CHAPTER 5

Forecasting, decoding and state prediction

Main results of this chapter:

conditional distributions p. 77

$$\Pr\left(X_t = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}\right) = \sum_i w_i(t) p_i(x)$$

forecast distributions p. 79

$$\Pr(X_{T+h} = x \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \sum_i \xi_i(h) p_i(x)$$

state probabilities and local decoding p. 81

$$\Pr(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \alpha_t(i)\beta_t(i)/L_T$$

global decoding maximize over $\mathbf{c}^{(T)}$: p. 82

$$\Pr(\mathbf{C}^{(T)} = \mathbf{c}^{(T)} \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$$

state prediction p. 86

$$\Pr(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h(, i)/L_T$$

Convenient expressions for conditional distributions and forecast distributions are available for HMMs. This makes it easy, for example, to check for outliers or to make interval forecasts. In this chapter, we first show (in Section 5.1) how to compute the conditional distribution of an observation under an HMM, i.e. the distribution of the observation at time t given the observations at all other times. In Section 5.2 we derive the forecast distribution of an HMM. Then, in Section 5.3, we demonstrate how, given the HMM and the observations, one can deduce information about the states occupied by the underlying Markov chain. Such inference is known as decoding. We continue to use the earthquakes series as our illustrative example. Our results are stated for the case of discrete observations X_t; if the observations are continuous, probabilities will need to be replaced by densities.

Note that in this chapter we do not assume stationarity of the Markov

chain $\{C_t\}$, only homogeneity: here the row vector $\boldsymbol{\delta}$ denotes the *initial* distribution, that of C_1, and is not assumed to be the stationary distribution. Of course the results also hold in the special case in which the Markov chain is stationary, in which case $\boldsymbol{\delta}$ is both the initial and the stationary distribution.

5.1 Conditional distributions

We now derive a formula for the distribution of X_t conditioned on all the other observations of the HMM. We use the notation $\mathbf{X}^{(-t)}$ for the observations at all times other than t; that is, we define

$$\mathbf{X}^{(-t)} \equiv (X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_T),$$

and similarly $\mathbf{x}^{(-t)}$.

Using the likelihood of an HMM as discussed in Section 2.3.2, and the definition of the forward and backward probabilities as in Section 4.1, it follows immediately, for $t = 2, 3, \ldots, T$, that

$$\begin{aligned} \Pr\left(X_t = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}\right) &= \frac{\boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{B}_2 \cdots \mathbf{B}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x)\mathbf{B}_{t+1} \cdots \mathbf{B}_T\mathbf{1}'}{\boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{B}_2 \cdots \mathbf{B}_{t-1}\boldsymbol{\Gamma}\mathbf{B}_{t+1} \cdots \mathbf{B}_T\mathbf{1}'} \\ &\propto \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x)\boldsymbol{\beta}_t'. \end{aligned} \tag{5.1}$$

Here, as before, $\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{B}_2 \cdots \mathbf{B}_t$, $\boldsymbol{\beta}_t' = \mathbf{B}_{t+1} \cdots \mathbf{B}_T\mathbf{1}'$ and $\boldsymbol{\beta}_T = \mathbf{1}$; recall that $\mathbf{B}_t$ is defined as $\boldsymbol{\Gamma}\mathbf{P}(x_t)$.

The result for the case $t = 1$ is

$$\begin{aligned} \Pr\left(X_1 = x \mid \mathbf{X}^{(-1)} = \mathbf{x}^{(-1)}\right) &= \frac{\boldsymbol{\delta}\mathbf{P}(x)\mathbf{B}_2 \cdots \mathbf{B}_T\mathbf{1}'}{\boldsymbol{\delta}\mathbf{I}\mathbf{B}_2 \cdots \mathbf{B}_T\mathbf{1}'} \\ &\propto \boldsymbol{\delta}\mathbf{P}(x)\boldsymbol{\beta}_1'. \end{aligned} \tag{5.2}$$

The above conditional distributions are ratios of two likelihoods of an HMM: the numerator is the likelihood of the observations except that the observation x_t is replaced by x, and the denominator (the reciprocal of the constant of proportionality) is the likelihood of the observations except that x_t is treated as missing.

We now show that these conditional probabilities can be expressed as mixtures of the m state-dependent probability distributions. In both Equations (5.1) and (5.2) the required conditional probability has the following form: a row vector multiplied by the $m \times m$ diagonal matrix $\mathbf{P}(x) = \text{diag}(p_1(x), \ldots, p_m(x))$, multiplied by a column vector. It follows, for $t = 1, 2, \ldots, T$, that

$$\Pr\left(X_1 = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}\right) \propto \sum_{i=1}^{m} d_i(t)p_i(x),$$

where, in the case of (5.1), $d_i(t)$ is the product of the ith entry of the

vector $\boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}$ and the ith entry of the vector $\boldsymbol{\beta}_t$; and in the case of (5.2), it is the product of the ith entry of the vector $\boldsymbol{\delta}$ and the ith entry of the vector $\boldsymbol{\beta}_1$. Hence

$$\Pr\left(X_t = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)}\right) = \sum_{i=1}^{m} w_i(t) p_i(x)\,, \tag{5.3}$$

where the mixing probabilities $w_i(t) = d_i(t)/\sum_{j=1}^{m} d_j(t)$ are functions of the observations $\mathbf{x}^{(-t)}$ and of the model parameters. The **R** code for such conditional distributions is given in A.2.9.

In Figure 5.1 we present the full array of conditional distributions for the earthquakes data. It is clear that each of the conditional distributions has a different shape, and the shape may change sharply from one time-point to the next. In addition, it is striking that some of the observed counts, which in Figure 5.1 are marked as bold bars, are extreme relative to their conditional distributions. This observation suggests using the conditional distributions for outlier checking, which will be demonstrated in Section 6.2.

5.2 Forecast distributions

We turn now to another type of conditional distribution, the forecast distribution of an HMM. Specifically we derive two expressions for the conditional distribution of X_{T+h} given $\mathbf{X}^{(T)} = \mathbf{x}^{(T)}$; h is termed the forecast horizon. Again we shall focus on the discrete case; the formulae for the continuous case are the same but with the probability functions replaced by density functions.

For discrete-valued observations the forecast distribution $\Pr(X_{T+h} = x \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$ of an HMM is very similar to the conditional distribution $\Pr(X_t = x \mid \mathbf{X}^{(-t)} = \mathbf{x}^{(-t)})$ just discussed, and can be computed in essentially the same way, as a ratio of likelihoods:

$$\begin{aligned}
\Pr(X_{T+h} = x \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) &= \frac{\Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, X_{T+h} = x)}{\Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)})} \\
&= \frac{\boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{B}_2\mathbf{B}_3\cdots\mathbf{B}_T\boldsymbol{\Gamma}^h\mathbf{P}(x)\mathbf{1}'}{\boldsymbol{\delta}\mathbf{P}(x_1)\mathbf{B}_2\mathbf{B}_3\cdots\mathbf{B}_T\mathbf{1}'} \\
&= \frac{\boldsymbol{\alpha}_T\boldsymbol{\Gamma}^h\mathbf{P}(x)\mathbf{1}'}{\boldsymbol{\alpha}_T\mathbf{1}'}\,.
\end{aligned}$$

Writing $\boldsymbol{\phi}_T = \boldsymbol{\alpha}_T/\boldsymbol{\alpha}_T\mathbf{1}'$ (see Section 3.2), we have

$$\Pr(X_{T+h} = x \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \boldsymbol{\phi}_T\boldsymbol{\Gamma}^h\mathbf{P}(x)\mathbf{1}'\,. \tag{5.4}$$

Expressions for joint distributions of several forecasts can be derived along the same lines. (See Exercise 5.)

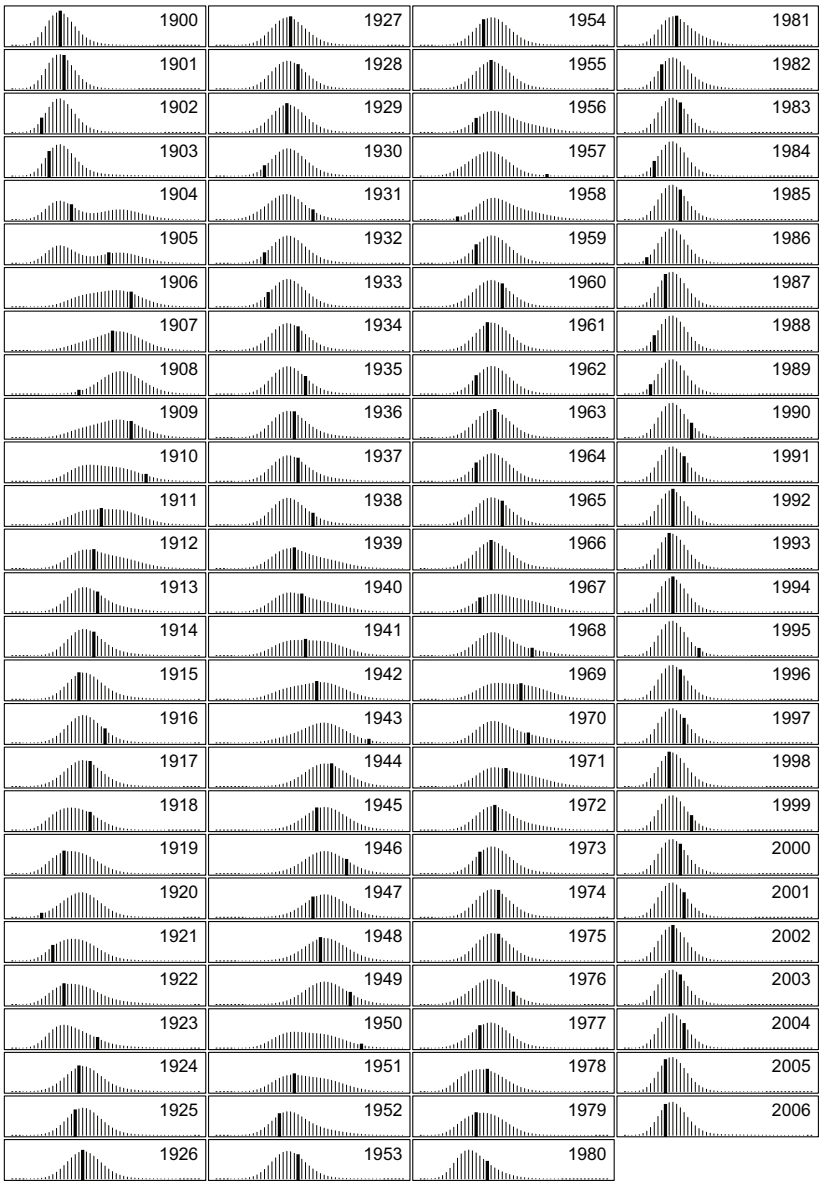

Figure 5.1 *Earthquakes data, three-state HMM: conditional distribution of the number of earthquakes in each year, given all the other observations. The bold bar corresponds to the actual number observed in that year.*

Table 5.1 *Earthquakes data, three-state Poisson–HMM: forecasts.*

year:	2007	2008	2009	2016	2026	2036
horizon:	1	2	5	10	20	30
forecast mode:	13	13	13	13	14	14
forecast median:	12.7	12.9	13.1	14.4	15.6	16.2
forecast mean:	13.7	14.1	14.5	16.4	17.5	18.0
nominal 90%						
forecast interval:	[8,21]	[8,23]	[8,25]	[8,30]	[8,32]	[9,32]
exact coverage:	0.908	0.907	0.907	0.918	0.932	0.910

The forecast distribution can therefore be written as a mixture of the state-dependent probability distributions:

$$\Pr(X_{T+h} = x \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \sum_{i=1}^{m} \xi_i(h) p_i(x)\,, \tag{5.5}$$

where the weight $\xi_i(h)$ is the i th entry of the vector $\boldsymbol{\phi}_T \boldsymbol{\Gamma}^h$. The **R** code for forecasts is given in A.2.8.

Since the entire probability distribution of the forecast is known, it is possible to make interval forecasts, and not only point forecasts. This is illustrated in Table 5.1, which lists statistics of some forecast distributions for the earthquake series fitted with a three-state Poisson HMM.

As the forecast horizon h increases, the forecast distribution converges to the marginal distribution of the stationary HMM, i.e.

$$\lim_{h\to\infty} \Pr(X_{T+h} = x \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \lim_{h\to\infty} \boldsymbol{\phi}_T \boldsymbol{\Gamma}^h \mathbf{P}(x) \mathbf{1}' = \boldsymbol{\delta}^* \mathbf{P}(x) \mathbf{1}'\,,$$

where here we temporarily use $\boldsymbol{\delta}^*$ to denote the stationary distribution of the Markov chain (in order to distinguish it from $\boldsymbol{\delta}$, the initial distribution). The limit follows from the observation that, for any nonnegative (row) vector $\boldsymbol{\eta}$ whose entries add to 1, the vector $\boldsymbol{\eta}\boldsymbol{\Gamma}^h$ approaches the stationary distribution of the Markov chain as $h \to \infty$, provided the Markov chain satisfies the usual regularity conditions of irreducibility and aperiodicity; see e.g. Feller (1968, p. 394). Sometimes the forecast distribution approaches its limiting distribution only slowly; see Figure 5.2, which displays six of the forecast distributions for the earthquakes series, compared with the limiting distribution. In other cases the approach can be relatively fast; for a case in point, consider the three-state model for the soap sales series introduced in Exercise 5 of Chapter 1. The rate of approach is determined by the size of the largest eigenvalue other than 1 of the t.p.m. $\boldsymbol{\Gamma}$.

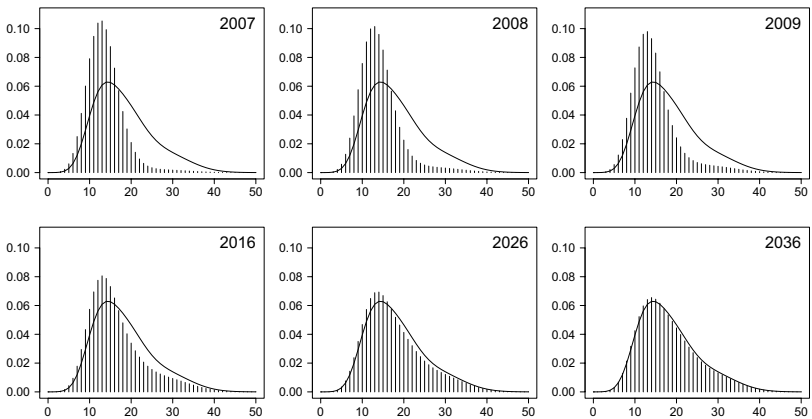

Figure 5.2 *Earthquakes data, three-state Poisson–HMM: forecast distributions for 1 to 30 years ahead, compared to limiting distribution, which is shown as a continuous line.*

5.3 Decoding

In speech recognition and other applications — see e.g. Fredkin and Rice (1992), or Guttorp (1995, p. 101) — it is of interest to determine the states of the Markov chain that are most likely (under the fitted model) to have given rise to the observation sequence. In the context of speech recognition this is known as the decoding problem: see Juang and Rabiner (1991). More specifically, 'local decoding' of the state at time t refers to the determination of that state which is most likely at that time. In contrast, 'global decoding' refers to the determination of the most likely sequence of states. These two are described in the next two sections.

5.3.1 State probabilities and local decoding

Consider again the vectors of forward and backward probabilities, $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$, as discussed in Section 4.1. For the derivation of the most likely state of the Markov chain at time t, we shall use the following result, which appears there as Equation (4.9):

$$\alpha_t(i)\beta_t(i) = \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i)\,.$$

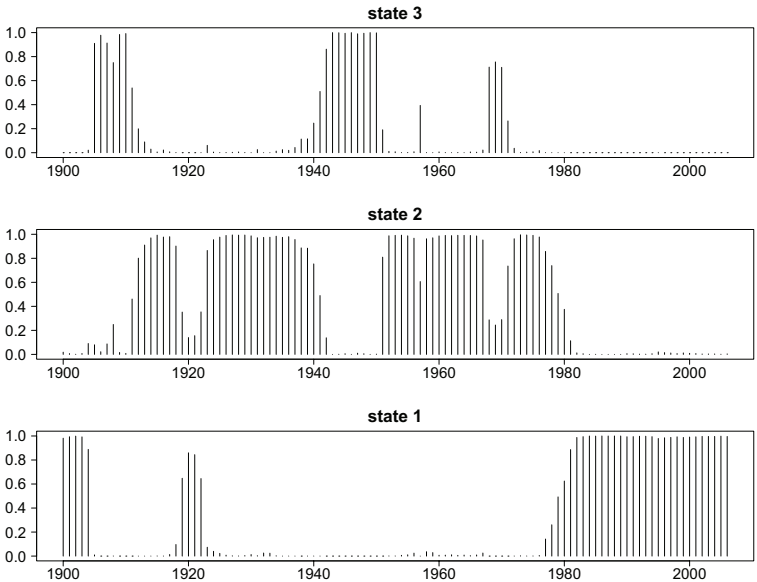

Figure 5.3 *Earthquakes data: state probabilities for fitted three-state HMM.*

Hence the conditional distribution of C_t given the observations can be obtained, for $i = 1, 2, \ldots, m$, as

$$\begin{aligned} \Pr(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) &= \frac{\Pr(C_t = i, \mathbf{X}^{(T)} = \mathbf{x}^{(T)})}{\Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)})} \\ &= \frac{\alpha_t(i)\beta_t(i)}{L_T}\,. \end{aligned} \tag{5.6}$$

Here L_T can be computed by the scaling method described in Section 3.2. Scaling is also necessary in order to prevent numerical underflow in the evaluation of the product $\alpha_t(i)\beta_t(i)$. The **R** code given in A.2.5 implements one method of doing this.

For each time $t \in \{1, \ldots, T\}$ one can therefore determine the distribution of the state C_t, given the observations $\mathbf{x}^{(T)}$, which for m states is a discrete probability distribution with support $\{1, \ldots, m\}$. In Figures 5.3 and 5.4 we display the state probabilities for the earthquakes series, based on the fitted three- and four-state Poisson–HMM models.

For each $t \in \{1, \ldots, T\}$ the most probable state i_t^*, given the observations, is defined as

$$i_t^* = \underset{i=1,\ldots,m}{\operatorname{argmax}} \Pr(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)})\,. \tag{5.7}$$

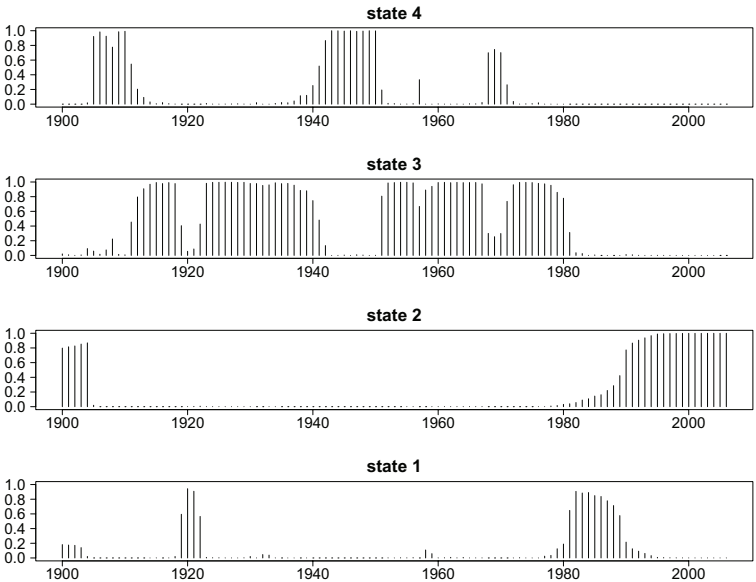

Figure 5.4 *Earthquakes data: state probabilities for fitted four-state HMM.*

This approach determines the most likely state separately for each t by maximizing the conditional probability $\Pr(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$ and is therefore called **local decoding**. In Figure 5.5 we display the results of applying local decoding to the earthquakes series, for the fitted three- and four-state models. The relevant **R** code is given in A.2.5 and A.2.6.

5.3.2 Global decoding

In many applications, e.g. speech recognition, one is not so much interested in the most likely state for each separate time t — as provided by local decoding — as in the most likely *sequence* of (hidden) states. Instead of maximizing $\Pr(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)})$ over i for each t, one seeks that sequence of states $c_1, c_2, \ldots, c_T$ which maximizes the conditional probability

$$\Pr(\mathbf{C}^{(T)} = \mathbf{c}^{(T)} \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}); \tag{5.8}$$

or equivalently, and more conveniently, the joint probability:

$$\Pr(\mathbf{C}^{(T)}, \mathbf{X}^{(T)}) \quad = \quad \delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1}, c_t} \prod_{t=1}^{T} p_{c_t}(x_t).$$

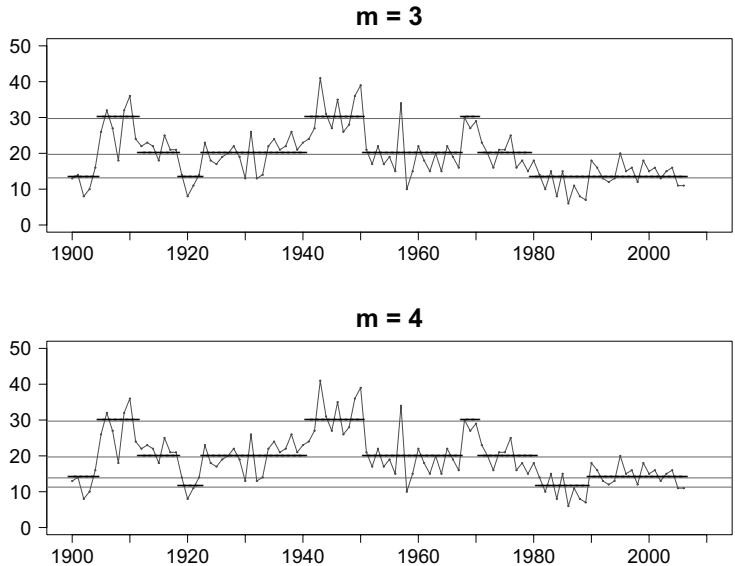

Figure 5.5 *Earthquakes data: local decoding according to three- and four-state HMMs. The horizontal lines indicate the state-dependent means.*

This is a subtly different maximization problem from that of local decoding, and is called **global decoding**. The results of local and global decoding are often very similar but not identical.

Maximizing (5.8) over all possible state sequences $c_1, c_2, \ldots, c_T$ by brute force would involve m^T function evaluations, which is clearly not feasible except for very small T. Fortunately one can use instead an efficient dynamic programming algorithm to determine the most likely sequence of states. The Viterbi algorithm (Viterbi, 1967; Forney, 1973) is such an algorithm.

We begin by defining

$$\xi_{1i} = \Pr(C_1 = i, X_1 = x_1) = \delta_i \, p_i(x_1),$$

and, for $t = 2, 3, \ldots, T$,

$$\xi_{ti} \quad = \quad \max_{c_1, c_2, \ldots, c_{t-1}} \Pr(\mathbf{C}^{(t-1)} = \mathbf{c}^{(t-1)}, C_t = i, \mathbf{X}^{(T)} = \mathbf{x}^{(T)}).$$

It can then be shown (see Exercise 1) that the probabilities ξ_{tj} satisfy the following recursion, for $t = 2, 3, \ldots, T$ and $i = 1, 2, \ldots, m$:

$$\xi_{tj} = \big(\max_i (\xi_{t-1,i} \, \gamma_{ij}) \big) p_j(x_t). \tag{5.9}$$

This provides an efficient means of computing the $T \times m$ matrix of values ξ_{tj}, as the computational effort is linear in T. The required maximizing sequence of states $i_1, i_2, \ldots, i_T$ can then be determined recursively from

$$i_T = \underset{i=1,\ldots,m}{\operatorname{argmax}}\, \xi_{Ti} \tag{5.10}$$

and, for $t = T-1, T-2, \ldots, 1$, from

$$i_t = \underset{i=1,\ldots,m}{\operatorname{argmax}}\, (\xi_{ti}\, \gamma_{i,i_{t+1}}). \tag{5.11}$$

Note that, since the quantity to be maximized in global decoding is simply a product of probabilities (as opposed to a sum of such products), one can choose to maximize its logarithm, in order to prevent numerical underflow; the Viterbi algorithm can easily be rewritten in terms of the logarithms of the probabilities. Alternatively a scaling similar to that used in the likelihood computation can be employed: in that case one scales each of the T rows of the matrix $\{\xi_{ti}\}$ to have row sum 1. The Viterbi algorithm is applicable to both stationary and nonstationary underlying Markov chains; there is no necessity to assume that the initial distribution $\boldsymbol{\delta}$ is the stationary distribution. For the relevant **R** code, see A.2.4.

Figure 5.6 displays, for the fitted three- and four-state models for the earthquakes series, the paths obtained by the Viterbi algorithm. The paths are very similar to those obtained by local decoding: compare with Figure 5.5. But they do differ. In the case of the three-state model, the years 1911, 1941 and 1980 differ. In the case of the four-state model, 1911 and 1941 differ. Notice also the nature of the difference between the upper and lower panels of Figure 5.6: allowing for a fourth state has the effect of splitting one of the states of the three-state model, that with the lowest mean. When four states are allowed, the 'Viterbi path' moves along the lowest state in the years 1919–1922 and 1981–1989 only.

Global decoding is the main objective in many applications, especially when there are substantive interpretations for the states. It is therefore of interest to investigate the performance of global decoding in identifying the correct states. This can be done by simulating a series from an HMM, applying the algorithm in order to decode the simulated observations, and then comparing the Viterbi path with the (known) series of simulated states. We present here an example of such a comparison, based on a simulated sequence of length 100 000 from the three-state (stationary) model for the earthquakes given on p. 51.

The 3×3 table displayed below with its marginals gives the simulated joint distribution of the true state i (rows) and the Viterbi estimate j of the state (columns). The row totals are close to (0.444, 0.405, 0.152), the stationary distribution of the model; this provides a partial check of the

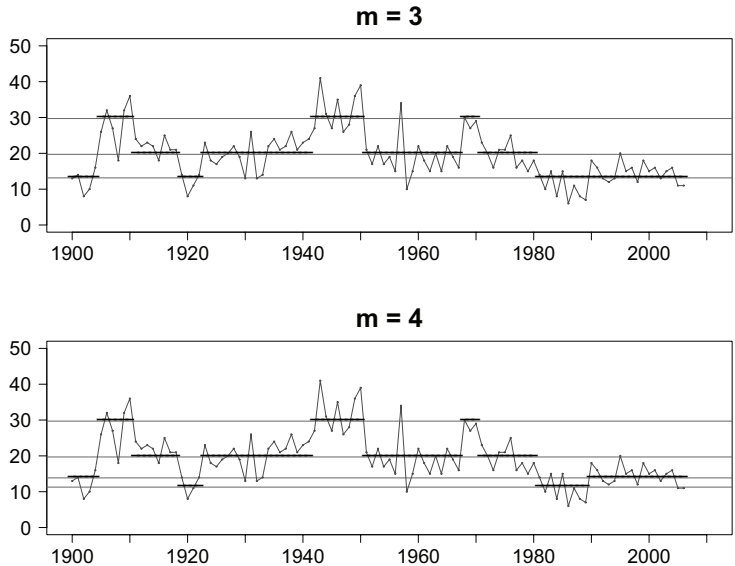

Figure 5.6 *Earthquakes: global decoding according to three- and four-state HMMs. The horizontal lines indicate the state-dependent means.*

simulation. The column totals (the distribution of the state as inferred by Viterbi) are also close to this stationary distribution.

	$j = 1$	2	3	
$i = 1$	0.431	0.017	0.000	0.448
2	0.018	0.366	0.013	0.398
3	0.000	0.020	0.134	0.154
	0.449	0.403	0.147	1.000

From this table one may conclude, for instance, that the estimated probability that the state inferred is 2, if the true state is 1, is $0.017/0.448 = 0.038$. More generally, the left-hand table below gives Pr(inferred state $= j \mid$ true state $= i$) and the right-hand table gives Pr(true state $= i \mid$ inferred state $= j$).

	$j = 1$	2	3
$i = 1$	0.961	0.038	0.000
2	0.046	0.921	0.033
3	0.002	0.130	0.868

	$j = 1$	2	3
$i = 1$	0.958	0.043	0.001
2	0.041	0.908	0.088
3	0.001	0.050	0.910

Ideally all the diagonal elements of the above two tables would be 1; here

Table 5.2 *Earthquakes data. State prediction using a three-state Poisson–HMM: the probability that the Markov chain will be in a given state in the specified year.*

year	2007	2008	2009	2016	2026	2036
state=1	0.951	0.909	0.871	0.674	0.538	0.482
2	0.028	0.053	0.077	0.220	0.328	0.373
3	0.021	0.038	0.052	0.107	0.134	0.145

they range from 0.868 to 0.961. Such a simulation exercise quantifies the expected accuracy of the Viterbi path and is therefore particularly recommended in applications in which the interpretation of that path is an important objective of the analysis.

5.4 State prediction

In Section 5.3.1 we derived an expression for the conditional distribution of the state C_t, for $t = 1, 2, \ldots, T$, given the observations $\mathbf{x}^{(T)}$. In so doing we considered only present or past states. However, it is also possible to provide the conditional distribution of the state C_t for $t > T$, i.e to perform 'state prediction'.

Given the observations $x_1, \ldots, x_T$, the following set of statements can be made about future, present and past states (respectively):

$$L_T \Pr(C_t = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \begin{cases} \boldsymbol{\alpha}_T \boldsymbol{\Gamma}^{t-T}(, i) & \text{for } t > T & \text{state prediction} \\ \alpha_T(i) & \text{for } t = T & \text{filtering} \\ \alpha_t(i)\beta_t(i) & \text{for } 1 \leq t < T & \text{smoothing,} \end{cases}$$

where $\boldsymbol{\Gamma}^{t-T}(, i)$ denotes the i th column of the matrix $\boldsymbol{\Gamma}^{t-T}$. The 'filtering' and 'smoothing' parts (for present or past states) are identical to the state probabilities as described in Section 5.3.1, and indeed could here be combined, since $\beta_T(i) = 1$ for all i. The 'state prediction' part is simply a generalization to $t > T$, the future, and can be restated as follows (see Exercise 6); for $i = 1, 2, \ldots, m$,

$$\Pr(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \boldsymbol{\alpha}_T \boldsymbol{\Gamma}^h(, i)/L_T = \boldsymbol{\phi}_T \boldsymbol{\Gamma}^h(, i), \qquad (5.12)$$

with $\boldsymbol{\phi}_T = \boldsymbol{\alpha}_T / \boldsymbol{\alpha}_T \mathbf{1}'$. Note that, as $h \to \infty$, $\boldsymbol{\phi}_T \boldsymbol{\Gamma}^h$ approaches the stationary distribution of the Markov chain.

Table 5.2 gives, for a range of years, the state predictions based on the three-state model for the earthquake series. The **R** code for state prediction is given in A.2.7.

Exercises

1. Prove the recursion (5.9):

$$\xi_{tj} = \big(\max_i(\xi_{t-1,i}\,\gamma_{ij})\big)p_j(x_t).$$

2. Apply local and global decoding to a three-state model for the soap sales series introduced in Exercise 5 of Chapter 1, and compare the results to see how much the conclusions differ.

3. Compute the h-step-ahead state predictions for the soap sales series, for $h = 1$ to 5. How close are these distributions to the stationary distribution of the Markov chain?

4.(a) Using the same sequence of random numbers in each case, generate sequences of length 1000 from the Poisson–HMMs with

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

and: (i) $\boldsymbol{\lambda} = (10, 20, 30)$; and (ii) $\boldsymbol{\lambda} = (15, 20, 25)$. Keep a record of the sequence of states, which should be the same in (i) and (ii).

(b) Use the Viterbi algorithm to infer the most likely sequence of states in each case, and compare these two sequences to the 'true' underlying sequence, i.e. the generated one.

(c) What conclusions do you draw about the accuracy of the Viterbi algorithm?

5. Bivariate forecast distributions for HMMs

(a) Find the joint distribution of X_{T+1} and X_{T+2}, given $\mathbf{X}^{(T)}$, in as simple a form as you can.

(b) For the earthquakes data, find $\Pr(X_{T+1} \leq 10, X_{T+2} \leq 10 \mid \mathbf{X}^{(T)})$.

6. Prove Equation (5.12):

$$\Pr(C_{T+h} = i \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \boldsymbol{\alpha}_T\mathbf{\Gamma}^h(,i)/L_T = \boldsymbol{\phi}_T\mathbf{\Gamma}^h(,i).$$