}^{l}\lambda_i = 1$, and $\mathbf{Q}=(q(j,k))$ is an $m\times m$ matrix with nonnegative entries and row sums equal to one, such that the right-hand side of equation (1.8) is bounded by zero and one for all j_0, j_1, $\dots$, $j_l \in M$. This last requirement, which generates m^{l+1} pairs of nonlinear constraints on the parameters, ensures that the conditional probabilities in equation (1.8) are indeed probabilities, and the condition on the row sums of $\mathbf{Q}$ ensures that the sum over j_0 of these conditional probabilities is one. Note that Raftery does not assume that the parameters λ_i are non-negative.

A variety of applications are presented by Raftery (1985a) and Raftery and Tavaré (1994). In several of the fitted models there are negative estimates of some of the coefficients λ_i. For further accounts of this class of models, see Haney (1993), Berchtold (2001), and Berchtold and Raftery (2002).

Azzalini and Bowman (1990) report the fitting of a second-order Markov chain model to the binary series they use to represent the lengths of successive eruptions of the Old Faithful geyser. Their analysis, and some alternative models, will be discussed in Chapter 17.

Exercises

1. (a) Let X be a random variable which is distributed as a (δ_1, δ_2)-mixture of two distributions with expectations μ_1, μ_2, and variances σ_1^2, σ_2^2, respectively, where $\delta_1 + \delta_2 = 1$.

   i. Show that $\text{Var}(X) = \delta_1\sigma_1^2 + \delta_2\sigma_2^2 + \delta_1\delta_2(\mu_1 - \mu_2)^2$.

   ii. Show that a (non-trivial) mixture X of two Poisson distributions with distinct means is overdispersed, that is, $\text{Var}(X) > \text{E}(X)$.

   (b) Now suppose that X is a mixture of $m \geq 2$ distributions, with means μ_i and variances σ_i^2, for $i = 1, 2, \ldots, m$. The mixing distribution is $\boldsymbol{\delta}$.

   i. Show that

   $$\text{Var}(X) = \sum_{i=1}^{m} \delta_i\sigma_i^2 + \sum_{i<j} \delta_i\delta_j(\mu_i - \mu_j)^2.$$

   Hint: use either $\text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2$ or the conditional variance formula,

   $$\text{Var}(X) = \text{E}(\text{Var}(X \mid C)) + \text{Var}(\text{E}(X \mid C)).$$

   ii. Describe the circumstances in which $\text{Var}(X)$ equals the linear combination $\sum_{i=1}^{m} \delta_i\sigma_i^2$.

2. A zero-inflated Poisson distribution is sometimes used as a model for unbounded counts displaying an excessive number of zeros relative to the Poisson. Such a model is a mixture of two distributions: one is a Poisson and the other is identically zero.

   (a) Is it ever possible for such a model to display underdispersion relative to Poisson?

   (b) Now consider the zero-inflated binomial. Is it possible in such a model that the variance is less than the mean?

3. Brown and Buckley (2015, p. 308) consider a Poisson mixture likelihood of the form

   $$L = \prod_{i=1}^{n} \sum_{j=1}^{k} w_j f(x_i \mid \mu_j).$$

   (Here $f(\cdot \mid \mu)$ denotes a Poisson probability function with mean μ.) They write that 'Even for moderate values of n and k, this takes a long time to evaluate as there are k^n terms when the inner sums are expanded', and do not pursue maximum likelihood estimation.

   Explain why it is in fact possible to evaluate L or its logarithm in computations which are of order kn rather than k^n.

4. (a) Write an **R** function to minimize minus the log-likelihood of a normal mixture model with m components, using the nonlinear minimizer `nlm`.

   Hint: first write a function to transform the parameters ($\boldsymbol{\delta}$ and the parameters of the m normal distributions) into unconstrained parameters. You will also need a function to reverse the transformation. (For the Poisson case, see the code on p. 10.)

   (b) Use your code to fit a mixture of two normals to the data appearing in Table 1.3, and compare your model with those displayed in that table.

5. Consider the following data, which appear in Lange (1995, 2002, 2004). (There they are quoted from Titterington, Smith and Makov (1985) and Hasselblad (1969), but the trail leads back via Schilling (1947) and Thorndike (1926) to Whitaker (1914), where in all eight similar data sets appear as Table XV on p. 67.)

   Here n_i denotes the number of days in 1910–1912 on which there appeared, in *The Times* of London, i death notices in respect of women aged 80 or over at death.

i	0	1	2	3	4	5	6	7	8	9
n_i	162	267	271	185	111	61	27	8	3	1

   (a) Use `nlm` or `optim` in **R** to fit a mixture of two Poisson distributions to these observations. (The parameter estimates reported by Lange (2002, p. 36; 2004, p. 151) are, in our notation: $\widehat{\delta}_1 = 0.3599$, $\widehat{\lambda}_1 = 1.2561$ and $\widehat{\lambda}_2 = 2.6634$.)

   (b) Fit also a single Poisson distribution to these data. Is a single Poisson distribution adequate as a model?

   (c) Fit a mixture of three Poisson distributions to these observations.

   (d) How many components do you think are necessary?

   (e) Repeat (a)–(d) for some of the other seven data sets of Whitaker.

6. Consider the series of weekly sales (in integer units) of a particular soap product in a supermarket, as shown in Table 1.4. The data were taken from a database[†] provided by the Kilts Center for Marketing, Graduate School of Business of the University of Chicago, at: `http://gsbwww.uchicago.edu/kilts/research/db/dominicks`. The product was ‘Zest White Water 15 oz.’, with code 3700031165, and the store number 67.

[†] That database is now at
`http://research.chicagobooth.edu/kilts/marketing-databases/dominicks`.

Table 1.4 *Weekly sales of the soap product; to be read across rows.*

1	6	9	18	14	8	8	1	6	7	3	3	1	3	4	12	8	10	8	2
17	15	7	12	22	10	4	7	5	0	2	5	3	4	4	7	5	6	1	3
4	5	3	7	3	0	4	5	3	3	4	4	4	4	4	3	5	5	5	7
4	0	4	3	2	6	3	8	9	6	3	4	3	3	3	3	2	1	4	5
5	2	7	5	2	3	1	3	4	6	8	8	5	7	2	4	2	7	4	15
15	12	21	20	13	9	8	0	13	9	8	0	6	2	0	3	2	4	4	6
3	2	5	5	3	2	1	1	3	1	2	6	2	7	3	2	4	1	5	6
8	14	5	3	6	5	11	4	5	9	9	7	9	8	3	4	8	6	3	5
6	3	1	7	4	9	2	6	6	4	6	6	13	7	4	8	6	4	4	4
9	2	9	2	2	2	13	13	4	5	1	4	6	5	4	2	3	10	6	15
5	9	9	7	4	4	2	4	2	3	8	15	0	0	3	4	3	4	7	5
7	6	0	6	4	14	5	1	6	5	5	4	9	4	14	2	2	1	5	2
6	4																		

Fit Poisson mixture models with one, two, three and four components. How many components do you think are necessary?

7. Consider a stationary two-state Markov chain with transition probability matrix given by

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}.$$

(a) Show that the stationary distribution is

$$(\delta_1, \delta_2) = \frac{1}{\gamma_{12} + \gamma_{21}} (\gamma_{21}, \gamma_{12}) \ .$$

(b) Consider the case

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix},$$

and the following two sequences of observations that are assumed to be generated by the above Markov chain.

Sequence 1:	1	1	1	2	2	1
Sequence 2:	2	1	1	2	1	1

Compute the probability of each of the sequences. Note that each sequence contains the same number of ones and twos. Why are these sequences not equally probable?

8. Consider a two-state Markov chain with transition probability matrix given by

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}.$$

Show that the k-step transition probability matrix, $\mathbf{\Gamma}^k$, is given by

$$\mathbf{\Gamma}^k = \begin{pmatrix} \delta_1 & \delta_2 \\ \delta_1 & \delta_2 \end{pmatrix} + w^k \begin{pmatrix} \delta_2 & -\delta_2 \\ -\delta_1 & \delta_1 \end{pmatrix},$$

where $w = 1 - \gamma_{12} - \gamma_{21}$ and δ_1 and δ_2 are as defined in Exercise 7. (Hint: one way of showing this is to diagonalize the transition probability matrix, but there is a quicker way.)

9. (a) This is one of several possible approaches to finding the stationary distribution of a Markov chain, plundered from Grimmett and Stirzaker (2001, Exercise 6.6.5).

   Suppose $\mathbf{\Gamma}$ is the transition probability matrix of a (discrete-time, homogeneous) Markov chain on m states, and that $\boldsymbol{\delta}$ is a non-negative row vector with m components. Show that $\boldsymbol{\delta}$ is a stationary distribution of the Markov chain if and only if

   $$\boldsymbol{\delta}(\mathbf{I}_m - \mathbf{\Gamma} + \mathbf{U}) = \mathbf{1},$$

   where $\mathbf{1}$ is a row vector of ones, and $\mathbf{U}$ is an $m \times m$ matrix of ones.

   (b) Write an **R** function `statdist(gamma)` that computes the stationary distribution of the Markov chain with t.p.m. `gamma`.

   (c) Use your function to find stationary distributions corresponding to the following transition probability matrices. One of them should cause a problem!

   i. $$\begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

   ii. $$\begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} \\ 0 & 1 & 0 \end{pmatrix}$$

   iii. $$\begin{pmatrix} 0 & 0.5 & 0 & 0.5 \\ 0.75 & 0 & 0.25 & 0 \\ 0 & 0.75 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0 \end{pmatrix}$$

   iv. $$\begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 & 0 \\ 0 & 0 & 0.25 & 0.75 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

   v. $$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0 & 0.75 & 0 & 0.25 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

10. Prove the Chapman–Kolmogorov equations.

11. Prove equation (1.4).

12. Let the quantities a_i be non-negative, with $\sum_i a_i > 0$. Using a Lagrange multiplier, maximize $S = \sum_{i=1}^{m} a_i \log \delta_i$ over $\delta_i \geq 0$, subject to $\sum_i \delta_i = 1$. (Check the second- as well as the first-derivative condition.)

13. (This exercise is based on Example 2 of Bisgaard and Travis (1991).) Consider the following sequence of 21 observations, assumed to arise from a two-state (homogeneous) Markov chain:

    11101 10111 10110 11111 1.

    (a) Estimate the transition probability matrix by ML, conditional on the first observation.

    (b) Estimate the t.p.m. by unconditional ML (assuming stationarity of the Markov chain).

    (c) Use the **R** functions `contour` and `persp` to produce contour and perspective plots of the unconditional log-likelihood (as a function of the two off-diagonal transition probabilities).

14. Consider the following two transition probability matrices, neither of which is diagonalizable:

    (a)

$$\mathbf{\Gamma} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix};$$

    (b)

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9 & 0.08 & 0 & 0.02 \\ 0 & 0.7 & 0.2 & 0.1 \\ 0 & 0 & 0.7 & 0.3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

    In each case, write $\mathbf{\Gamma}$ in Jordan canonical form, and so find an explicit expression for the t-step transition probabilities (t =1, 2, ...).

15. Consider the following (very) short DNA sequence, taken from Singh (2003, p. 358):

    AACGT CTCTA TCATG CCAGG ATCTG

    (a) Fit a homogeneous Markov chain to these data by:

        i. maximizing the likelihood conditioned on the first observation;

        ii. assuming stationarity and maximizing the unconditional likelihood of all 25 observations.

    (b) Compare your estimates of the t.p.m. with each other and with the estimate displayed as Table 1 of Singh (2003, p. 360).

(c) Now repeat (a) for the following 50-nucleotide sequence, taken from Singh (2003, p. 367):

ATTAG GCACG CATTA TAATG GGCAC
CCGGA AATAA CCAGA GTTAC GGCCA.

16. Write an **R** function `rMC(n,m,gamma,delta=NULL)` that generates a series of length `n` from an `m`-state Markov chain with t.p.m. `gamma`. If the initial state distribution is given, then it should be used; otherwise the stationary distribution should be used as the initial distribution. (Use your function `statdist` from Exercise 9(b).)