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. $\mathbf{\Gamma}$, the initial distribution $\boldsymbol{\delta}$ is also estimated; it is not assumed that $\boldsymbol{\delta}\mathbf{\Gamma} = \boldsymbol{\delta}$. 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:

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

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 j th 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, for $t = 1, 2, \dots, T$, is defined by

$$\boldsymbol{\beta}_t' = \boldsymbol{\Gamma}\mathbf{P}(x_{t+1})\boldsymbol{\Gamma}\mathbf{P}(x_{t+2})\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}' = \left(\prod_{s=t+1}^{T}\boldsymbol{\Gamma}\mathbf{P}(x_s)\right)\mathbf{1}', \qquad (4.2)$$

with the convention that an empty product is the identity matrix; the case $t = T$ therefore yields $\boldsymbol{\beta}_T = \mathbf{1}$. We shall show that $\beta_t(j)$, the j th component of $\boldsymbol{\beta}_t$, can be identified as the *conditional* probability $\Pr(X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \dots, X_T = x_T \mid C_t = j)$. It will then follow that, for $t = 1, 2, \dots, T$,

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

4.1.1 Forward probabilities

It follows immediately from the definition of $\boldsymbol{\alpha}_t$ that, for $t = 1, 2, \dots, T-1$, $\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t\boldsymbol{\Gamma}\mathbf{P}(x_{t+1})$ or, in scalar form,

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{m}\alpha_t(i)\gamma_{ij}\right)p_j(x_{t+1}). \qquad (4.3)$$

We shall now use the above recursion, and Equation (B.1) in Appendix B, to prove the following result by induction.

Proposition 2 *For $t = 1, 2, \dots, T$ and $j = 1, 2, \dots, m$,*

$$\alpha_t(j) = \Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = j).$$

Proof. Since $\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(x_1)$, we have

$$\alpha_1(j) = \delta_j\, p_j(x_1) = \Pr(C_1 = j)\Pr(X_1 = x_1 \mid C_1 = j),$$

hence $\alpha_1(j) = \Pr(X_1 = x_1, C_1 = j)$; i.e. the proposition holds for $t = 1$. We now show that, if the proposition holds for some $t \in \mathbb{N}$, then it also holds for $t + 1$.

$$\begin{aligned}
\alpha_{t+1}(j) &= \sum_{i=1}^{m}\alpha_t(i)\gamma_{ij}p_j(x_{t+1}) \qquad \text{(see (4.3))}\\
&= \textstyle\sum_i \Pr(\mathbf{X}^{(t)} = \mathbf{x}^{(t)}, C_t = i)\Pr(C_{t+1} = j \mid C_t = i)\\
&\qquad\qquad \times \Pr(X_{t+1} = x_{t+1} \mid C_{t+1} = j)\\
&= \textstyle\sum_i \Pr(\mathbf{X}^{(t+1)} = \mathbf{x}^{(t+1)}, C_t = i, C_{t+1} = j) \qquad (4.4)\\
&= \Pr(\mathbf{X}^{(t+1)} = \mathbf{x}^{(t+1)}, C_{t+1} = j),
\end{aligned}$$

as required. The crux is the line numbered (4.4); Equation (B.1) provides the justification thereof. □

4.1.2 Backward probabilities

It follows immediately from the definition of $\boldsymbol{\beta}_t$ that $\boldsymbol{\beta}'_t = \boldsymbol{\Gamma}\mathbf{P}(x_{t+1})\boldsymbol{\beta}'_{t+1}$, for $t = 1, 2\ , \ldots, T-1$.

Proposition 3 *For* $t = 1, 2, \ldots, T-1$ *and* $i = 1,\ 2,\ \ldots,\ m$,

$$\beta_t(i) = \Pr(X_{t+1} = x_{t+1}, X_{t+2} = x_{t+2}, \ldots, X_T = x_T \mid C_t = i),$$

provided that $\Pr(C_t = i) > 0$. *In a more compact notation:*

$$\beta_t(i) = \Pr(\mathbf{X}_{t+1}^T = \mathbf{x}_{t+1}^T \mid C_t = i),$$

where $\mathbf{X}_a^b$ *denotes the vector* $(X_a, X_{a+1}, \ldots, X_b)$.

This proposition identifies $\beta_t(i)$ as a conditional probability: the probability of the observations being $x_{t+1}, \ldots, x_T$, given that the Markov chain is in state i at time t. (Recall that the forward probabilities are joint probabilities, not conditional.)

Proof. The proof is by induction, but essentially comes down to Equations (B.5) and (B.6) of Appendix B. These are

$$\Pr(X_{t+1} \mid C_{t+1}) \Pr(\mathbf{X}_{t+2}^T \mid C_{t+1}) = \Pr(\mathbf{X}_{t+1}^T \mid C_{t+1}), \tag{B.5}$$

and

$$\Pr(\mathbf{X}_{t+1}^T \mid C_{t+1}) = \Pr(\mathbf{X}_{t+1}^T \mid C_t, C_{t+1}). \tag{B.6}$$

To establish validity for $T = t-1$, note that, since $\boldsymbol{\beta}'_{T-1} = \boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'$,

$$\beta_{T-1}(i) = \textstyle\sum_j \Pr(C_T = j \mid C_{T-1} = i) \Pr(X_T = x_T \mid C_T = j). \tag{4.5}$$

But, by (B.6),

$$\begin{aligned} \Pr(C_T \mid C_{T-1}) \Pr(X_T \mid C_T) &= \Pr(C_T \mid C_{T-1}) \Pr(X_T \mid C_{T-1}, C_T) \\ &= \Pr(X_T, C_{T-1}, C_T) / \Pr(C_{T-1}). \quad (4.6) \end{aligned}$$

Substitute from (4.6) into (4.5), and the result is

$$\begin{aligned} \beta_{T-1}(i) &= \frac{1}{\Pr(C_{T-1} = i)} \textstyle\sum_j \Pr(X_T = x_T, C_{T-1} = i, C_T = j) \\ &= \Pr(X_T = x_T, C_{T-1} = i) / \Pr(C_{T-1} = i) \\ &= \Pr(X_T = x_T \mid C_{T-1} = i), \end{aligned}$$

as required.

To show that validity for $t+1$ implies validity for t, first note that the recursion for $\boldsymbol{\beta}_t$, and the inductive hypothesis, establish that

$$\beta_t(i) = \textstyle\sum_j \gamma_{ij} \Pr(X_{t+1} = x_{t+1} \mid C_{t+1} = j) \Pr(\mathbf{X}_{t+2}^T = \mathbf{x}_{t+2}^T \mid C_{t+1} = j). \tag{4.7}$$

But (B.5) and (B.6) imply that

$$\Pr(X_{t+1} \mid C_{t+1}) \Pr(\mathbf{X}_{t+2}^T \mid C_{t+1}) = \Pr(\mathbf{X}_{t+1}^T \mid C_t, C_{t+1}). \qquad (4.8)$$

Substitute from (4.8) into (4.7), and the result is

$$\begin{aligned}\beta_t(i) &= \textstyle\sum_j \Pr(C_{t+1} = j \mid C_t = i) \Pr(\mathbf{X}_{t+1}^T = \mathbf{x}_{t+1}^T \mid C_t = i, C_{t+1} = j)\\ &= \frac{1}{\Pr(C_t = i)} \textstyle\sum_j \Pr(\mathbf{X}_{t+1}^T = \mathbf{x}_{t+1}^T, C_t = i, C_{t+1} = j)\\ &= \Pr(\mathbf{X}_{t+1}^T = \mathbf{x}_{t+1}^T, C_t = i) / \Pr(C_t = i),\end{aligned}$$

which is the required conditional probability. □

Note that the backward probabilities require a backward pass through the data for their evaluation, just as the forward probabilities require a forward pass; hence the names.

4.1.3 Properties of forward and backward probabilities

We now establish a result relating the forward and backward probabilities $\alpha_t(i)$ and $\beta_t(i)$ to the probabilities $\Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i)$. This we shall use in applying the EM algorithm to HMMs, and in local decoding: see Section 5.3.1.

Proposition 4 *For $t = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, m$,*

$$\alpha_t(i)\beta_t(i) = \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, C_t = i), \qquad (4.9)$$

and consequently $\boldsymbol{\alpha}_t \boldsymbol{\beta}_t' = \Pr(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = L_T$, for each such t.

Proof. By the preceding two propositions,

$$\begin{aligned}\alpha_t(i)\beta_t(i) &= \Pr(\mathbf{X}_1^T, C_t = i) \Pr(\mathbf{X}_{t+1}^T \mid C_t = i)\\ &= \Pr(C_t = i) \Pr(\mathbf{X}_1^t \mid C_t = i) \Pr(\mathbf{X}_{t+1}^T \mid C_t = i).\end{aligned}$$

Now apply the conditional independence of $\mathbf{X}_1^t$ and $\mathbf{X}_{t+1}^T$ given C_t (see Equation (B.7) of Appendix B), and the result is that

$$\alpha_t(i)\beta_t(i) = \Pr(C_t = i) \Pr(\mathbf{X}_1^t, \mathbf{X}_{t+1}^T \mid C_t = i) = \Pr(\mathbf{X}^{(T)}, C_t = i).$$

Summation of this equation over i yields the second conclusion. □

The second conclusion also follows immediately from the matrix expression for the likelihood and the definitions of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$:

$$\begin{aligned}L_T &= \big(\boldsymbol{\delta}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\ldots\boldsymbol{\Gamma}\mathbf{P}(x_t)\big) \big(\boldsymbol{\Gamma}\mathbf{P}(x_{t+1})\ldots\boldsymbol{\Gamma}\mathbf{P}(x_T)\mathbf{1}'\big)\\ &= \boldsymbol{\alpha}_t\boldsymbol{\beta}_t'.\end{aligned}$$

Note that we now have available T routes to the computation of the likelihood L_T, one for each possible value of t. But the route we have used so far (the case $t = T$, yielding $L_T = \boldsymbol{\alpha}_T \mathbf{1}'$) seems the most convenient, as it requires the computation of forward probabilities only, and only a single pass (forward) through the data.

In applying the EM algorithm to HMMs we shall also need the following two properties.

Proposition 5 *Firstly, for $t = 1, \ldots, T$,*

$$\Pr(C_t = j \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \alpha_t(j)\beta_t(j)/L_T; \tag{4.10}$$

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

$$\Pr(C_{t-1} = j, C_t = k \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) = \alpha_{t-1}(j)\,\gamma_{jk}\,p_k(x_t)\,\beta_t(k)/L_T. \tag{4.11}$$

Proof. The first assertion follows immediately from (4.9) above. The second is an application of Equations (B.4) and (B.5) of Appendix B, and the proof proceeds as follows.

$$\begin{aligned}
&\Pr(C_{t-1} = j, C_t = k \mid \mathbf{X}^{(T)} = \mathbf{x}^{(T)}) \\
&= \Pr(\mathbf{X}^{(T)}, C_{t-1} = j, C_t = k)/L_T \\
&= \Pr(\mathbf{X}^{(t-1)}, C_{t-1} = j)\ \Pr(C_t = k \mid C_{t-1} = j)\ \Pr(\mathbf{X}_t^T \mid C_t = k)/L_T \\
&\qquad \text{by (B.4)} \\
&= \alpha_{t-1}(j)\,\gamma_{jk}\Big(\Pr(X_t \mid C_t = k)\ \Pr(\mathbf{X}_{t+1}^T \mid C_t = k)\Big)/L_T \\
&\qquad \text{by (B.5)} \\
&= \alpha_{t-1}(j)\,\gamma_{jk}\,p_k(x_t)\,\beta_t(k)/L_T. \qquad \square
\end{aligned}$$

4.2 The EM algorithm

Since the sequence of states occupied by the Markov-chain component of an HMM is not observed, a very natural approach to parameter estimation in HMMs is to treat those states as missing data and to employ the EM algorithm (Dempster, Laird and Rubin, 1977) in order to find maximum likelihood estimates of the parameters. Indeed the pioneering work of Leonard Baum and his co-authors (see Baum *et al.*, 1970; Baum, 1972; Welch, 2003) on what later were called HMMs was an important precursor of the work of Dempster *et al.*

4.2.1 EM in general

The EM algorithm is an iterative method for performing maximum likelihood estimation when some of the data are missing, and exploits the fact that the complete-data log-likelihood may be straightforward to maximize even if the likelihood of the observed data is not. By ‘complete-data log-likelihood’ (CDLL) we mean the log-likelihood of the parameters of interest $\boldsymbol{\theta}$, based on both the observed data and the missing data.

The algorithm may be described informally as follows (see e.g. Little and Rubin, 2002, pp. 166–168). Choose starting values for the parameters $\boldsymbol{\theta}$ you wish to estimate. Then repeat the following steps.

- **E step** Compute the conditional expectations of the missing data given the observations and given the current estimate of $\boldsymbol{\theta}$. More precisely, compute the conditional expectations of *those functions of* the missing data that appear in the complete-data log-likelihood.
- **M step** Maximize, with respect to $\boldsymbol{\theta}$, the complete-data log-likelihood with the functions of the missing data replaced in it by their conditional expectations.

These two steps are repeated until some convergence criterion has been satisfied, e.g. until the resulting change in $\boldsymbol{\theta}$ is less than some threshold. The resulting value of $\boldsymbol{\theta}$ is then a stationary point of the likelihood of the observed data. In some cases, however, the stationary point reached can be a local (as opposed to global) maximum or a saddle point.

Little and Rubin (p. 168) stress the point that it is not (necessarily) the missing data themselves that are replaced in the CDLL by their conditional expectations, but those functions of the missing data that appear in the CDLL; they describe this as the ‘key idea of EM’.

4.2.2 EM for HMMs

In the case of an HMM it is here convenient to represent the sequence of states $c_1, c_2, \dots, c_T$ followed by the Markov chain by the zero-one random variables defined as follows:

$$u_j(t) = 1 \text{ if and only if } c_t = j, \quad (t = 1, 2, \dots, T)$$

and

$$v_{jk}(t) = 1 \text{ if and only if } c_{t-1} = j \text{ and } c_t = k \quad (t = 2, 3, \dots, T).$$

With this notation, the complete-data log-likelihood of an HMM — i.e. the log-likelihood of the observations $x_1, x_2, \dots, x_T$ plus the missing data $c_1, c_2, \dots, c_T$ — is given by

$$\begin{aligned}\log\left(\Pr(\mathbf{x}^{(T)}, \mathbf{c}^{(T)})\right) &= \log\left(\delta_{c_1} \prod_{t=2}^{T} \gamma_{c_{t-1},c_t} \prod_{t=1}^{T} p_{c_t}(x_t)\right) \\ &= \log \delta_{c_1} + \sum_{t=2}^{T} \log \gamma_{c_{t-1},c_t} + \sum_{t=1}^{T} \log p_{c_t}(x_t).\end{aligned}$$

Hence the CDLL is

$$\begin{aligned}
&\log\left(\Pr(\mathbf{x}^{(T)},\mathbf{c}^{(T)})\right)\\
&= \sum_{j=1}^{m} u_j(1)\log\delta_j + \sum_{j=1}^{m}\sum_{k=1}^{m}\left(\sum_{t=2}^{T} v_{jk}(t)\right)\log\gamma_{jk}\\
&\qquad + \sum_{j=1}^{m}\sum_{t=1}^{T} u_j(t)\log p_j(x_t) \qquad (4.12)\\
&= \text{term } 1 + \text{term } 2 + \text{term } 3.
\end{aligned}$$

Here $\boldsymbol{\delta}$ is to be understood as the *initial* distribution of the Markov chain, the distribution of C_1, not necessarily the stationary distribution. Of course it is not reasonable to try to estimate the initial distribution from just one observation at time 1, especially as the state of the Markov chain itself is not observed. It is therefore interesting to see how the EM algorithm responds to this unreasonable request: see Section 4.2.4. We shall later (Section 4.2.5) make the additional assumption that the Markov chain is stationary, and not merely homogeneous; $\boldsymbol{\delta}$ will then denote the stationary distribution implied by $\boldsymbol{\Gamma}$, and the question of estimating $\boldsymbol{\delta}$ will fall away.

The EM algorithm for HMMs proceeds as follows.

- **E step** Replace all the quantities $v_{jk}(t)$ and $u_j(t)$ by their conditional expectations given the observations $\mathbf{x}^{(T)}$ (and given the current parameter estimates):

$$\hat{u}_j(t) = \Pr(C_t=j \mid \mathbf{x}^{(T)}) = \alpha_t(j)\beta_t(j)/L_T; \qquad (4.13)$$

  and

$$\hat{v}_{jk}(t) = \Pr(C_{t-1}=j, C_t=k \mid \mathbf{x}^{(T)}) = \alpha_{t-1}(j)\,\gamma_{jk}\,p_k(x_t)\,\beta_t(k)/L_T. \qquad (4.14)$$

  (See Section 4.1.3, Equations (4.10) and (4.11), for justification of the above equalities.) Note that in this context we need the forward probabilities as computed for an HMM that does *not* assume stationarity of the underlying Markov chain $\{C_t\}$; the backward probabilities are however not affected by the stationarity or otherwise of $\{C_t\}$.

- **M step** Having replaced $v_{jk}(t)$ and $u_j(t)$ by $\hat{v}_{jk}(t)$ and $\hat{u}_j(t)$, maximize the CDLL, expression (4.12), with respect to the three sets of parameters: the initial distribution $\boldsymbol{\delta}$, the transition probability matrix $\boldsymbol{\Gamma}$, and the parameters of the state-dependent distributions (e.g. $\lambda_1, \ldots, \lambda_m$ in the case of a simple Poisson–HMM).

Examination of (4.12) reveals that the M step splits neatly into three separate maximizations, since (of the parameters) term 1 depends only on the initial distribution $\boldsymbol{\delta}$, term 2 on the transition probability matrix

$\mathbf{\Gamma}$, and term 3 on the 'state-dependent parameters'. We must therefore maximize:

1. $\sum_{j=1}^m \hat{u}_j(1) \log \delta_j$ with respect to $\boldsymbol{\delta}$;
2. $\sum_{j=1}^m \sum_{k=1}^m \left(\sum_{t=2}^T \hat{v}_{jk}(t)\right) \log \gamma_{jk}$ with respect to $\mathbf{\Gamma}$; and
3. $\sum_{j=1}^m \sum_{t=1}^T \hat{u}_j(t) \log p_j(x_t)$ with respect to the state-dependent parameters. Notice here that the only parameters on which the term $\sum_{t=1}^T \hat{u}_j(t) \log p_j(x_t)$ depends are those of the j th state-dependent distribution, p_j; this further simplifies the problem.

The solution is as follows.

1. Set $\delta_j = \hat{u}_j(1) / \sum_{j=1}^m \hat{u}_j(1) = \hat{u}_j(1)$. (See Exercise 11 of Chapter 1 for justification.)
2. Set $\gamma_{jk} = f_{jk} / \sum_{k=1}^m f_{jk}$, where $f_{jk} = \sum_{t=2}^T \hat{v}_{jk}(t)$. (Apply the result of Exercise 11 of Chapter 1 to each row.)
3. The maximization of the third term may be easy or difficult, depending on the nature of the state-dependent distributions assumed. It is essentially the standard problem of maximum likelihood estimation for the distributions concerned. In the case of Poisson and normal distributions, closed-form solutions are available: see Section 4.2.3. In some other cases, e.g. the gamma distributions and the negative binomial, numerical maximization will be necessary to carry out this part of the M step.

From point 2 above, we see that it is not the quantities $\hat{v}_{jk}(t)$ themselves that are needed, but their sums f_{jk}. It is worth noting that the computation of the forward and backward probabilities is susceptible to under- or overflow error, as are the computation and summation of the quantities $\hat{v}_{jk}(t)$. In applying EM as described here, precautions (e.g. scaling) therefore have to be taken in order to prevent, or at least reduce the risk of, such error. Code for computing the logarithms of the forward and backward probabilities of a Poisson–HMM, and for computing MLEs via the EM algorithm, appears in A.2.2 and A.2.3.

4.2.3 M step for Poisson– and normal–HMMs

Here we give part 3 of the M step explicitly for the cases of Poisson and normal state-dependent distributions. The state-dependent part of the CDLL (term 3 of expression (4.12)) is

$$\sum_{j=1}^m \sum_{t=1}^T \hat{u}_j(t) \log p_j(x_t).$$

For a Poisson–HMM, $p_j(x) = e^{-\lambda_j}\lambda_j^x/x!$, so in that case term 3 is maximized by setting

$$0 = \sum_t \hat{u}_j(t)(-1 + x_t/\lambda_j);$$

that is, by

$$\hat{\lambda}_j = \sum_{t=1}^{T} \hat{u}_j(t)x_t \Big/ \sum_{t=1}^{T} \hat{u}_j(t).$$

For a normal–HMM the state-dependent density is of the form $p_j(x) = (2\pi\sigma_j^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_j^2}(x-\mu_j)^2\right)$, and the maximizing values of the state-dependent parameters μ_j and σ_j^2 are

$$\hat{\mu}_j = \sum_{t=1}^{T} \hat{u}_j(t)x_t \Big/ \sum_{t=1}^{T} \hat{u}_j(t),$$

and

$$\hat{\sigma}_j^2 = \sum_{t=1}^{T} \hat{u}_j(t)(x_t - \hat{\mu}_j)^2 \Big/ \sum_{t=1}^{T} \hat{u}_j(t).$$

4.2.4 Starting from a specified state

What is sometimes done, however, notably by Leroux and Puterman (1992), is instead to condition the likelihood of the observations — i.e. the ‘incomplete-data’ likelihood — on the Markov chain starting in a particular state: that is, to assume that $\boldsymbol{\delta}$ is a unit vector $(0, \dots, 0, 1, 0, \dots, 0)$ rather than a vector whose components δ_j all require estimation. But, since it is known (see Levinson, Rabiner and Sondhi, 1983, p. 1055) that, at a maximum of the likelihood, $\boldsymbol{\delta}$ is one of the m unit vectors, maximizing the conditional likelihood of the observations over the m possible starting states is equivalent to maximizing the unconditional likelihood over $\boldsymbol{\delta}$; see also Exercise 1. Furthermore, it requires considerably less computational effort. If one wishes to use EM in order to fit an HMM in which the Markov chain is not assumed stationary, this does seem to be the most sensible approach. In our EM examples, however, we shall treat $\boldsymbol{\delta}$ as a vector of parameters requiring estimation, as it is instructive to see what emerges.

4.2.5 EM for the case in which the Markov chain is stationary

Now assume in addition that the underlying Markov chain is stationary, and not merely homogeneous. This is often a desirable assumption in time series applications. The initial distribution $\boldsymbol{\delta}$ is then such that $\boldsymbol{\delta} =$

$\boldsymbol{\delta\Gamma}$ and $\boldsymbol{\delta 1}' = 1$, or equivalently

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

with $\mathbf{U}$ being a square matrix of ones. In this case, $\boldsymbol{\delta}$ is completely determined by the transition probabilities $\boldsymbol{\Gamma}$, and the question of estimating $\boldsymbol{\delta}$ falls away. However, the M step then gives rise to the following optimization problem: maximize, with respect to $\boldsymbol{\Gamma}$, the sum of terms 1 and 2 of expression (4.12), i.e. maximize

$$\sum_{j=1}^{m} \hat{u}_j(1) \log \delta_j + \sum_{j=1}^{m} \sum_{k=1}^{m} \left(\sum_{t=2}^{T} \hat{v}_{jk}(t) \right) \log \gamma_{jk}. \qquad (4.15)$$

Notice that here term 1 also depends on $\boldsymbol{\Gamma}$. Even in the case of only two states, analytic maximization would require the solution of a pair of quadratic equations in two variables, viz. two of the transition probabilities; see Exercise 3. Numerical solution is in general therefore needed for this part of the M step if stationarity is assumed. This is a slight disadvantage of the use of EM, as the stationary version of the models is important in a time series context.

4.3 Examples of EM applied to Poisson–HMMs

4.3.1 Earthquakes

We now present two- and three-state models fitted by the EM algorithm, as described above, to the earthquakes data. For the two-state model, the starting values of the off-diagonal transition probabilities are taken to be 0.1, and the starting value of $\boldsymbol{\delta}$, the initial distribution, is (0.5, 0.5). Since 19.36 is the sample mean, 10 and 30 are plausible starting values for the state-dependent means λ_1 and λ_2.

In the tables shown, 'iteration 0' refers to the starting values, and 'stationary model' to the parameter values and log-likelihood of the comparable stationary model fitted via `nlm` by direct numerical maximization.

Several features of Table 4.1 are worth noting. Firstly, the likelihood value of the stationary model is slightly lower than that fitted here by EM (i.e. $-l$ is higher). This is to be expected, as constraining the initial distribution $\boldsymbol{\delta}$ to be the stationary distribution can only decrease the maximal value of the likelihood. Secondly, the estimates of the transition probabilities and the state-dependent means are not identical for the two models, but close; this, too, is to be expected. Thirdly, although we know from Section 4.2.4 that $\boldsymbol{\delta}$ will approach a unit vector, it is noticeable just how quickly, starting from (0.5, 0.5), it approaches $(1, 0)$.

Table 4.2 displays similar information for the corresponding three-state models. In this case as well, the starting values of the off-diagonal

Table 4.1 *Two-state model for earthquakes, fitted by EM.*

Iteration	γ_{12}	γ_{21}	λ_1	λ_2	δ_1	$-l$
0	0.100000	0.10000	10.000	30.000	0.50000	413.27542
1	0.138816	0.11622	13.742	24.169	0.99963	343.76023
2	0.115510	0.10079	14.090	24.061	1.00000	343.13618
30	0.071653	0.11895	15.419	26.014	1.00000	341.87871
50	0.071626	0.11903	15.421	26.018	1.00000	341.87870
convergence	0.071626	0.11903	15.421	26.018	1.00000	341.87870
stationary model	0.065961	0.12851	15.472	26.125	0.66082	342.31827

Table 4.2 *Three-state model for earthquakes, fitted by EM.*

Iteration	λ_1	λ_2	λ_3	δ_1	δ_2	$-l$
0	10.000	20.000	30.000	0.33333	0.33333	342.90781
1	11.699	19.030	29.741	0.92471	0.07487	332.12143
2	12.265	19.078	29.581	0.99588	0.00412	330.63689
30	13.134	19.713	29.710	1.00000	0.00000	328.52748
convergence	13.134	19.713	29.710	1.00000	0.00000	328.52748
stationary model	13.146	19.721	29.714	0.44364	0.40450	329.46028

transition probabilities are all taken to be 0.1 and the starting $\boldsymbol{\delta}$ is uniform over the states.

We now present more fully the 'EM' and the stationary versions of the three-state model, which are only summarized in Table 4.2.

- Three-state model with initial distribution (1,0,0), fitted by EM:

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.9393 & 0.0321 & 0.0286 \\ 0.0404 & 0.9064 & 0.0532 \\ 0.0000 & 0.1903 & 0.8097 \end{pmatrix},$$

$$\boldsymbol{\lambda} = (13.134, 19.713, 29.710).$$

- Three-state model based on stationary Markov chain, fitted by direct numerical maximization:

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.9546 & 0.0244 & 0.0209 \\ 0.0498 & 0.8994 & 0.0509 \\ 0.0000 & 0.1966 & 0.8034 \end{pmatrix},$$

$$\boldsymbol{\delta} = (0.4436, 0.4045, 0.1519),$$

$$\text{and } \boldsymbol{\lambda} = (13.146, 19.721, 29.714).$$

One of the many things that can be done by means of the **R** package `msm` (Jackson *et al.*, 2003) is to fit HMMs with a range of state-dependent distributions, including the Poisson. The default initial distribution used by `msm` assigns a probability of 1 to state 1; the resulting models are therefore directly comparable to the models which we have fitted by EM. The two-state model fitted by `msm` corresponds closely to our model as given in Table 4.1; it has $-l = 341.8787$, state-dependent means 15.420 and 26.017, and transition probability matrix

$$\begin{pmatrix} 0.9283 & 0.0717 \\ 0.1189 & 0.8811 \end{pmatrix}.$$

The three-state models do not correspond quite so closely. The 'best' three-state model we have found by `msm` has $-l = 328.6208$ (cf. 328.5275), state-dependent means 13.096, 19.708 and 29.904, and transition probability matrix

$$\begin{pmatrix} 0.9371 & 0.0296 & 0.0333 \\ 0.0374 & 0.9148 & 0.0479 \\ 0.0040 & 0.1925 & 0.8035 \end{pmatrix}.$$

4.3.2 Foetal movement counts

Leroux and Puterman (1992) used EM to fit (among other models) Markov-dependent mixtures to a time series of unbounded counts, a series which has subsequently been analysed by Chib (1996), Robert and Titterington (1998), Robert and Casella (1999, p. 432) and Scott (2002). The series consists of counts of movements by a foetal lamb in 240 consecutive 5-second intervals, and was taken from Wittmann, Rurak and Taylor (1984).

We present here a two-state HMM fitted by EM, and two HMMs fitted by direct numerical maximization of the likelihood, and we compare these with the models in Table 4 of Leroux and Puterman and Table 2 of Robert and Titterington. Robert and Titterington omit any comparison with the results of Leroux and Puterman, a comparison which might have been informative: the estimates of λ_1 differ, and Leroux and Puterman's likelihood value is somewhat higher, as are our values.

The models of Leroux and Puterman start with probability 1 from one of the states; i.e. the initial distribution of the Markov chain is a unit vector. But it is of course easy to fit such models by direct numerical maximization if one so wishes. In any code which takes the initial distribution of the Markov chain to be the stationary distribution, one merely replaces that stationary distribution by the relevant unit vector. Our Table 4.3 presents (*inter alia*) three models we have fitted, the second of which is of this kind. The first we have fitted by direct numerical

Table 4.3 *Six two-state models fitted to the foetal movement counts time series.*

	DNM (stat.)	DNM (state 2)	EM (this work)	EM (L&P)	SEM	prior feedback
λ_1	3.1148	3.1007	3.1007	3.1006	2.93	2.84
λ_2	0.2564	0.2560	0.2560	0.2560	0.26	0.25
γ_{12}	0.3103	0.3083	0.3083	0.3083	0.28	0.32
γ_{21}	0.0113	0.0116	0.0116	0.0116	0.01	0.015
$-l^*$	–	150.7007	150.7007	150.70	–	–
$-l$	177.5188	177.4833	177.4833	(177.48)	(177.58)	(177.56)
$10^{78}L$	8.0269	8.3174	8.3174	(8.3235)	7.539	7.686

Key to notation and abbreviations:

DNM (stat.)	model starting from stationary distribution of the Markov chain, fitted by direct numerical maximization
DNM (state 2)	model starting from state 2 of the Markov chain, fitted by direct numerical maximization
EM (this work)	model fitted by EM, as described in Sections 4.2.2–4.2.3
EM (L&P)	model fitted by EM by Leroux and Puterman (1992)
SEM	model fitted by stochastic EM algorithm of Chib (1996); from Table 2 of Robert and Titterington (1998)
prior feedback	model fitted by maximum likelihood via ‘prior feedback’; from Table 2 of Robert and Titterington (1998)
l^*	log-likelihood omitting constant terms
l	log-likelihood
L	likelihood
–	not needed; deliberately omitted

Figures appearing in brackets in the last three columns of this table have been deduced from figures published in the works cited, but do not themselves appear there; they are at best as accurate as the figures on which they are based. For instance, $10^{78}L = 8.3235$ is based on $-l^* = 150.70$. The figures 150.70, 7.539 and 7.686 in those three columns are exactly as published.

maximization of the likelihood based on the initial distribution being the stationary distribution, the second by direct numerical maximization of the likelihood starting from state 2, and the third by EM. In Table 4.3 they are compared to three models appearing in the published literature. The very close correspondence between the models we fitted by EM and by direct maximization (starting in state 2) is reassuring.

The results displayed in the table suggest that, if one wishes to fit models by maximum likelihood, then EM and direct numerical maximization are superior to the other methods considered. Robert and Titterington suggest that Chib's algorithm may not have converged, or may have converged to a different local optimum of the likelihood; their prior feedback results appear also to be suboptimal, however.

4.4 Discussion

As Bulla and Berzel (2008) point out, researchers and practitioners tend to use either EM or direct numerical maximization, but not both, to perform maximum likelihood estimation in HMMs, and each approach has its merits. However, one of the merits rightly claimed for EM in some generality turns out to be illusory in the context of HMMs. McLachlan and Krishnan (1997, p. 33) state of the EM algorithm in general — i.e. not specifically in the context of HMMs — that it is '[...] easy to program, since no evaluation of the likelihood nor its derivatives is involved.' If one applies EM as described above, one has to compute both the forward and the backward probabilities; on the other hand one needs only the forward probabilities to compute the likelihood, which can then be maximized numerically. In effect, EM does all that is needed to compute the likelihood via the forward probabilities, and then does more. Especially if one has available an optimization routine, such as `nlm`, `optim` or `constrOptim` in **R**, which does not demand the specification of derivatives, ease of programming seems in the present context to be more a characteristic of direct numerical maximization than of EM.

In our experience, it is a major advantage of direct numerical maximization without analytical derivatives that one can, with a minimum of programming effort, repeatedly modify a model in an interactive search for the best model. Often all that is needed is a small change to the code that evaluates the likelihood. It is also usually straightforward to replace one optimizer by another if an optimizer fails, or if one wishes to check in any way the output of a particular optimizer.

Note the experience reported by Altman and Petkau (2005). In their applications direct maximization of the likelihood produced the MLEs far more quickly than did the EM algorithm. See also Turner (2008), who provides a detailed study of direct maximization by the Levenberg–Marquardt algorithm. In two examples he finds that this algorithm is much faster (in the sense of CPU time) than is EM, and it is also clearly faster than `optim`, both with and without the provision of analytical derivatives. In our opinion the disadvantage of using analytical derivatives in exploratory modelling is the work involved in recoding those derivatives, and checking the code, when one alters a model. Of course for standard models such as the Poisson–HMM, which are likely to be used repeatedly, the advantage of having efficient code would make such labour worthwhile. Cappé *et al.* (2005, p. 358) provide a discussion of the relative merits of EM and direct maximization of the likelihood of an HMM by gradient-based methods.

Exercises

1.(a) Suppose $L_i > 0$ for $i = 1, 2, \ldots, m$. Maximize $L = \sum_{i=1}^{m} a_i L_i$ over $a_i \geq 0$, $\sum_{i=1}^{m} a_i = 1$.

(b) Consider an HMM with initial distribution $\boldsymbol{\delta}$, and consider the likelihood as a function of $\boldsymbol{\delta}$. Show that, at a maximum of the likelihood, $\boldsymbol{\delta}$ is a unit vector.

2. Consider the example on pp. 186–187 of Visser, Raijmakers and Molenaar (2002). There a series of length 1000 is simulated from an HMM with states S_1 and S_2 and the three observation symbols 1, 2 and 3. The transition probability matrix is

$$\mathbf{A} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix},$$

the initial probabilities are $\boldsymbol{\pi} = (0.5, 0.5)$, and the state-dependent distribution in state i is row i of the matrix

$$\mathbf{B} = \begin{pmatrix} 0.7 & 0.0 & 0.3 \\ 0.0 & 0.4 & 0.6 \end{pmatrix}.$$

The parameters $\mathbf{A}$, $\mathbf{B}$ and $\boldsymbol{\pi}$ are then estimated by EM; the estimates of $\mathbf{A}$ and $\mathbf{B}$ are close to $\mathbf{A}$ and $\mathbf{B}$, but that of $\boldsymbol{\pi}$ is $(1, 0)$. This estimate for $\boldsymbol{\pi}$ is explained as follows: 'The reason for this is that the sequence of symbols that was generated actually starts with the symbol 1 which can only be produced from state S_1.'

Do you agree with the above statement? What if the probability of symbol 1 in state S_2 had been (say) 0.1 rather than 0.0?

3. Consider the fitting by EM of a two-state HMM based on a *stationary* Markov chain. In the M step, the sum of terms 1 and 2 must be maximized with respect to $\boldsymbol{\Gamma}$; see the expression labelled (4.15).

Write term 1 + term 2 as a function of γ_{12} and γ_{21}, the off-diagonal transition probabilities, and differentiate to find the equations satisfied by these probabilities at a stationary point. (You should find that the equations are quadratic in both γ_{12} and γ_{21}.)

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

Use the EM algorithm to fit Poisson–HMMs with two, three and four states to these data.

5. Let $\{X_t\}$ be an HMM on m states.

(a) Suppose the state-dependent distributions are binomial. More precisely, assume that

$$\Pr(X_t = x \mid C_t = j) = \binom{n_t}{x} p_j^x (1 - p_j)^{n_t - x}.$$

Find the value for p_j that will maximize the third term of Equation (4.12). (This is needed in order to carry out the M step of EM for a binomial–HMM.)

(b) Now suppose instead that the state-dependent distributions are exponential, with means $1/\lambda_j$. Find the value for λ_j that will maximize the third term of Equation (4.12).

6. Modify the code given in A.2 for Poisson–HMMs, in order to fit normal–, binomial–, and exponential–HMMs by EM.