(a) Estimate the transition probability matrix by ML (maximum likelihood) 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).

13. 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 ($t = 1, 2, \ldots$).

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

AACGT CTCTA TCATG CCAGG ATCTG

Fit a homogeneous Markov chain to these data by

(a) maximizing the likelihood conditioned on the first observation;

(b) assuming stationarity and maximizing the unconditional likelihood of all 25 observations.

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

15. 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 8(b).)

CHAPTER 2

Hidden Markov models: definition and properties

2.1 A simple hidden Markov model

Consider again the observed earthquake series displayed in Figure 1.1 on p. 4. The observations are unbounded counts, making the Poisson distribution a natural choice to describe them. However, the sample variance of the observations is substantially greater than the sample mean, indicating overdispersion relative to the Poisson. In Exercise 1 of Chapter 1 we saw that one can accommodate overdispersion by using a mixture model, specifically a mixture of Poisson distributions for this series.

We suppose that each count is generated by one of m Poisson distributions, with means $\lambda_1, \lambda_2, \ldots, \lambda_m$, where the choice of mean is made by a second random mechanism, the parameter process. The mean λ_i is selected with probability δ_i, where $i = 1, 2, \ldots, m$ and $\sum_{i=1}^{m} \delta_i = 1$. The variance of the mixture model is greater than its expectation, which takes care of the problem of overdispersion.

An independent mixture model will not do for the earthquake series because — by definition — it does not allow for the serial dependence in the observations. The sample autocorrelation function, displayed in

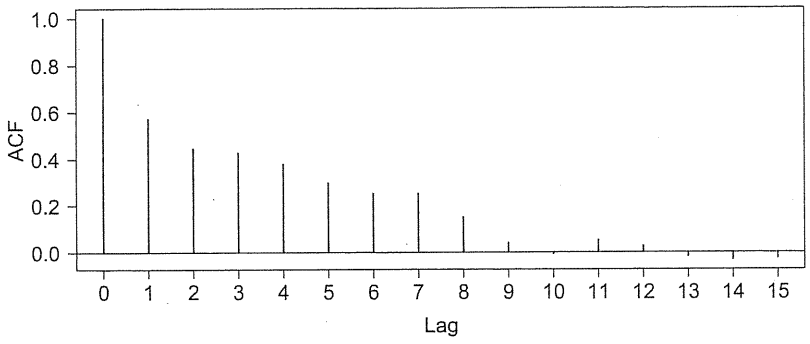

Figure 2.1 *Earthquakes series: sample autocorrelation function (ACF).*

Figure 2.1, gives a clear indication that the observations are serially dependent. One way of allowing for serial dependence in the observations is to relax the assumption that the parameter process is serially independent. A simple and mathematically convenient way to do so is to assume that it is a Markov chain. The resulting model for the observations is called a Poisson–hidden Markov model, a simple example of the class of models discussed in the rest of this book, namely hidden Markov models (HMMs).

We shall not give an account here of the (interesting) history of such models, but two valuable sources of information on HMMs that go far beyond the scope of this book, and include accounts of the history, are Ephraim and Merhav (2002) and Cappé, Moulines and Rydén (2005).

2.2 The basics

2.2.1 Definition and notation

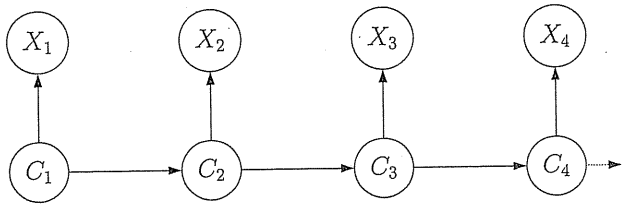

Figure 2.2 *Directed graph of basic HMM.*

A **hidden Markov model** $\{X_t : t \in \mathbb{N}\}$ is a particular kind of dependent mixture. With $\mathbf{X}^{(t)}$ and $\mathbf{C}^{(t)}$ representing the histories from time 1 to time t, one can summarize the simplest model of this kind by:

$$\Pr(C_t \mid \mathbf{C}^{(t-1)}) = \Pr(C_t \mid C_{t-1}), \ t = 2, 3, \ldots \tag{2.1}$$

$$\Pr(X_t \mid \mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = \Pr(X_t \mid C_t), \ t \in \mathbb{N}. \tag{2.2}$$

The model consists of two parts: firstly, an unobserved 'parameter process' $\{C_t : t = 1, 2, \ldots\}$ satisfying the Markov property, and secondly the 'state-dependent process' $\{X_t : t = 1, 2, \ldots\}$ such that, when C_t is known, the distribution of X_t depends only on the current state C_t and not on previous states or observations. This structure is represented by the directed graph in Figure 2.2. If the Markov chain $\{C_t\}$ has m states, we call $\{X_t\}$ an m-state HMM. Although it is the usual terminology in speech-processing applications, the name 'hidden Markov model' is by no means the only one used for such models or similar ones. For instance, Ephraim and Merhav (2002) argue for 'hidden Markov process', Leroux

and Puterman (1992) use 'Markov-dependent mixture', and others use 'Markov-switching model' (especially for models with extra dependencies at the level of the observations X_t), 'models subject to Markov regime', or 'Markov mixture model'.

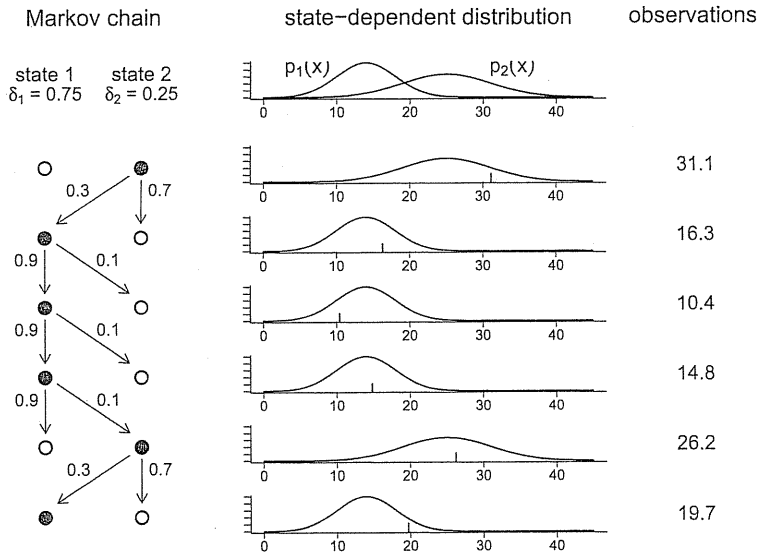

Figure 2.3 *Process generating the observations in a two-state HMM. The chain followed the path 2,1,1,1,2,1, as indicated on the left. The corresponding state-dependent distributions are shown in the middle. The observations are generated from the corresponding active distributions.*

The process generating the observations is demonstrated again in Figure 2.3, for state-dependent distributions p_1 and p_2, stationary distribution $\boldsymbol{\delta} = (0.75, 0.25)$, and t.p.m. $\boldsymbol{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$. In contrast to the case of an independent mixture, here the distribution of C_t, the state at time t, does depend on C_{t-1}. As is also true of independent mixtures, there is for each state a different distribution, discrete or continuous.

We now introduce some notation which will cover both discrete- and continuous-valued observations. In the case of discrete observations we define, for $i = 1, 2, \ldots, m$,

$$p_i(x) = \Pr(X_t = x \mid C_t = i).$$

That is, p_i is the probability mass function of X_t if the Markov chain is in state i at time t. The continuous case is treated similarly: there we define p_i to be the probability *density* function of X_t if the Markov

chain is in state i at time t. We refer to the m distributions p_i as the **state-dependent distributions** of the model. Many of our results are stated only in the discrete form, but, if probabilities are interpreted as densities, apply also to the continuous case.

2.2.2 Marginal distributions

We shall often need the distribution of X_t and also higher-order marginal distributions, such as that of (X_t, X_{t+k}). We shall derive the results for the case in which the Markov chain is homogeneous but not necessarily stationary, and then give them as well for the special case in which the Markov chain is stationary. For convenience the derivation is given only for discrete state-dependent distributions; the continuous case can be derived analogously.

Univariate distributions

For discrete-valued observations X_t, defining $u_i(t) = \Pr(C_t = i)$ for $t = 1, \ldots, T$, we have

$$\begin{aligned}\Pr(X_t = x) &= \sum_{i=1}^{m} \Pr(C_t = i) \Pr(X_t = x \mid C_t = i) \\ &= \sum_{i=1}^{m} u_i(t) p_i(x).\end{aligned}$$

This expression can conveniently be rewritten in matrix notation:

$$\begin{aligned}\Pr(X_t = x) &= (u_1(t), \ldots, u_m(t)) \begin{pmatrix} p_1(x) & & 0 \\ & \ddots & \\ 0 & & p_m(x) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \mathbf{u}(t)\mathbf{P}(x)\mathbf{1}',\end{aligned}$$

where $\mathbf{P}(x)$ is defined as the diagonal matrix with ith diagonal element $p_i(x)$. It follows from Equation (1.3) that $\mathbf{u}(t) = \mathbf{u}(1)\mathbf{\Gamma}^{t-1}$, and hence that

$$\Pr(X_t = x) = \mathbf{u}(1)\mathbf{\Gamma}^{t-1}\mathbf{P}(x)\mathbf{1}'. \tag{2.3}$$

Equation (2.3) holds if the Markov chain is merely homogeneous, and not necessarily stationary. If, as we shall often assume, the Markov chain is stationary, with stationary distribution $\boldsymbol{\delta}$, then the result is simpler: in that case $\boldsymbol{\delta}\mathbf{\Gamma}^{t-1} = \boldsymbol{\delta}$ for all $t \in \mathbb{N}$, and so

$$\Pr(X_t = x) = \boldsymbol{\delta}\mathbf{P}(x)\mathbf{1}'. \tag{2.4}$$

Bivariate distributions

The calculation of many of the distributions relating to an HMM is most easily done by first noting that, in any directed graphical model, the joint distribution of a set of random variables V_i is given by

$$\Pr(V_1, V_2, \ldots, V_n) = \prod_{i=1}^{n} \Pr(V_i \mid \text{pa}(V_i)), \tag{2.5}$$

where $\text{pa}(V_i)$ denotes all the 'parents' of V_i in the set $V_1, V_2, \ldots, V_n$; see e.g. Davison (2003, p. 250) or Jordan (2004).

Examining the directed graph of the four random variables X_t, X_{t+k}, C_t, C_{t+k}, for positive integer k, we see that $\text{pa}(C_t)$ is empty, $\text{pa}(X_t) = \{C_t\}$, $\text{pa}(C_{t+k}) = \{C_t\}$ and $\text{pa}(X_{t+k}) = \{C_{t+k}\}$. It therefore follows that

$$\Pr(X_t, X_{t+k}, C_t, C_{t+k}) = \Pr(C_t) \Pr(X_t|C_t) \Pr(C_{t+k}|C_t) \Pr(X_{t+k}|C_{t+k}),$$

and hence that

$$\begin{aligned}&\Pr(X_t = v, X_{t+k} = w) \\ &\quad= \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr(X_t = v, X_{t+k} = w, C_t = i, C_{t+k} = j) \\ &\quad= \sum_{i=1}^{m}\sum_{j=1}^{m} \underbrace{\Pr(C_t = i)}_{u_i(t)} p_i(v) \underbrace{\Pr(C_{t+k} = j \mid C_t = i)}_{\gamma_{ij(k)}} p_j(w) \\ &\quad= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t) p_i(v) \gamma_{ij}(k) p_j(w).\end{aligned}$$

Writing the above double sum as a product of matrices yields

$$\Pr(X_t = v, X_{t+k} = w) = \mathbf{u}(t)\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{1}'. \tag{2.6}$$

If the Markov chain is stationary, this reduces to

$$\Pr(X_t = v, X_{t+k} = w) = \boldsymbol{\delta}\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{1}'. \tag{2.7}$$

Similarly one can obtain expressions for the higher-order marginal distributions; in the stationary case, the formula for a trivariate distribution is, for positive integers k and l,

$$\Pr(X_t = v, X_{t+k} = w, X_{t+k+l} = z) = \boldsymbol{\delta}\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{\Gamma}^l\mathbf{P}(z)\mathbf{1}'.$$

2.2.3 Moments

First we note that

$$\mathrm{E}(X_t) = \sum_{i=1}^{m} \mathrm{E}(X_t \mid C_t = i) \Pr(C_t = i) = \sum_{i=1}^{m} u_i(t) \mathrm{E}(X_t \mid C_t = i),$$

which, in the stationary case, reduces to

$$\mathrm{E}(X_t) = \sum_{i=1}^{m} \delta_i \mathrm{E}(X_t \mid C_t = i).$$

More generally, analogous results hold for $\mathrm{E}(g(X_t))$ and $\mathrm{E}(g(X_t, X_{t+k}))$, for any functions g for which the relevant state-dependent expectations exist. In the stationary case

$$\mathrm{E}(g(X_t)) = \sum_{i=1}^{m} \delta_i \mathrm{E}(g(X_t) \mid C_t = i); \tag{2.8}$$

and

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g(X_t, X_{t+k}) \mid C_t = i, C_{t+k} = j)\, \delta_i \gamma_{ij}(k), \tag{2.9}$$

where $\gamma_{ij}(k) = (\mathbf{\Gamma}^k)_{ij}$, for $k \in \mathbb{N}$. Often we shall be interested in a function g which factorizes as $g(X_t, X_{t+k}) = g_1(X_t) g_2(X_{t+k})$, in which case Equation (2.9) becomes

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g_1(X_t) \mid C_t = i)\, \mathrm{E}(g_2(X_{t+k}) \mid C_{t+k} = j)\, \delta_i \gamma_{ij}(k). \tag{2.10}$$

These expressions enable us, for instance, to find covariances and correlations without too much trouble; convenient explicit expressions exist in many cases. For instance, the following conclusions result in the case of a stationary two-state Poisson–HMM:

- $\mathrm{E}(X_t) = \delta_1 \lambda_1 + \delta_2 \lambda_2$;
- $\mathrm{Var}(X_t) = \mathrm{E}(X_t) + \delta_1 \delta_2 (\lambda_2 - \lambda_1)^2 \geq \mathrm{E}(X_t)$;
- $\mathrm{Cov}(X_t, X_{t+k}) = \delta_1 \delta_2 (\lambda_2 - \lambda_1)^2 (1 - \gamma_{12} - \gamma_{21})^k$, for $k \in \mathbb{N}$.

Notice that the resulting formula for the correlation of X_t and X_{t+k} is of the form $\rho(k) = A(1 - \gamma_{12} - \gamma_{21})^k$ with $A \in [0, 1)$, and that $A = 0$ if $\lambda_1 = \lambda_2$. For more details, and for more general results, see Exercises 3 and 4.

2.3 The likelihood

The aim of this section is to develop an explicit (and computable) formula for the likelihood L_T of T consecutive observations $x_1, x_2, \ldots, x_T$ assumed to be generated by an m-state HMM. That such a formula exists is indeed fortunate, but by no means obvious. We shall see that the computation of the likelihood, consisting as it does of a sum of m^T terms, each of which is a product of $2T$ factors, appears to require $O(Tm^T)$ operations, and several authors have come to the conclusion that straightforward calculation of the likelihood is infeasible. However, it has long been known in several contexts that the likelihood is computable: see e.g. Baum (1972), Lange and Boehnke (1983), and Cosslett and Lee (1985). What we describe here is in fact a special case of a much more general theory: see Smyth, Heckerman and Jordan (1997) or Jordan (2004).

It is our purpose here to demonstrate that L_T can in general be computed relatively simply in $O(Tm^2)$ operations. Once it is clear that the likelihood is simple to compute, the way will be open to estimate parameters by numerical maximization of the likelihood.

First the likelihood of a two-state model will be explored, and then the general formula will be presented.

2.3.1 The likelihood of a two-state Bernoulli–HMM

Example: Consider the two-state HMM with t.p.m.

$$\mathbf{\Gamma} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

and state-dependent distributions given by

$$\Pr(X_t = x \mid C_t = 1) = \frac{1}{2} \quad (\text{for } x = 0, 1)$$

and

$$\Pr(X_t = 1 \mid C_t = 2) = 1.$$

We call a model of this kind a Bernoulli–HMM. The stationary distribution of the Markov chain is $\boldsymbol{\delta} = \frac{1}{3}(1, 2)$. Then the probability that $X_1 = X_2 = X_3 = 1$ can be calculated as follows. First, note that, by Equation (2.5),

$$\begin{aligned} &\Pr(X_1, X_2, X_3, C_1, C_2, C_3) \\ &= \Pr(C_1) \Pr(X_1 \mid C_1) \Pr(C_2 \mid C_1) \Pr(X_2 \mid C_2) \Pr(C_3 \mid C_2) \Pr(X_3 \mid C_3); \end{aligned}$$

Table 2.1 *Example of a likelihood computation.*

i	j	k	$p_i(1)$	$p_j(1)$	$p_k(1)$	δ_i	γ_{ij}	γ_{jk}	product
1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{96}$
1	1	2	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{48}$
1	2	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{96}$
1	2	2	$\frac{1}{2}$	1	1	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
2	1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{48}$
2	1	2	1	$\frac{1}{2}$	1	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{24}$
2	2	1	1	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{16}$
2	2	2	1	1	1	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{4}$	$\frac{3}{8}$
									$\frac{29}{48}$

and then sum over the values assumed by C_1, C_2, C_3. The result is

$$\begin{aligned}
&\Pr(X_1 = 1, X_2 = 1, X_3 = 1) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \Pr(X_1 = 1, X_2 = 1, X_3 = 1, C_1 = i, C_2 = j, C_3 = k) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \delta_i p_i(1)\gamma_{ij} p_j(1)\gamma_{jk} p_k(1). \qquad (2.11)
\end{aligned}$$

Notice that the triple sum (2.11) has $m^T = 2^3$ terms, each of which is a product of $2T = 2 \times 3$ factors. To evaluate the required probability, the different possibilities for the values of i, j and k can be listed and the sum (2.11) calculated as in Table 2.1.

Summation of the last column of Table 2.1 tells us that $\Pr(X_1 = 1, X_2 = 1, X_3 = 1) = \frac{29}{48}$. In passing we note that the largest element in the last column is $\frac{3}{8}$; the state sequence ijk that maximizes the joint probability

$$\Pr(X_1 = 1, X_2 = 1, X_3 = 1, C_1 = i, C_2 = j, C_3 = k)$$

is therefore the sequence 222. Equivalently, it maximizes the conditional probability $\Pr(C_1 = i, C_2 = j, C_3 = k \mid X_1 = 1, X_2 = 1, X_3 = 1)$. This is an example of 'global decoding', which will be discussed in Section 5.3.2: see p. 82.

But a more convenient way to present the sum is to use matrix nota-

tion. Let $\mathbf{P}(u)$ be defined (as before) as $\mathrm{diag}(p_1(u), p_2(u))$. Then

$$\mathbf{P}(0) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{P}(1) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix},$$

and the triple sum (2.11) can be written as

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \delta_i p_i(1)\gamma_{ij} p_j(1)\gamma_{jk} p_k(1) = \boldsymbol{\delta}\mathbf{P}(1)\boldsymbol{\Gamma}\mathbf{P}(1)\boldsymbol{\Gamma}\mathbf{P}(1)\mathbf{1}'.$$

2.3.2 The likelihood in general

Here we consider the likelihood of an HMM in general. We suppose there is an observation sequence $x_1, x_2, \ldots, x_T$ generated by such a model. We seek the probability L_T of observing that sequence, as calculated under an m-state HMM which has *initial* distribution $\boldsymbol{\delta}$ and t.p.m. $\boldsymbol{\Gamma}$ for the Markov chain, and state-dependent probability (density) functions p_i. In many of our applications we shall assume that $\boldsymbol{\delta}$ is the stationary distribution implied by $\boldsymbol{\Gamma}$, but it is not necessary to make that assumption in general.

Proposition 1 *The likelihood is given by*

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'. \qquad (2.12)$$

If $\boldsymbol{\delta}$, the distribution of C_1, is the stationary distribution of the Markov chain, then in addition

$$L_T = \boldsymbol{\delta}\boldsymbol{\Gamma}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'. \qquad (2.13)$$

Before proving the above proposition, we rewrite the conclusions in a notation which is sometimes useful. For $t = 1, \ldots, T$, let the matrix $\mathbf{B}_t$ be defined by $\mathbf{B}_t = \boldsymbol{\Gamma}\mathbf{P}(x_t)$. Equations (2.12) and (2.13) (respectively) can then be written as

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{B}_2\mathbf{B}_3\cdots\mathbf{B}_T\mathbf{1}'$$

and

$$L_T = \boldsymbol{\delta}\mathbf{B}_1\mathbf{B}_2\mathbf{B}_3\cdots\mathbf{B}_T\mathbf{1}'.$$

Note that in the first of these equations $\boldsymbol{\delta}$ represents the initial distribution of the Markov chain, and in the second the stationary distribution.

Proof. We present only the case of discrete observations. First note that

$$L_T = \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \sum_{c_1, c_2, \ldots, c_T = 1}^{m} \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, \mathbf{C}^{(T)} = \mathbf{c}^{(T)}),$$

and that, by Equation (2.5),

$$\Pr(\mathbf{X}^{(T)}, \mathbf{C}^{(T)}) = \Pr(C_1) \prod_{k=2}^{T} \Pr(C_k \mid C_{k-1}) \prod_{k=1}^{T} \Pr(X_k \mid C_k). \quad (2.14)$$

It follows that

$$\begin{aligned} L_T &= \sum_{c_1,\ldots,c_T=1}^{m} \left(\delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_3}\cdots\gamma_{c_{T-1},c_T}\right)\left(p_{c_1}(x_1)p_{c_2}(x_2)\cdots p_{c_T}(x_T)\right) \\ &= \sum_{c_1,\ldots,c_T=1}^{m} \delta_{c_1}p_{c_1}(x_1)\gamma_{c_1,c_2}p_{c_2}(x_2)\gamma_{c_2,c_3}\cdots\gamma_{c_{T-1},c_T}p_{c_T}(x_T) \\ &= \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}', \end{aligned}$$

i.e. Equation (2.12). If $\boldsymbol{\delta}$ is the stationary distribution of the Markov chain, we have $\boldsymbol{\delta}\mathbf{P}(x_1) = \boldsymbol{\delta}\boldsymbol{\Gamma}\mathbf{P}(x_1) = \boldsymbol{\delta}\mathbf{B}_1$, hence Equation (2.13), which involves an extra factor of $\boldsymbol{\Gamma}$ but may be slightly simpler to code. □

In order to set out the likelihood computation in the form of an algorithm, let us now define the vector $\boldsymbol{\alpha}_t$, for $t = 1, 2, \ldots, T$, by

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\mathbf{P}(x_3)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_t) = \boldsymbol{\delta}\mathbf{P}(x_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\mathbf{P}(x_s), \quad (2.15)$$

with the convention that an empty product is the identity matrix. It follows immediately from this definition that

$$L_T = \boldsymbol{\alpha}_T\mathbf{1}', \quad \text{and} \quad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for } t \geq 2.$$

Accordingly, we can conveniently set out as follows the computations involved in the likelihood formula (2.12):

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(x_1);$$
$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for } t = 2, 3, \ldots, T;$$
$$L_T = \boldsymbol{\alpha}_T\mathbf{1}'.$$

That the number of operations involved is of order Tm^2 can be deduced thus. For each of the values of t in the loop, there are m elements of $\boldsymbol{\alpha}_t$ to be computed, and each of those elements is a sum of m products of three quantities: an element of $\boldsymbol{\alpha}_{t-1}$, a transition probability γ_{ij}, and a state-dependent probability (or density) $p_j(x_t)$.

The corresponding scheme for computation of (2.13) (i.e. if $\boldsymbol{\delta}$, the distribution of C_1, is the stationary distribution of the Markov chain) is

$$\boldsymbol{\alpha}_0 = \boldsymbol{\delta};$$
$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \quad \text{for } t = 1, 2, \ldots, T;$$
$$L_T = \boldsymbol{\alpha}_T\mathbf{1}'.$$

The elements of $\boldsymbol{\alpha}_t$ are usually referred to as **forward probabilities**; the reason for this name will appear only later, in Section 4.1.1.

HMMs are not Markov processes

HMMs do not in general satisfy the Markov property. This we can now establish via a simple counterexample. Let X_t and C_t be as defined in the example in Section 2.3.1. We already know that

$$\Pr(X_1 = 1, X_2 = 1, X_3 = 1) = \frac{29}{48},$$

and from the above general expression for the likelihood, or otherwise, it can be established that $\Pr(X_2 = 1) = \frac{5}{6}$, and that

$$\Pr(X_1 = 1, X_2 = 1) = \Pr(X_2 = 1, X_3 = 1) = \frac{17}{24}.$$

It therefore follows that

$$\begin{aligned} \Pr(X_3 = 1 \mid X_1 = 1, X_2 = 1) &= \frac{\Pr(X_1 = 1, X_2 = 1, X_3 = 1)}{\Pr(X_1 = 1, X_2 = 1)} \\ &= \frac{29/48}{17/24} = \frac{29}{34}, \end{aligned}$$

and that

$$\begin{aligned} \Pr(X_3 = 1 \mid X_2 = 1) &= \frac{\Pr(X_2 = 1, X_3 = 1)}{\Pr(X_2 = 1)} \\ &= \frac{17/24}{5/6} = \frac{17}{20}. \end{aligned}$$

Hence $\Pr(X_3 = 1 \mid X_2 = 1) \neq \Pr(X_3 = 1 \mid X_1 = 1, X_2 = 1)$; this HMM does not satisfy the Markov property. That some HMMs do satisfy the property, however, is clear. For instance, a two-state Bernoulli–HMM can degenerate in obvious fashion to the underlying Markov chain; one simply identifies each of the two observable values with one of the two underlying states. For the conditions under which an HMM will itself satisfy the Markov property, see Spreij (2001).

2.3.3 The likelihood when data are missing at random

In a time series context it is potentially awkward if some of the data are missing. In the case of hidden Markov time series models, however, the adjustment that needs to be made to the likelihood computation if data are missing turns out to be a simple one.

Suppose, for example, that one has available the observations x_1, x_2, x_4, x_7, x_8, …, x_T of an HMM, but the observations x_3, x_5 and x_6 are

missing at random. Then the likelihood is given by

$$\begin{aligned}\Pr(X_1 = x_1, X_2 = x_2, X_4 = x_4, X_7 = x_7, \ldots, X_T = x_T) \\ = \sum \delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_4}(2)\gamma_{c_4,c_7}(3)\gamma_{c_7,c_8}\cdots\gamma_{c_{T-1},c_T} \\ \times p_{c_1}(x_1)p_{c_2}(x_2)p_{c_4}(x_4)p_{c_7}(x_7)\cdots p_{c_T}(x_T),\end{aligned}$$

where (as before) $\gamma_{ij}(k)$ denotes a k-step transition probability, and the sum is taken over all c_t other than c_3, c_5 and c_6. But this is just

$$\begin{aligned}\sum \delta_{c_1}p_{c_1}(x_1)\gamma_{c_1,c_2}p_{c_2}(x_2)\gamma_{c_2,c_4}(2)p_{c_4}(x_4)\gamma_{c_4,c_7}(3)p_{c_7}(x_7) \\ \cdots \times \gamma_{c_{T-1},c_T}p_{c_T}(x_T) \\ = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}^2\mathbf{P}(x_4)\boldsymbol{\Gamma}^3\mathbf{P}(x_7)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'.\end{aligned}$$

With $L_T^{-(3,5,6)}$ denoting the likelihood of the observations other than x_3, x_5 and x_6, our conclusion is therefore that

$$L_T^{-(3,5,6)} = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}^2\mathbf{P}(x_4)\boldsymbol{\Gamma}^3\mathbf{P}(x_7)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'.$$

The easiest way to summarize this conclusion is to say that, in the expression for the likelihood, the diagonal matrices $\mathbf{P}(x_t)$ corresponding to missing observations x_t are replaced by the identity matrix; equivalently, the corresponding state-dependent probabilities $p_i(x_t)$ are replaced by 1 for all states i.

The fact that, even in the case of missing observations, the likelihood of an HMM can be easily computed is especially useful in the derivation of conditional distributions, as will be shown in Section 5.1.

2.3.4 The likelihood when observations are interval-censored

Suppose that we wish to fit a Poisson–HMM to a series of counts, some of which are interval-censored. For instance, the exact value of x_t may be known only for $4 \leq t \leq T$, with the information $x_1 \leq 5$, $2 \leq x_2 \leq 3$ and $x_3 > 10$ available about the remaining observations. For simplicity, let us first assume that the Markov chain has only two states. In that case, one replaces the diagonal matrix $\mathbf{P}(x_1)$ in the likelihood expression (2.12) by the matrix

$$\text{diag}(\Pr(X_1 \leq 5 \mid C_1 = 1), \Pr(X_1 \leq 5 \mid C_1 = 2)),$$

and similarly for $\mathbf{P}(x_2)$ and $\mathbf{P}(x_3)$.

More generally, suppose that $a \leq x_t \leq b$, where a may be $-\infty$ (although that is not relevant to the Poisson case), b may be ∞, and the Markov chain has m states. One replaces $\mathbf{P}(x_t)$ in the likelihood by the $m \times m$ diagonal matrix of which the ith diagonal element is $\Pr(a \leq X_t \leq b \mid C_t = i)$. See Exercise 12. It is worth noting that missing data can be regarded as an extreme case of such interval-censoring.

Exercises

1. Consider a stationary two-state Poisson–HMM with parameters

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.1 & 0.9 \\ 0.4 & 0.6 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\lambda} = (1, 3).$$

In each of the following ways, compute the probability that the first three observations from this model are 0, 2, 1.

(a) Consider all possible sequences of states of the Markov chain that could have occurred. Compute the probability of each sequence, and the probability of the observations given each sequence.

(b) Apply the formula

$$\Pr(X_1 = 0, X_2 = 2, X_3 = 1) = \boldsymbol{\delta}\mathbf{P}(0)\boldsymbol{\Gamma}\mathbf{P}(2)\boldsymbol{\Gamma}\mathbf{P}(1)\mathbf{1}',$$

where

$$\mathbf{P}(s) = \begin{pmatrix} \lambda_1^s e^{-\lambda_1}/s! & 0 \\ 0 & \lambda_2^s e^{-\lambda_2}/s! \end{pmatrix} = \begin{pmatrix} 1^s e^{-1}/s! & 0 \\ 0 & 3^s e^{-3}/s! \end{pmatrix}.$$

2. Consider again the model defined in Exercise 1. In that question you were asked to compute $\Pr(X_1 = 0, X_2 = 2, X_3 = 1)$. Now compute $\Pr(X_1 = 0, X_3 = 1)$ in each of the following ways.

(a) Consider all possible sequences of states of the Markov chain that could have occurred. Compute the probability of each sequence, and the probability of the observations given each sequence.

(b) Apply the formula

$$\Pr(X_1 = 0, X_3 = 1) = \boldsymbol{\delta}\mathbf{P}(0)\boldsymbol{\Gamma}\mathbf{I}_2\boldsymbol{\Gamma}\mathbf{P}(1)\mathbf{1}' = \boldsymbol{\delta}\mathbf{P}(0)\boldsymbol{\Gamma}^2\mathbf{P}(1)\mathbf{1}',$$

and check that this probability is equal to your answer in (a).

3. Consider an m-state HMM $\{X_t : t = 1, 2, \ldots\}$, based on a stationary Markov chain with transition probability matrix $\boldsymbol{\Gamma}$ and stationary distribution $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$, and having (univariate) state-dependent distributions $p_i(x)$. Let μ_i and σ_i^2 denote the mean and variance of the distribution p_i, $\boldsymbol{\mu}$ the vector $(\mu_1, \mu_2, \ldots, \mu_m)$, and $\mathbf{M}$ the matrix $\text{diag}(\boldsymbol{\mu})$.

Derive the following results for the moments of $\{X_t\}$. (Sometimes, but not always, it is useful to express such results in matrix form.)

(a) $\mathrm{E}(X_t) = \sum_{i=1}^m \delta_i\mu_i = \boldsymbol{\delta}\boldsymbol{\mu}'$.

(b) $\mathrm{E}(X_t^2) = \sum_{i=1}^m \delta_i(\sigma_i^2 + \mu_i^2)$.

(c) $\mathrm{Var}(X_t) = \sum_{i=1}^m \delta_i(\sigma_i^2 + \mu_i^2) - (\boldsymbol{\delta}\boldsymbol{\mu}')^2$.

(d) If $m = 2$, $\mathrm{Var}(X_t) = \delta_1\sigma_1^2 + \delta_2\sigma_2^2 + \delta_1\delta_2(\mu_1 - \mu_2)^2$.

(e) For $k \in \mathbb{N}$, i.e. for positive integers k,
$\mathrm{E}(X_t X_{t+k}) = \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_i \mu_i \gamma_{ij}(k) \mu_j = \boldsymbol{\delta} \mathbf{M} \boldsymbol{\Gamma}^k \boldsymbol{\mu}'$.

(f) For $k \in \mathbb{N}$,

$$\rho(k) = \mathrm{Corr}(X_t, X_{t+k}) = \frac{\boldsymbol{\delta} \mathbf{M} \boldsymbol{\Gamma}^k \boldsymbol{\mu}' - (\boldsymbol{\delta}\boldsymbol{\mu}')^2}{\mathrm{Var}(X_t)}.$$

Note that, if the eigenvalues of $\boldsymbol{\Gamma}$ are distinct, this is a linear combination of the kth powers of those eigenvalues.

Timmermann (2000) and Frühwirth-Schnatter (2006, pp. 308–312) are useful references for moments.

4. (Marginal moments and autocorrelation function of a Poisson–HMM: special case of Exercise 3.) Consider a stationary m-state Poisson–HMM $\{X_t : t = 1, 2, \ldots\}$ with transition probability matrix $\boldsymbol{\Gamma}$ and state-dependent means $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)$. Let $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$ be the stationary distribution of the Markov chain. Let $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$. Derive the following results.

(a) $\mathrm{E}(X_t) = \boldsymbol{\delta}\boldsymbol{\lambda}'$.
(b) $\mathrm{E}(X_t^2) = \sum_{i=1}^{m} (\lambda_i^2 + \lambda_i)\delta_i = \boldsymbol{\delta}\boldsymbol{\Lambda}\boldsymbol{\lambda}' + \boldsymbol{\delta}\boldsymbol{\lambda}'$.
(c) $\mathrm{Var}(X_t) = \boldsymbol{\delta}\boldsymbol{\Lambda}\boldsymbol{\lambda}' + \boldsymbol{\delta}\boldsymbol{\lambda}' - (\boldsymbol{\delta}\boldsymbol{\lambda}')^2 = \mathrm{E}(X_t) + \boldsymbol{\delta}\boldsymbol{\Lambda}\boldsymbol{\lambda}' - (\boldsymbol{\delta}\boldsymbol{\lambda}')^2 \geq \mathrm{E}(X_t)$.
(d) For $k \in \mathbb{N}$, $\mathrm{E}(X_t X_{t+k}) = \boldsymbol{\delta}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^k\boldsymbol{\lambda}'$.
(e) For $k \in \mathbb{N}$,

$$\rho(k) = \mathrm{Corr}(X_t, X_{t+k}) = \frac{\boldsymbol{\delta}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^k\boldsymbol{\lambda}' - (\boldsymbol{\delta}\boldsymbol{\lambda}')^2}{\boldsymbol{\delta}\boldsymbol{\Lambda}\boldsymbol{\lambda}' + \boldsymbol{\delta}\boldsymbol{\lambda}' - (\boldsymbol{\delta}\boldsymbol{\lambda}')^2}.$$

(f) For the case $m = 2$, $\rho(k) = Aw^k$, where

$$A = \frac{\delta_1\delta_2(\lambda_2 - \lambda_1)^2}{\delta_1\delta_2(\lambda_2 - \lambda_1)^2 + \boldsymbol{\delta}\boldsymbol{\lambda}'}$$

and $w = 1 - \gamma_{12} - \gamma_{21}$.

5. Consider the three-state Poisson–HMM $\{X_t\}$ with state-dependent means λ_i $(i = 1, 2, 3)$ and transition probability matrix

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

Assume that the Markov chain is stationary.

Show that the autocorrelation $\rho(k) = \mathrm{Corr}(X_t, X_{t+k})$ is given by

$$\frac{(-\frac{1}{3})^k\left\{3(-5\lambda_1-3\lambda_2+8\lambda_3)^2+180(\lambda_2-\lambda_1)^2\right\}+k(-\frac{1}{3})^{k-1}\left\{4(-5\lambda_1-3\lambda_2+8\lambda_3)(\lambda_2-\lambda_1)\right\}}{32\left\{15(\lambda_1^2+\lambda_1)+9(\lambda_2^2+\lambda_2)+8(\lambda_3^2+\lambda_3)\right\}-(15\lambda_1+9\lambda_2+8\lambda_3)^2}.$$

Notice that, as a function of k, this (rather tedious!) expression is a linear combination of $(-\frac{1}{3})^k$ and $k(-\frac{1}{3})^{k-1}$. (This is an example of a non-diagonalizable t.p.m. In practice such cases are not likely to be of interest.)

6. We have the general expression

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'$$

for the likelihood of an HMM, e.g. of Poisson type. Consider the special case in which the Markov chain degenerates to a sequence of independent random variables, i.e. an independent mixture model.

Show that, in this case, the likelihood simplifies to the expression given in Equation (1.1) for the likelihood of an *independent* mixture.

7. Consider a multiple sum S of the following general form:

$$S = \sum_{i_1=1}^{m}\sum_{i_2=1}^{m}\cdots\sum_{i_T=1}^{m} f_1(i_1)\prod_{t=2}^{T} f_t(i_{t-1}, i_t).$$

For $i_1 = 1, 2, \ldots, m$, define

$$\alpha_1(i_1) \equiv f_1(i_1);$$

and for $r = 1, 2, \ldots, T-1$ and $i_{r+1} = 1, 2, \ldots, m$, define

$$\alpha_{r+1}(i_{r+1}) \equiv \sum_{i_r=1}^{m} \alpha_r(i_r)\, f_{r+1}(i_r, i_{r+1}).$$

That is, the row vector $\boldsymbol{\alpha}_{r+1}$ is defined by, and can be computed as, $\boldsymbol{\alpha}_{r+1} = \boldsymbol{\alpha}_r\, \mathbf{F}_{r+1}$, where the $m \times m$ matrix $\mathbf{F}_t$ has (i, j) element equal to $f_t(i, j)$.

(a) Show by induction that $\alpha_T(i_T)$ is precisely the sum over all but i_T, i.e. that

$$\alpha_T(i_T) = \sum_{i_1}\sum_{i_2}\cdots\sum_{i_{T-1}} f_1(i_1)\prod_{t=2}^{T} f_t(i_{t-1}, i_t).$$

(b) Hence show that $S = \sum_{i_T} \alpha_T(i_T) = \boldsymbol{\alpha}_T\mathbf{1}' = \boldsymbol{\alpha}_1\mathbf{F}_2\mathbf{F}_3\cdots\mathbf{F}_T\mathbf{1}'$.
(c) Does this result generalize to nonconstant m?

8.(a) In Section 1.3.3 we defined reversibility for a random process, and showed that the stationary Markov chain with the t.p.m. $\boldsymbol{\Gamma}$ given below is not reversible.

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

Now let $\{X_t\}$ be the stationary HMM with $\mathbf{\Gamma}$ as above, and having Poisson state-dependent distributions with means 1, 5 and 10; e.g. in state 1 the observation X_t is distributed Poisson with mean 1.

By finding the probabilities $\Pr(X_t = 0, X_{t+1} = 1)$ and $\Pr(X_t = 1, X_{t+1} = 0)$, or otherwise, show that $\{X_t\}$ is irreversible.

(b) Show that, if the Markov chain underlying a stationary HMM is reversible, the HMM is also reversible.

(c) Suppose that the Markov chain underlying a stationary HMM is irreversible. Does it follow that the HMM is irreversible?

9. Write a function `pois-HMM.moments(m,lambda,gamma,lag.max=10)` that computes the expectation, variance and autocorrelation function (for lags 0 to `lag.max`) of an `m`-state stationary Poisson–HMM with t.p.m. `gamma` and state-dependent means `lambda`.

10. Write the three functions listed below, relating to the marginal distribution of an `m`-state Poisson–HMM with parameters `lambda`, `gamma`, and possibly `delta`. In each case, if `delta` is not specified, the stationary distribution should be used. You can use your function `statdist` (see Exercise 8(b) of Chapter 1) to provide the stationary distribution.

```
dpois.HMM(x, m, lambda, gamma, delta=NULL)
ppois.HMM(x, m, lambda, gamma, delta=NULL)
dpois.HMM(p, m, lambda, gamma, delta=NULL)
```

The function `dpois.HMM` computes the probability function at the arguments specified by the vector `x`, `ppois.HMM` the distribution function, and `ppois.HMM` the inverse distribution function.

11. Consider the function `pois.HMM.generate_sample` in A.2.1 that generates observations from a stationary m-state Poisson–HMM. Test the function by generating a long sequence of observations (10 000, say), and then check whether the sample mean, variance, ACF and relative frequencies correspond to what you expect.

12. Interval-censored observations

(a) Suppose that, in a series of unbounded counts $x_1, \ldots, x_T$, only the observation x_t is interval-censored, and $a \leq x_t \leq b$, where b may be ∞.

Prove the statement made in Section 2.3.4 that the likelihood of a Poisson–HMM with m states is obtained by replacing $\mathbf{P}(x_t)$ in the expression (2.12) by the $m \times m$ diagonal matrix of which the ith diagonal element is $\Pr(a \leq X_t \leq b \mid C_t = i)$.

(b) Extend part (a) to allow for any number of interval-censored observations.

CHAPTER 3

Estimation by direct maximization of the likelihood

3.1 Introduction

We saw in Equation (2.12) that the likelihood of an HMM is given by

$$L_T = \Pr\left(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) = \boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{\Gamma}\mathbf{P}(x_2)\cdots\mathbf{\Gamma}\mathbf{P}(x_T)\mathbf{1}',$$

where $\boldsymbol{\delta}$ is the initial distribution (that of C_1) and $\mathbf{P}(x)$ the $m \times m$ diagonal matrix with ith diagonal element the state-dependent probability or density $p_i(x)$. In principle we can therefore compute $L_T = \boldsymbol{\alpha}_T\mathbf{1}'$ recursively via

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(x_1)$$

and

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\mathbf{\Gamma}\mathbf{P}(x_t) \quad \text{for } t = 2, 3, \ldots, T.$$

If the Markov chain is assumed to be stationary (in which case $\boldsymbol{\delta} = \boldsymbol{\delta}\mathbf{\Gamma}$), we can choose to use instead

$$\boldsymbol{\alpha}_0 = \boldsymbol{\delta}$$

and

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\mathbf{\Gamma}\mathbf{P}(x_t) \quad \text{for } t = 1, 2, \ldots, T.$$

We shall first consider the stationary case.

The number of operations involved is of order Tm^2, making the evaluation of the likelihood quite feasible even for large T. Parameter estimation can therefore be performed by numerical maximization of the likelihood with respect to the parameters.

But there are several problems that need to be addressed when the likelihood is computed in this way and maximized numerically in order to estimate parameters. The main problems are numerical underflow, constraints on the parameters, and multiple local maxima in the likelihood function. In this chapter we first discuss how to overcome these problems, in order to arrive at a general strategy for computing MLEs which is easy to implement. Then we discuss the estimation of standard errors for parameters. We defer to the next chapter the EM algorithm, which necessitates some discussion of the forward and backward probabilities.

3.2 Scaling the likelihood computation

In the case of discrete state-dependent distributions, the elements of $\boldsymbol{\alpha}_t$, being made up of products of probabilities, become progressively smaller as t increases, and are eventually rounded to zero. In fact, with probability 1 the likelihood approaches 0 or ∞ exponentially fast; see Leroux and Puterman (1992). The problem is therefore not confined to the discrete case and underflow; overflow may occur in a continuous case. The remedy is, however, the same for over- and underflow, and we confine our attention to underflow.

Since the likelihood is a product of matrices, not of scalars, it is not possible to circumvent numerical underflow simply by computing the log of the likelihood as the sum of logs of its factors. In this respect the computation of the likelihood of an independent mixture model is simpler than that of an HMM.

To solve the problem, Durbin *et al.* (1998, p. 78) suggest (*inter alia*) a method of computation that relies on the following approximation. Suppose we wish to compute $\log(p+q)$, where $p > q$. Write $\log(p+q)$ as

$$\log p + \log(1 + q/p) = \log p + \log(1 + \exp(\tilde{q} - \tilde{p})),$$

where $\tilde{p} = \log p$ and $\tilde{q} = \log q$. The function $\log(1 + e^x)$ is then approximated by interpolation from a table of its values; apparently quite a small table will give a reasonable degree of accuracy.

We prefer to compute the logarithm of L_T using a strategy of scaling the vector of forward probabilities $\boldsymbol{\alpha}_t$. Define, for $t = 0, 1, \ldots, T$, the vector

$$\boldsymbol{\phi}_t = \boldsymbol{\alpha}_t / w_t,$$

where $w_t = \sum_i \alpha_t(i) = \boldsymbol{\alpha}_t \mathbf{1}'$.

First we note certain immediate consequences of the definitions of $\boldsymbol{\phi}_t$ and w_t:

$$\begin{aligned} w_0 = \boldsymbol{\alpha}_0 \mathbf{1}' &= \boldsymbol{\delta}\mathbf{1}' = 1; \\ \boldsymbol{\phi}_0 &= \boldsymbol{\delta}; \\ w_t \boldsymbol{\phi}_t &= w_{t-1}\boldsymbol{\phi}_{t-1}\mathbf{B}_t; \\ L_T = \boldsymbol{\alpha}_T \mathbf{1}' &= w_T(\boldsymbol{\phi}_T \mathbf{1}') = w_T. \end{aligned} \tag{3.1}$$

Hence $L_T = w_T = \prod_{t=1}^{T}(w_t/w_{t-1})$. From (3.1) it follows that

$$w_t = w_{t-1}(\boldsymbol{\phi}_{t-1}\mathbf{B}_t\mathbf{1}'),$$

and so we conclude that

$$\log L_T = \sum_{t=1}^{T} \log(w_t/w_{t-1}) = \sum_{t=1}^{T} \log(\boldsymbol{\phi}_{t-1}\mathbf{B}_t\mathbf{1}').$$

The computation of the log-likelihood is summarized below in the form

of an algorithm. Note that $\boldsymbol{\Gamma}$ and $\mathbf{P}(x_t)$ are $m \times m$ matrices, $\mathbf{v}$ and $\boldsymbol{\phi}_t$ are vectors of length m, u is a scalar, and l is the scalar in which the log-likelihood is accumulated.

set $\boldsymbol{\phi}_0 \leftarrow \boldsymbol{\delta}$ and $l \leftarrow 0$
for $t = 1, 2, \ldots, T$
  $\mathbf{v} \leftarrow \boldsymbol{\phi}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t)$
  $u \leftarrow \mathbf{v}\mathbf{1}'$
  $l \leftarrow l + \log u$
  $\boldsymbol{\phi}_t \leftarrow \mathbf{v}/u$
return l

The required log-likelihood, $\log L_T$, is then given by the final value of l. This procedure will avoid underflow in many cases. Clearly, variations of the technique are possible: for instance, the scale factor w_t could be chosen instead to be the largest element of the vector being scaled, or the mean of its elements (as opposed to the sum). See A.1.3 (in Appendix A) for an implementation of the above algorithm.

The algorithm is easily modified to compute the log-likelihood without assuming stationarity of the Markov chain. If $\boldsymbol{\delta}$ is the initial distribution, replace the first two lines above by

set $w_1 \leftarrow \boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{1}'$, $\boldsymbol{\phi}_1 \leftarrow \boldsymbol{\delta}\mathbf{P}(x_1)/w_1$ and $l \leftarrow \log w_1$
for $t = 2, 3, \ldots, T$

Of course, if the initial distribution happens to be the stationary distribution, the more general algorithm still applies.

3.3 Maximization of the likelihood subject to constraints

3.3.1 Reparametrization to avoid constraints

The elements of $\boldsymbol{\Gamma}$ and those of $\boldsymbol{\lambda}$, the vector of state-dependent means in a Poisson–HMM, do not range over the whole of $\mathbb{R}$, the set of all real numbers. Neither therefore should any sensible estimates of the parameters. In particular, the row sums of $\boldsymbol{\Gamma}$, and any estimate thereof, should equal one. Thus when maximizing the likelihood we have a constrained optimization problem to solve, not an unconstrained one.

Special-purpose software, e.g. NPSOL (Gill *et al.*, 1986) or the corresponding NAG routine E04UCF, can be used to maximize a function of several variables which are subject to constraints. The advice of Gill, Murray and Wright (1981, p. 267) is that it is 'rarely appropriate to alter linearly constrained problems'. However — depending on the implementation and the nature of the data — constrained optimization can be slow. For example, the constrained optimizer `constrOptim` available in **R** is acknowledged to be slow if the optimum lies on the boundary of the parameter space. We shall focus on the use of the unconstrained op-

timizer `nlm`. Exercise 3 and A.4 explore the use of `constrOptim`, which can minimize a function subject to linear inequality constraints.

In general, there are two groups of constraints: those that apply to the parameters of the state-dependent distributions and those that apply to the parameters of the Markov chain. The first group of constraints depends on which state-dependent distribution(s) are chosen; e.g. the 'success probability' of a binomial distribution lies between 0 and 1.

In the case of a Poisson–HMM the relevant constraints are:

- the means λ_i of the state-dependent distributions must be nonnegative, for $i = 1, \ldots, m$;
- the rows of the transition probability matrix $\boldsymbol{\Gamma}$ must add to 1, and all the parameters γ_{ij} must be nonnegative.

Here the constraints can be circumvented by making certain transformations. The transformation of the parameters λ_i is relatively easy. Define $\eta_i = \log \lambda_i$, for $i = 1, \ldots, m$. Then $\eta_i \in \mathbb{R}$. After we have maximized the likelihood with respect to the unconstrained parameters, the constrained parameter estimates can be obtained by transforming back: $\hat{\lambda}_i = \exp \hat{\eta}_i$.

The reparametrization of the matrix $\boldsymbol{\Gamma}$ requires more work, but can be accomplished quite elegantly. Note that $\boldsymbol{\Gamma}$ has m^2 entries but only $m(m-1)$ free parameters, as there are m row sum constraints

$$\gamma_{i1} + \gamma_{i2} + \cdots + \gamma_{im} = 1 \quad (i = 1, \ldots, m).$$

We shall show one possible transformation between the m^2 constrained probabilities γ_{ij} and $m(m-1)$ unconstrained real numbers $\tau_{ij}, i \neq j$.

For the sake of readability we show the case $m = 3$. We begin by defining the matrix

$$\mathbf{T} = \begin{pmatrix} - & \tau_{12} & \tau_{13} \\ \tau_{21} & - & \tau_{23} \\ \tau_{31} & \tau_{32} & - \end{pmatrix}, \text{ a matrix with } m(m-1) \text{ entries } \tau_{ij} \in \mathbb{R}.$$

Now let $g : \mathbb{R} \to \mathbb{R}^+$ be a strictly increasing function, e.g.

$$g(x) = e^x \quad \text{or} \quad g(x) = \begin{cases} e^x & x \leq 0 \\ x + 1 & x \geq 0. \end{cases}$$

Define

$$\varrho_{ij} = \begin{cases} g(\tau_{ij}) & \text{for } i \neq j \\ 1 & \text{for } i = j. \end{cases}$$

We then set $\gamma_{ij} = \varrho_{ij} / \sum_{k=1}^{3} \varrho_{ik}$ (for $i, j = 1, 2, 3$) and $\boldsymbol{\Gamma} = (\gamma_{ij})$. It is left to the reader as an exercise to verify that the resulting matrix $\boldsymbol{\Gamma}$ satisfies the constraints of a transition probability matrix. We shall refer to the parameters η_i and τ_{ij} as **working parameters**, and to the parameters λ_i and γ_{ij} as **natural parameters**.

Using the above transformations of $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$, we can perform the calculation of the likelihood-maximizing parameters in two steps.

1. Maximize L_T with respect to the working parameters $\mathbf{T} = \{\tau_{ij}\}$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$. These are all unconstrained.
2. Transform the estimates of the working parameters to estimates of the natural parameters:

$$\hat{\mathbf{T}} \to \hat{\boldsymbol{\Gamma}}, \ \hat{\boldsymbol{\eta}} \to \hat{\boldsymbol{\lambda}}.$$

See A.1.1 and A.1.2 for **R** functions that transform natural parameters to working and vice versa.

As an illustration we consider the first row of $\boldsymbol{\Gamma}$ for the case $g(x) = e^x$ and $m = 3$. We have

$$\begin{aligned} \gamma_{11} &= 1/\left(1 + \exp(\tau_{12}) + \exp(\tau_{13})\right), \\ \gamma_{12} &= \exp(\tau_{12})/\left(1 + \exp(\tau_{12}) + \exp(\tau_{13})\right), \\ \gamma_{13} &= \exp(\tau_{13})/\left(1 + \exp(\tau_{12}) + \exp(\tau_{13})\right). \end{aligned}$$

The transformation in the opposite direction is

$$\begin{aligned} \tau_{12} &= \log\left(\gamma_{12}/\left(1 - \gamma_{12} - \gamma_{13}\right)\right) = \log\left(\gamma_{12}/\gamma_{11}\right), \\ \tau_{13} &= \log\left(\gamma_{13}/\left(1 - \gamma_{12} - \gamma_{13}\right)\right) = \log\left(\gamma_{13}/\gamma_{11}\right). \end{aligned}$$

This generalization of the logit and inverse logit transforms has long been used in (for example) the context of compositional data: see Aitchison (1982), where several other transforms are described as well.

3.3.2 Embedding in a continuous-time Markov chain

A different reparametrization is described by Zucchini and MacDonald (1998). In a continuous-time Markov chain on a finite state-space, the transition probability matrix $\mathbf{P}_t$ is given by $\mathbf{P}_t = \exp(t\mathbf{Q})$, where $\mathbf{Q}$ is the matrix of transition intensities. The row sums of $\mathbf{Q}$ are zero, but the only constraint on the off-diagonal elements of $\mathbf{Q}$ is that they be nonnegative. The one-step transition probabilities in a discrete-time Markov chain can therefore be parametrized via $\boldsymbol{\Gamma} = \exp(\mathbf{Q})$. This is effectively the parametrization used in the **R** package `msm` (Jackson *et al.*, 2003).

3.4 Other problems

3.4.1 Multiple maxima in the likelihood

The likelihood of an HMM is a complicated function of the parameters and frequently has several local maxima. The goal of course is to find the global maximum, but there is no simple method of determining in

general whether a numerical maximization algorithm has reached the global maximum. Depending on the starting values, it can easily happen that the algorithm identifies a local, but not the global, maximum. This applies also to the main alternative method of estimation, the EM algorithm, which is discussed in Chapter 4. A sensible strategy is therefore to use a range of starting values for the maximization, and to see whether the same maximum is identified in each case.

3.4.2 Starting values for the iterations

It is often easy to find plausible starting values for some of the parameters of an HMM: for instance, if one seeks to fit a Poisson–HMM with two states, and the sample mean is 10, one could try 8 and 12, or 5 and 15, for the values of the two state-dependent means. More systematic strategies based on the quantiles of the observations are possible, however: e.g. if the model has three states, use as the starting values of the state-dependent means the lower quartile, median and upper quartile of the observed counts.

It is less easy to guess values of the transition probabilities γ_{ij}. One strategy is to assign a common starting value (e.g. 0.01 or 0.05) to all the off-diagonal transition probabilities. A consequence of such a choice, perhaps convenient, is that the corresponding stationary distribution is uniform over the states; this follows by symmetry. Choosing good starting values for parameters tends to steer one away from numerical instability.

3.4.3 Unbounded likelihood

In the case of HMMs with continuous state-dependent distributions, just as in the case of independent mixtures (see Section 1.2.3), it may happen that the likelihood is unbounded in the vicinity of certain parameter combinations. As before, we suggest that, if this creates difficulties, one maximizes the discrete likelihood instead of the joint density. This has the advantage in any case that it applies more generally to data known only up to some interval. Applications of this kind are described in Sections 10.3 and 10.4; code for the latter is in A.3.

3.5 Example: earthquakes

Figure 3.1 shows the result of fitting (stationary) Poisson–hidden Markov models with two and three states to the earthquakes series by means of the unconstrained optimizer `nlm`. The relevant code appears in A.1. The

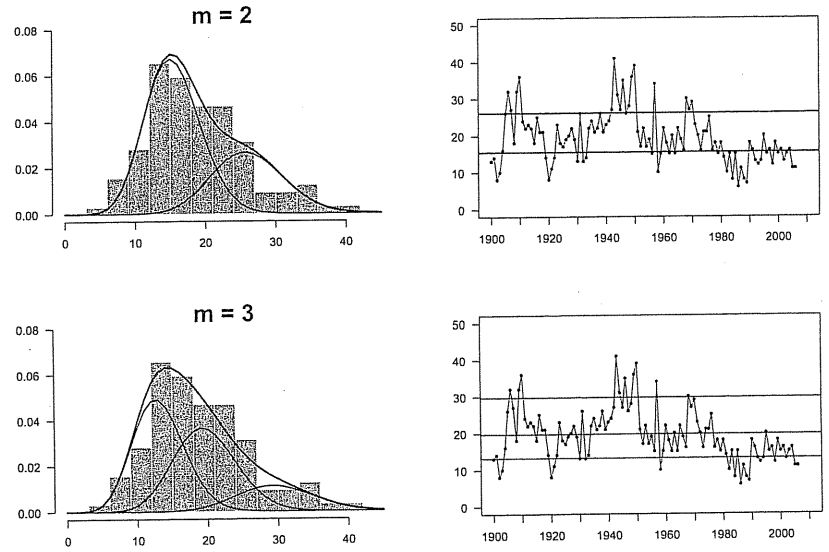

Figure 3.1 *Earthquakes series. Left: marginal distributions of Poisson–HMMs with 2 and 3 states, and their components, compared with a histogram of the observations. Right: the state-dependent means (horizontal lines) compared to the observations.*

two-state model is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9340 & 0.0660 \\ 0.1285 & 0.8715 \end{pmatrix},$$

with $\boldsymbol{\delta} = (0.6608, 0.3392)$, $\boldsymbol{\lambda} = (15.472, 26.125)$, and log-likelihood given by $l = -342.3183$. It is clear that the fitted (Markov-dependent) mixture of two Poisson distributions provides a much better fit to the marginal distribution of the observations than does a single Poisson distribution, but the fit can be further improved by using a mixture of three or four Poisson distributions.

The three-state model is

$$\mathbf{\Gamma} = \begin{pmatrix} 0.955 & 0.024 & 0.021 \\ 0.050 & 0.899 & 0.051 \\ 0.000 & 0.197 & 0.803 \end{pmatrix},$$

with $\boldsymbol{\delta} = (0.4436, 0.4045, 0.1519)$, $\boldsymbol{\lambda} = (13.146, 19.721, 29.714)$ and $l =$

-329.4603, and the four-state is as follows:

$$\mathbf{\Gamma} = \begin{pmatrix} 0.805 & 0.102 & 0.093 & 0.000 \\ 0.000 & 0.976 & 0.000 & 0.024 \\ 0.050 & 0.000 & 0.902 & 0.048 \\ 0.000 & 0.000 & 0.188 & 0.812 \end{pmatrix},$$

with $\boldsymbol{\delta} = (0.0936, 0.3983, 0.3643, 0.1439)$, $\boldsymbol{\lambda} = (11.283,\ 13.853,\ 19.695,\ 29.700)$, and $l = -327.8316$.

The means and variances of the marginal distributions of the four models compare as follows with those of the observations. By a one-state Poisson–HMM we mean a model that assumes that the observations are realizations of independent Poisson random variables with common mean.

	mean	variance
observations:	19.364	51.573
'one-state HMM':	19.364	19.364
two-state HMM:	19.086	44.523
three-state HMM:	18.322	50.709
four-state HMM:	18.021	49.837

As regards the autocorrelation functions of the models, i.e. $\rho(k) = \text{Corr}(X_{t+k}, X_t)$, we have the following results, valid for all $k \in \mathbb{N}$, based on the conclusions of Exercise 4 of Chapter 2:

- two states: $\rho(k) = 0.5713 \times 0.8055^k$;
- three states: $\rho(k) = 0.4447 \times 0.9141^k + 0.1940 \times 0.7433^k$;
- four states: $\rho(k) = 0.2332 \times 0.9519^k + 0.3682 \times 0.8174^k + 0.0369 \times 0.7252^k$.

In all these cases the ACF is just a linear combination of the kth powers of the eigenvalues other than 1 of the transition probability matrix.

For model selection, e.g. choosing between competing models such as HMMs and independent mixtures, or choosing the number of components in either, see Section 6.1.

A phenomenon that is noticeable when one fits models with three or more states to relatively short series is that the estimates of one or more of the transition probabilities turn out to be very close to zero; see the three-state model above (one such probability, γ_{13}) and the four-state model (six of the twelve off-diagonal transition probabilities).

This phenomenon can be explained as follows. In a stationary Markov chain, the expected number of transitions from state i to state j in a series of T observations is $(T-1)\delta_i\gamma_{ij}$. For $\delta_3 = 0.152$ and $T = 107$ (as in our three-state model), this expectation will be less than 1 if $\gamma_{31} < 0.062$. In such a series, therefore, it is likely that if γ_{31} is fairly small there will be no transitions from state 3 to state 1, and so when we

seek to estimate γ_{31} in an HMM the estimate is likely to be effectively zero. As m increases, the probabilities δ_i and γ_{ij} get smaller on average; this makes it increasingly likely that at least one estimated transition probability is effectively zero.

3.6 Standard errors and confidence intervals

Relatively little is known about the properties of the maximum likelihood estimators of HMMs; only asymptotic results are available. To exploit these results one requires estimates of the variance-covariance matrix of the estimators of the parameters. One can estimate the standard errors from the Hessian of the log-likelihood at the maximum, but this approach runs into difficulties when some of the parameters are on the boundary of their parameter space, which occurs quite often when HMMs are fitted. An alternative here is the parametric bootstrap, for which see Section 3.6.2. The algorithm is easy to code (see A.2.1), but the computations are time-consuming.

3.6.1 Standard errors via the Hessian

Although the point estimates $\widehat{\boldsymbol{\Theta}} = (\widehat{\mathbf{\Gamma}}, \widehat{\boldsymbol{\lambda}})$ are easy to compute, exact interval estimates are not available. Cappé *et al.* (2005, Chapter 12) show that, under certain regularity conditions, the MLEs of HMM parameters are consistent, asymptotically normal and efficient. Thus, if we can estimate the standard errors of the MLEs, then, using the asymptotic normality, we can also compute approximate confidence intervals. However, as pointed out by Frühwirth-Schnatter (2006, p. 53) in the context of independent mixture models, 'The regularity conditions are often violated, including cases of great practical concern, among them small data sets, mixtures with small component weights, and overfitting mixtures with too many components.' Furthermore, McLachlan and Peel (2000, p. 68) warn: 'In particular for mixture models, it is well known that the sample size n has to be very large before the asymptotic theory of maximum likelihood applies.'

With the above caveats in mind we can, to estimate the standard errors of the MLEs of an HMM, use the approximate Hessian of minus the log-likelihood at the minimum, e.g. as supplied by `nlm`. We can invert it and so estimate the asymptotic variance-covariance matrix of the estimators of the parameters. A problem with this suggestion is that, if the parameters have been transformed, the Hessian available will be that which refers to the working parameters ϕ_i, not the original, more readily interpretable, natural parameters θ_i ($\mathbf{\Gamma}$ and $\boldsymbol{\lambda}$ in the case of a Poisson–HMM).

The situation is therefore that we have the Hessian

$$\mathbf{H} = -\left(\frac{\partial^2 l}{\partial\phi_i \partial\phi_j}\right)$$

available at the minimum of $-l$, and what we really need is the Hessian with respect to the natural parameters:

$$\mathbf{G} = -\left(\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right).$$

There is, however, the following relationship between the two Hessians at the minimum:

$$\mathbf{H} = \mathbf{M}\mathbf{G}\mathbf{M}' \quad \text{and} \quad \mathbf{G}^{-1} = \mathbf{M}'\mathbf{H}^{-1}\mathbf{M}, \tag{3.2}$$

where $\mathbf{M}$ is defined by $m_{ij} = \partial\theta_j/\partial\phi_i$. (Note that all the derivatives appearing here are as evaluated at the minimum.) In the case of a Poisson–HMM, the elements of $\mathbf{M}$ are quite simple: see Exercise 7 for details.

With $\mathbf{M}$ at our disposal, we can use (3.2) to deduce $\mathbf{G}^{-1}$ from $\mathbf{H}^{-1}$, and use $\mathbf{G}^{-1}$ to find standard errors for the natural parameters, provided such parameters are not on the boundary of the parameter space. It is however true in many applications that some of the estimated parameters lie on or very close to the boundary; this limits the usefulness of the above results.

As already pointed out on p. 52, for series of moderate length the estimates of some transition probabilities are expected to be close to zero. This is true of $\hat{\gamma}_{13}$ in the three-state model for the earthquakes series. An additional example of this type can be found in Section 10.2.2. In Section 12.2.1, several of the estimates of the parameters in the state-dependent distributions are practically zero, their lower bound; see Table 12.1. The same phenomenon is apparent in Section 16.9.2; see Table 16.1.

Recursive computation of the Hessian

An alternative method of computing the Hessian is that of Lystig and Hughes (2002). They present the forward algorithm $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t)$ in a form which incorporates automatic or 'natural' scaling, and then extend that approach in order to compute (in a single pass, along with the log-likelihood) its Hessian and gradient with respect to the natural parameters, those we have denoted above by θ_i. Turner (2008) has used this approach in order to find the analytical derivatives needed to maximize HMM likelihoods directly by the Levenberg–Marquardt algorithm.

While this may be a more efficient and more accurate method of computing the Hessian than that outlined above, it does not solve the fundamental problem that the use of the Hessian to compute standard errors

(and thence confidence intervals) is unreliable if some of the parameters are on or near the boundary of their parameter space.

3.6.2 Bootstrap standard errors and confidence intervals

As an alternative to the technique described in Section 3.6.1 one may use the **parametric bootstrap** (Efron and Tibshirani, 1993). Roughly speaking, the idea of the parametric bootstrap is to assess the properties of the model with parameters $\boldsymbol{\Theta}$ by using those of the model with parameters $\widehat{\boldsymbol{\Theta}}$. The following steps are performed to estimate the variance-covariance matrix of $\widehat{\boldsymbol{\Theta}}$.

1. Fit the model, i.e. compute $\widehat{\boldsymbol{\Theta}}$.
2. (a) Generate a sample, called a bootstrap sample, of observations from the fitted model, i.e. the model with parameters $\widehat{\boldsymbol{\Theta}}$. The length should be the same as the original number of observations.
   (b) Estimate the parameters $\boldsymbol{\Theta}$ by $\widehat{\boldsymbol{\Theta}}^*$ for the bootstrap sample.
   (c) Repeat steps (a) and (b) B times (with B 'large') and record the values $\widehat{\boldsymbol{\Theta}}^*$.

The variance-covariance matrix of $\widehat{\boldsymbol{\Theta}}$ is then estimated by the sample variance-covariance matrix of the bootstrap estimates $\widehat{\boldsymbol{\Theta}}^*(b)$, $b = 1, 2, \ldots, B$:

$$\widehat{\text{Var-Cov}}(\widehat{\boldsymbol{\Theta}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\boldsymbol{\Theta}}^*(b) - \widehat{\boldsymbol{\Theta}}^*(.)\right)'\left(\widehat{\boldsymbol{\Theta}}^*(b) - \widehat{\boldsymbol{\Theta}}^*(.)\right),$$

where $\widehat{\boldsymbol{\Theta}}^*(.) = B^{-1}\sum_{b=1}^{B}\widehat{\boldsymbol{\Theta}}^*(b)$.

The parametric bootstrap requires code to generate realizations from a fitted model; for a Poisson–HMM this is given in A.2.1. Since code to fit models is available, that same code can be used to fit models to the bootstrap sample.

The bootstrap method can be used to estimate confidence intervals directly. In the example given in the next section we use the well-known 'percentile method' (Efron and Tibshirani, 1993); other options are available.

3.7 Example: the parametric bootstrap applied to the three-state model for the earthquakes data

A bootstrap sample of size 500 was generated from the three-state model for the earthquakes data, which appears on p. 51. In fitting models to the bootstrap samples, we noticed that, in two cases out of the 500, the starting values we were in general using caused numerical instability or

Table 3.1 *Earthquakes data: bootstrap confidence intervals for the parameters of the three-state HMM.*

parameter	MLE	90% conf. limits	
λ_1	13.146	11.463	14.253
λ_2	19.721	13.708	21.142
λ_3	29.714	20.929	33.160
γ_{11}	0.954	0.750	0.988
γ_{12}	0.024	0.000	0.195
γ_{13}	0.021	0.000	0.145
γ_{21}	0.050	0.000	0.179
γ_{22}	0.899	0.646	0.974
γ_{23}	0.051	0.000	0.228
γ_{31}	0.000	0.000	0.101
γ_{32}	0.197	0.000	0.513
γ_{33}	0.803	0.481	0.947
δ_1	0.444	0.109	0.716
δ_2	0.405	0.139	0.685
δ_3	0.152	0.042	0.393

Table 3.2 *Earthquakes data: bootstrap estimates of the correlations of the estimators of λ_i, for $i = 1, 2, 3$.*

	λ_1	λ_2	λ_3
λ_1	1.000	0.483	0.270
λ_2		1.000	0.688
λ_3			1.000

convergence problems. By choosing better starting values for these two cases we were able to fit models successfully and complete the exercise. The resulting sample of parameter values then produced the 90% confidence intervals for the parameters that are displayed in Table 3.1, and the estimated parameter correlations that are displayed in Table 3.2. What is noticeable is that the intervals for the state-dependent means λ_i overlap, the intervals for the stationary probabilities δ_i are very wide, and the estimators $\hat{\lambda}_i$ are quite strongly correlated.

These results, in particular the correlations shown in Table 3.2, should make one wary of over-interpreting a model with nine parameters based on only 107 (dependent) observations. In particular, they suggest that the states are not well defined, and one should be cautious of attaching a substantive interpretation to them.

Exercises

1. Consider the following parametrization of the t.p.m. of an m-state Markov chain. Let $\tau_{ij} \in \mathbb{R}$ $(i, j = 1, 2, \ldots, m; i \neq j)$ be $m(m-1)$ arbitrary real numbers. Let $g : \mathbb{R} \to \mathbb{R}^+$ be some strictly increasing function, e.g. $g(x) = e^x$. Define ϱ_{ij} and γ_{ij} as on p. 48.

   (a) Show that the matrix $\mathbf{\Gamma}$ with entries γ_{ij} that are constructed in this way is a t.p.m., i.e. show that $0 \leq \gamma_{ij} \leq 1$ for all i and j, and that the row sums of $\mathbf{\Gamma}$ are equal to one.

   (b) Given a t.p.m. $\mathbf{\Gamma} = \{\gamma_{ij} : i, j = 1, 2, \ldots, m\}$, derive an expression for the parameters τ_{ij}, for $i, j = 1, 2, \ldots, m$; $i \neq j$.

2. The purpose of this exercise is to investigate the numerical behaviour of an 'unscaled' evaluation of the likelihood of an HMM, and to compare this with the behaviour of an alternative algorithm that applies scaling.

   Consider the stationary two-state Poisson–HMM with parameters

   $$\mathbf{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}, \qquad (\lambda_1, \lambda_2) = (1, 5).$$

   Compute the likelihood, L_{10}, of the following sequence of ten observations in two ways: 2, 8, 6, 3, 6, 1, 0, 0, 4, 7.

   (a) Use the unscaled method $L_{10} = \boldsymbol{\alpha}_{10}\mathbf{1}'$, where $\boldsymbol{\alpha}_0 = \boldsymbol{\delta}$ and $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\mathbf{B}_t$;

   $$\mathbf{B}_t = \mathbf{\Gamma} \begin{pmatrix} p_1(x_t) & 0 \\ 0 & p_2(x_t) \end{pmatrix};$$

   and

   $$p_i(x_t) = \lambda_i^{x_t} e^{-\lambda_i} / x_t!, \quad i = 1, 2; \quad t = 1, 2, \ldots, 10.$$

   Examine the numerical values of the vectors $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{10}$.

   (b) Use the algorithm given in Section 3.2 to compute $\log L_{10}$.

   Examine the numerical values of the vectors $\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{10}$. (It is easiest to store these vectors as rows in an 11×2 matrix.)

3. Use the **R** function `constrOptim` to fit HMMs with two to four states to the earthquakes data, and compare your models with those given in Section 3.5.

4. Consider the following transformation:

   $$\begin{aligned} w_1 &= \sin^2 \theta_1 \\ w_i &= \left(\textstyle\prod_{j=1}^{i-1} \cos^2 \theta_j\right) \sin^2 \theta_i \qquad i = 2, \ldots, m-1 \\ w_m &= \textstyle\prod_{i=1}^{m-1} \cos^2 \theta_i. \end{aligned}$$

Show how this transformation can be used to convert the constraints

$$\sum_{i=1}^{m} w_i = 1, \qquad w_i \geq 0 \qquad i = 1, \ldots, m$$

into simple 'box constraints', i.e. constraints of the form $a \leq \theta_i \leq b$. How could this be used in the context of estimation in HMMs?

5.(a) Consider a stationary Markov chain, with t.p.m. $\boldsymbol{\Gamma}$ and stationary distribution $\boldsymbol{\delta}$. Show that the expected number of transitions from state i to state j in a series of T observations (i.e. in $T-1$ transitions) is $(T-1)\delta_i\gamma_{ij}$.

Hint: this expectation is $\sum_{t=2}^{T} \Pr(X_{t-1} = i, X_t = j)$.

(b) Show that, for $\delta_3 = 0.152$ and $T = 107$, this expectation is less than 1 if $\gamma_{31} < 0.062$.

6. Prove the relation (3.2) between the Hessian $\mathbf{H}$ of $-l$ with respect to the working parameters and the Hessian $\mathbf{G}$ of $-l$ with respect to the natural parameters, both being evaluated at the minimum of $-l$.

7. (See Section 3.6.1.) Consider an m-state Poisson–HMM, with natural parameters γ_{ij} and λ_i, and working parameters τ_{ij} and η_i defined as in Section 3.3.1, with $g(x) = e^x$.

(a) Show that

$$\partial\gamma_{ij}/\partial\tau_{ij} = \gamma_{ij}(1-\gamma_{ij}), \text{ for all } i,j; \quad \partial\gamma_{ij}/\partial\tau_{il} = -\gamma_{ij}\gamma_{il}, \text{ for } j \neq l;$$
$$\partial\gamma_{ij}/\partial\tau_{kl} = 0, \text{ for } i \neq k; \text{ and} \quad \partial\lambda_i/\partial\eta_i = e^{\eta_i} = \lambda_i, \text{ for all } i.$$

(b) Hence find the matrix $\mathbf{M}$ in this case.

8. Modify the **R** code in A.1 in order to fit a Poisson–HMM to interval-censored observations. (Assume that the observations are available as a $T \times 2$ matrix of which the first column contains the lower bound of the observation and the second the upper bound, possibly `Inf`.)

9. Verify the autocorrelation functions given on p. 52 for the two-, three- and four-state models for the earthquakes data. (Hint: use the **R** function `eigen` to find the eigenvalues and -vectors of the relevant transition probability matrices.)

10. Consider again the soap sales series introduced in Exercise 5 of Chapter 1.

(a) Fit stationary Poisson–HMMs with two, three and four states to these data.

(b) Find the marginal means and variances, and the ACFs, of these models, and compare them with their sample equivalents.

CHAPTER 4

Estimation by the EM algorithm

I know many statisticians are deeply in love with the EM algorithm [...]

Speed (2008)

A commonly used method of fitting HMMs is the EM algorithm, which we shall describe in Section 4.2, the crux of this chapter. The tools we need to do so are the forward and the backward probabilities, which are also used for decoding and state prediction in Chapter 5. In establishing some useful propositions concerning the forward and backward probabilities we invoke several properties of HMMs which are fairly obvious given the structure of an HMM; we defer the proofs of such properties to Appendix B.

In the context of HMMs the EM algorithm is known as the Baum–Welch algorithm. The Baum–Welch algorithm is designed to estimate the parameters of an HMM whose Markov chain is homogeneous but not necessarily stationary. Thus, in addition to the parameters of the state-dependent distributions and the t.p.m. $\boldsymbol{\Gamma}$, the initial distribution $\boldsymbol{\delta}$ is also estimated; it is not assumed that $\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\delta}$. Indeed the method has to be modified if this assumption is made; see Section 4.2.5.

4.1 Forward and backward probabilities

In Section 2.3.2 we have, for $t = 1, 2, \ldots, T$, defined the (row) vector $\boldsymbol{\alpha}_t$ as follows:

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_t) = \boldsymbol{\delta}\mathbf{P}(x_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\mathbf{P}(x_s), \tag{4.1}$$

with $\boldsymbol{\delta}$ denoting the initial distribution of the Markov chain. We have referred to the elements of $\boldsymbol{\alpha}_t$ as **forward probabilities**, but we have given no reason even for their description as probabilities. One of the purposes of this section is to show that $\alpha_t(j)$, the jth component of $\boldsymbol{\alpha}_t$, is indeed a probability, the joint probability $\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_t = x_t, C_t = j)$.

We shall also need the vector of **backward probabilities** $\boldsymbol{\beta}_t$ which,