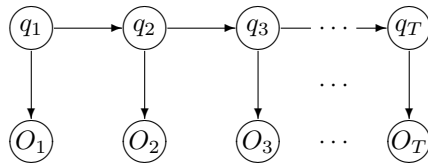


Chapter 1

Hidden Markov Models

1.1 Introduction

A hidden Markov model (HMM) is a statistical model for ordered data. The observed data is assumed to have been generated by a unobservable statistical process of a particular form. This process is such that each observation is coincident with the system being in a particular state. Furthermore it is a first order Markov process: the next state is dependent only the current state. The model is completely described by the initial state probabilities, the first order Markov chain state-to-state transition probabilities, and the probability distributions of observable outputs associated with each state.



Partially observed Markov chain.

Figure 1.1: A representation of the hidden Markov model, with hidden nodes in underlying system states q , and observable variables O .

1.2 Notation

Our notation is similar to that employed by Rabiner [11] and is as follows: a hidden Markov model λ with N states is composed of initial state probabilities $\pi = (\pi_1, \dots, \pi_N)$, state-to-state transition probabilities $A = (a_{11}, \dots, a_{ij}, \dots, a_{NN})$, and the observable output probability distributions $B = (b_1, \dots, b_N)$. The observable outputs can be either discrete or continuous. In the discrete case, the

output probability distributions are denoted by $b_i(m)$, where m is one of M discrete output symbols. In the continuous case, the output probability distributions are denoted by $b_i(y, \theta_{i1}, \dots, \theta_{ij}, \dots, \theta_{iM})$ where y is the real-valued observable output (scalar or vector) and the θ_{ij} s are the parameters describing the output probability distribution. For the normal distribution we have $b_i(y, \mu_i, \Sigma_i)$.

1.3 Model optimization problem

For the series of observations $O = O_1 O_2 \cdots O_T$, we consider the possible model state sequences $Q = q_1 q_2 \cdots q_T$ to which this series of observations could be assigned. For a given fixed state sequence Q , the probability of the observation sequence O is given by

$$P(O|Q, \lambda) = \prod_{t=1}^T P(O_t|q_t, \lambda). \quad (1.1)$$

Assuming statistical independence of observations,

$$P(O|Q, \lambda) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \quad (1.2)$$

The probability of the given state sequence Q is

$$P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}. \quad (1.3)$$

The joint probability of O and Q is the product of the above, so that

$$P(O, Q|\lambda) = P(O|Q, \lambda) P(Q|\lambda), \quad (1.4)$$

and the probability of O given the model is obtained by summing this joint probability over all possible state sequences Q :

$$P(O|\lambda) = \sum_{\text{all } Q=q_1 q_2 \cdots q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T). \quad (1.5)$$

Although other optimization criteria are possible, most commonly we wish to optimize the model parameters so as to maximize this likelihood $P(O|\lambda)$. We

can pose this as non-convex, non-linear optimization problem:

$$\begin{aligned}
& \text{Maximize :} && P(O|\lambda) \\
& \text{Subject to :} && \sum_{i=1}^N \pi_i = 1 \\
& && \pi_i \geq 0, \quad i = 1, \dots, N \\
& && \sum_{j=1}^N a_{ij} = 1, \quad i = 1, \dots, N \\
& && a_{ij} \geq 0, \quad i = 1, \dots, N, \quad j = 1, \dots, N \\
& && \sum_{m=1}^M b_i(m) = 1, \quad i = 1, \dots, N \\
& && b_i(m) \geq 0, \quad i = 1, \dots, N, \quad m = 1, \dots, M.
\end{aligned} \tag{1.6}$$

Note that the above is for the discrete output case. In the case of continuous outputs, the last two constraints are replaced by

$$\begin{aligned}
& \int_Y b_i(y) dy = 1, \quad i = 1, \dots, N \\
& b_i(y) \geq 0, \quad i = 1, \dots, N, \quad y \in Y.
\end{aligned} \tag{1.7}$$

This problem is often presented in terms of equivalent problem of maximizing the *log likelihood* $\log P(O|\lambda)$.

1.4 Expectation-Maximization

The most common optimization technique employed to solve this problem is the Expectation-Maximization (EM) algorithm [7, 10]. We can pose the EM algorithm generally as follows: we wish to maximize a likelihood $P(\lambda)$ where λ is a set of model parameters. Given $p(x, \lambda)$, a positive real-valued function on $x \times \Lambda$ measurable in x for fixed λ with measure μ , we define

$$P(\lambda) = E[p(x, \lambda)|\lambda] = \int_X p(x, \lambda) d\mu(x) \tag{1.8}$$

and

$$Q(\lambda, \lambda') = E[\log p(x, \lambda')|\lambda] = \int_X p(x, \lambda) \log p(x, \lambda') d\mu(x). \tag{1.9}$$

Here x is the so-called *hidden variable*, while $p(x, \lambda)$ is often referred to as the *complete data likelihood*. The function Q is often referred to as the *Q-function*. Note that the function p may be a function of the observable outputs y as well as the parameters of the model λ , so $p = p(x, y, \lambda)$. In this case, the integrals are over $X \rightarrow Y(X)$.

Assume $Q(\lambda, \bar{\lambda}) \geq Q(\lambda, \lambda)$. Then $P(\bar{\lambda}) \geq P(\lambda)$. Proof:

$$\begin{aligned} \log P(\bar{\lambda})/P(\lambda) &= \log \int_X p(x, \bar{\lambda}) d\mu(x)/P(\lambda) \\ &= \log \int_X [p(x, \lambda) d\mu(x)/P(\lambda)] p(x, \bar{\lambda})/p(x, \lambda) \\ &\geq \int_X [p(x, \lambda) d\mu(x)/P(\lambda)] \log [p(x, \bar{\lambda})/p(x, \lambda)] \\ &= (P(\lambda))^{-1} [Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda)] \geq 0. \end{aligned}$$

From this we can show that for a transformation \mathcal{F} that if $\mathcal{F}(\lambda)$ is a critical point of $Q(\lambda, \lambda')$ as a function of λ' , then the fixed points of \mathcal{F} are critical points of P . This gives us the EM algorithm:

1. Start with $k = 0$ and pick a starting $\lambda^{(k)}$.
2. Calculate $Q(\lambda^{(k)}, \lambda)$ (expectation step).
3. Maximize $Q(\lambda^{(k)}, \lambda)$ over λ (maximization step). This gives us the transformation \mathcal{F} .
4. Set $\lambda^{(k+1)} = \mathcal{F}(\lambda^{(k)})$. If $Q(\lambda^{(k+1)}, \lambda) - Q(\lambda^{(k)}, \lambda)$ is below some threshold, stop. Otherwise, go to step 2.

Note that this method is inherently sensitive to the initial conditions $\lambda^{(0)}$, and only guarantees eventual convergence to a local maxima of the objective function, not the global maximum. Nevertheless, it is widely used in practice and often achieves good results.

1.5 Optimization procedure for the HMM

We now present a procedure for calculating the optimal HMM parameters, based on that first suggested by Baum and colleagues [1–5]. For the hidden Markov model, we have

$$p(q, O, \lambda) = \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T), \quad (1.10)$$

with $P(\lambda) = E[p(q, O, \lambda) | \lambda]$ defined as in (1.5). If we let z be a set of state-indicator indicator vectors $z = (z_1, \dots, z_T)$ such that $z_{it} = 1$ if $q_t = i$, $z_{it} = 0$ otherwise, then we can represent the complete data log likelihood as

$$\sum_{i=1}^N z_{i1} \log \pi_i + \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^{T-1} z_{it} z_{j,t+1} \log a_{ij} + \sum_{i=1}^N \sum_{t=1}^T z_{it} \log b_i(O_t). \quad (1.11)$$

From this we can calculate

$$Q(\lambda, \lambda^{(k)}) = \sum_{i=1}^N \tau_{i1}^{(k)} \log \pi_i + \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij} + \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)} \log b_i(O_t) \quad (1.12)$$

where

$$\tau_{ijt} = P(Z_{it} = 1, Z_{j,t+1} = 1 | O, \lambda) \quad t = 1, \dots, T-1, \quad (1.13)$$

$$\tau_{it} = P(Z_{it} = 1 | O, \lambda) \quad t = 1, \dots, T, \quad (1.14)$$

and Z is a probabilistic component indicator variable analogous to z .

We wish to maximize $Q(\lambda, \lambda^{(k)})$ over λ . We can view Q as the sum of three separable components, $Q = Q_1 + Q_2 + Q_3$:

$$Q_1(\lambda, \lambda^{(k)}) = \sum_{i=1}^N \tau_{i1}^{(k)} \log \pi_i, \quad (1.15)$$

$$Q_2(\lambda, \lambda^{(k)}) = \sum_{i=1}^N \sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}^{(k)} \log a_{ij}, \quad (1.16)$$

$$Q_3(\lambda, \lambda^{(k)}) = \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)} \log b_i(O_t). \quad (1.17)$$

Maximization of each component may be pursued separately. However, a direct solution by calculation of the critical points of the first two components is not possible. For instance,

$$\frac{\partial Q_1}{\partial \pi_i} = \frac{\partial \sum_{i=1}^N \tau_{i1}^{(k)} \log \pi_i}{\partial \pi_i} = \frac{\tau_{i1}^{(k)}}{\pi_i} = 0 \quad (1.18)$$

is clearly not useful, and derivatives of Q_2 fare similarly.

Instead we will solve the general convex optimization problem

$$\begin{aligned} \text{Minimize :} \quad & f(x) = - \sum_{i=1}^N c_i \log x_i \\ \text{Subject to :} \quad & \sum_{i=1}^N x_i = 1, \end{aligned} \quad (1.19)$$

with constants $c_i \geq 0$. First, we calculate the Lagrangian

$$L(x, \nu) = - \sum_{i=1}^N c_i \log x_i + \nu \left(\sum_{i=1}^N x_i - 1 \right), \quad (1.20)$$

which has a maximum in the x_i s at $x_i = c_i/\nu$. The dual problem is then

$$\text{Maximize :} \quad g(\nu) = - \sum_{i=1}^N c_i \log(c_i/\nu) + \sum_{i=1}^N c_i - \nu. \quad (1.21)$$

The maximum of the dual problem can be found by setting the derivative equal to zero and solving. We find that $\nu^* = \sum_{i=1}^N c_i$ is the maximum, with $g(\nu^*) =$

$-\sum_{i=1}^N c_i \log(c_i / \sum_{i=1}^N c_i)$. Since $f(x) = g(\nu^*)$ is feasible with $x_i = c_i / \sum_{i=1}^N c_i$, this is the minimizing solution to the primal problem.

Since τ_{i1} is dependent only on the first observation, we can calculate using Bayes' Theorem:

$$\tau_{i1} = \frac{\pi_i b_i(O_1)}{\sum_{j=1}^N \pi_j b_j(O_1)}. \quad (1.22)$$

Our solution to (1.19) gives us the values π_i which maximize Q_1 :

$$\pi_i = \frac{\tau_{i1}^{(k)}}{\sum_{j=1}^N \tau_{j1}^{(k)}} = \tau_{i1}^{(k)} = \frac{\pi_i^{(k)} b_i^{(k)}(O_1)}{\sum_{j=1}^N \pi_j^{(k)} b_j^{(k)}(O_1)}. \quad (1.23)$$

Similarly, we can use the solution to (1.19) to give us the values a_{ij} which maximize Q_2 :

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}{\sum_{j=1}^N \sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}. \quad (1.24)$$

Noting that $\tau_{it} = \sum_{j=1}^N \tau_{ijt}$ for $t = 1, \dots, T-1$, we have

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \tau_{ijt}^{(k)}}{\sum_{t=1}^{T-1} \tau_{it}^{(k)}}. \quad (1.25)$$

If the outputs of the model are discrete, we can apply our solution to (1.19) once more by noting that

$$Q_3 = \sum_{i=1}^N \sum_{t=1}^T \sum_{m=1}^M \tau_{it}^{(k)} \delta(O_t - m) \log b_i(m), \quad (1.26)$$

and so therefore the output probability distributions that maximize Q_3 are

$$b_i(m) = \frac{\sum_{t=1}^T \tau_{it}^{(k)} \delta(O_t - m)}{\sum_{t=1}^T \tau_{it}^{(k)}}. \quad (1.27)$$

where m is a possible output symbol. If the outputs of the model are continuous, then there is no general explicit formula for the maximum value of the output distribution parameters. However, for certain special forms of the output distribution, the maximizing values can be calculated analytically. For example, in the case of multivariate Gaussian output distributions ($b_i(y) = n(\det(\Sigma_i))^{-1/2} \exp(-(y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i) / 2)$, where n is a normalizing factor), we have:

$$\begin{aligned} Q_3 &= \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)} \left(\log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2} (O_t - \mu_i)^T \Sigma_i^{-1} (O_t - \mu_i) \right) \\ &= \sum_{i=1}^N \sum_{t=1}^T \tau_{it}^{(k)} \left(\log n - \frac{1}{2} \log \det(\Sigma_i) - \frac{1}{2} (m_i - \mu_i)^T \Sigma_i^{-1} (m_i - \mu_i) \right. \\ &\quad \left. - \frac{1}{2} (O_t - m_i)^T \Sigma_i^{-1} (O_t - m_i) \right), \end{aligned} \quad (1.28)$$

where $m_i = \sum_{t=1}^T \tau_{it}^{(k)} O_t / \sum_{t=1}^T \tau_{it}^{(k)}$. Let

$$S_i = \frac{\sum_{t=1}^T \tau_{it}^{(k)} (O_t - m_i)(O_t - m_i)^T}{\sum_{t=1}^T \tau_{it}^{(k)}}. \quad (1.29)$$

Then

$$Q_3 = \sum_{t=1}^T \sum_{i=1}^N \tau_{it}^{(k)} \left(\log n + \frac{1}{2} \log \det(\Sigma_i^{-1}) - \frac{1}{2} (m_i - \mu_i)^T \Sigma_i^{-1} (m_i - \mu_i) - \frac{1}{2} \text{Tr} \Sigma_i^{-1} S_i \right). \quad (1.30)$$

Since Σ_i is positive definite, we see that Q_3 is maximized in the μ_i s when

$$\mu_i = m_i = \frac{\sum_{t=1}^T \tau_{it}^{(k)} O_t}{\sum_{t=1}^T \tau_{it}^{(k)}}. \quad (1.31)$$

Given a maximizing solution for μ_i , we can solve for Σ_i directly by taking the derivative. For a D -by- D matrix A , let the matrix B be such that

$$\{B\}_{ij} = \text{cof}_{ji}(A). \quad (1.32)$$

Then we have

$$\begin{aligned} AB &= \det(A)I \\ B &= \det(A)A^{-1} \\ \{B\}_{ij} &= \det(A)\{A^{-1}\}_{ij}, \end{aligned} \quad (1.33)$$

and therefore

$$\frac{\partial \det(A)}{\partial \{A\}_{ij}} = \frac{\partial}{\partial \{A\}_{ij}} \sum_{i=1}^D \{A\}_{ij} \text{cof}_{ij}(A) = \text{cof}_{ij}(A) = \{B\}_{ji} = \det(A)\{A^{-1}\}_{ji}, \quad (1.34)$$

and

$$\frac{\partial \log \det(A)}{\partial \{A\}_{ij}} = \{A^{-1}\}_{ji}. \quad (1.35)$$

Using these relations, we calculate the derivative of Q_3 with respect to each element of the Σ_i^{-1} s (neglecting constant factors) and set the result equal to zero:

$$\frac{\partial Q_3}{\partial \{\Sigma_i^{-1}\}_{ab}} = \{\Sigma_i\}_{ba} - \{S_i\}_{ba} = 0. \quad (1.36)$$

From this we see that Q_3 has a critical point in the Σ_i s at

$$\Sigma_i = S_i = \frac{\sum_{t=1}^T \tau_{it}^{(k)} (O_t - \mu_i^{(k+1)})(O_t - \mu_i^{(k+1)})^T}{\sum_{t=1}^T \tau_{it}^{(k)}}. \quad (1.37)$$

Since Q_3 is concave this is a global maximum.

How do we calculate the probabilities τ_{it} and τ_{ijt} ? To do so, we make use of the lattice structure of the HMM to perform an iterative calculation, known as the *forward-backward* procedure. Consider the forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(O_1 \cdots O_t, Z_{it} = 1 | \lambda). \quad (1.38)$$

This is the probability of observing the partial sequence $O_1 \cdots O_t$ and that the system is in state i at time t , given the model λ . We can solve for $\alpha_t(i)$ inductively as follows:

1. Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad i = 1, \dots, N. \quad (1.39)$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(O_{t+1}), \quad t = 1, \dots, T-1, \\ j = 1, \dots, N. \quad (1.40)$$

This is an $O(N^2T)$ computation. Note that it also gives us an efficient way to calculate the value of the objective function, since

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i). \quad (1.41)$$

As the second part of the forward-backward procedure, we consider the backward variable $\beta_t(i)$ defined as

$$\beta_t(i) = P(O_{t+1} \cdots O_T | Z_{it} = 1, \lambda). \quad (1.42)$$

This is the probability of observing the partial sequence $O_{t+1} \cdots O_T$, given that the system is in state i at time t and the model λ . Once again we can solve for $\beta_t(i)$ inductively:

1. Initialization:

$$\beta_T(i) = 1, \quad i = 1, \dots, N. \quad (1.43)$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad t = T-1, \dots, 1, \\ i = 1, \dots, N. \quad (1.44)$$

This is also an $O(N^2T)$ computation.

Now we can calculate the probabilities τ using the forward and backwards variables.

$$\begin{aligned}
\tau_{it} &= P(Z_{it} = 1|O, \lambda) \\
&= \frac{P(Z_{it} = 1|O, \lambda)P(O|\lambda)}{P(O|\lambda)} \\
&= \frac{P(Z_{it} = 1|\lambda)}{P(O|\lambda)} \\
&= \frac{P(O_1 \cdots O_t, Z_{it} = 1|\lambda)P(O_{t+1} \cdots O_T|Z_{it} = 1, \lambda)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^N \alpha_t(i)\beta_t(i)} \tag{1.45}
\end{aligned}$$

is the probability of being in state i at time t , given the observation sequence and the model. Note that we can use τ_{it} to solve for the individually most likely state q_t at time t , as

$$q_t = \operatorname{argmax}_{1 \leq i \leq N}(\tau_{it}), \quad t = 1, \dots, T. \tag{1.46}$$

We can also calculate τ_{ijt} , the probability of being in state i in time t and state j at time $t + 1$, given the model and the observation sequence. Using our definitions of the forward-backward variables, we can write

$$\begin{aligned}
\tau_{ijt} &= P(Z_{it} = 1, Z_{j,t+1} = 1|O, \lambda) \\
&= \frac{P(Z_{it} = 1, Z_{j,t+1} = 1, O|\lambda)}{P(O|\lambda)} \\
&= \frac{P(O_1 \cdots O_t, Z_{it} = 1|\lambda)P(O_{t+1} \cdots O_T, Z_{j,t+1} = 1|Z_{it} = 1, \lambda)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)P(O_{t+1}, Z_{j,t+1} = 1|Z_{it} = 1, \lambda)P(O_{t+2} \cdots O_T|Z_{j,t+1} = 1, \lambda)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)P(Z_{j,t+1} = 1|Z_{it}, \lambda)P(O_{t+1}|Z_{it} = 1, Z_{j,t+1} = 1, \lambda)\beta_{t+1}(j)}{P(O|\lambda)} \\
&= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}. \tag{1.47}
\end{aligned}$$

1.6 Finite mixture models and HMMs

A finite mixture model [15] λ_{fmm} with R components is composed of the mixture parameters $w = (w_1, \dots, w_R)$ and the observation probability density functions associated with each mixture component, $b_r(m)$ for discrete output symbols,

or $b_r(y, \theta_{r1}, \dots, \theta_{rM})$ for continuous outputs. In general, we wish to solve the following problem:

$$\begin{aligned}
\text{Maximize : } & \prod_{t=1}^T P(O_t | \lambda_{fmm}) \\
\text{Subject to : } & \sum_{r=1}^R w_r = 1 \\
& w_r \geq 0, \quad r = 1, \dots, R \\
& \sum_{m=1}^M b_r(m) = 1, \quad r = 1, \dots, R \\
& b_r(m) \geq 0, \quad r = 1, \dots, R, \quad m = 1, \dots, M.
\end{aligned} \tag{1.48}$$

We can express the objective function in terms of the model parameters as follows:

$$\prod_{t=1}^T P(O_t | \lambda_{fmm}) = \prod_{t=1}^T \sum_{r=1}^R w_r b_r(O_t). \tag{1.49}$$

In the case of continuous outputs, the last constraint is replaced by

$$\begin{aligned}
\int_Y b_r(y) dy &= 1, \quad r = 1, \dots, R \\
b_r(y) &\geq 0, \quad r = 1, \dots, R, \quad y \in Y.
\end{aligned} \tag{1.50}$$

Once again we use the EM method to solve this optimization problem [12]. We can represent the complete data log likelihood for the finite mixture model as

$$\sum_{r=1}^R \sum_{t=1}^T z_{rt} \log w_r b_r(O_t) \tag{1.51}$$

where $z = (z_1, \dots, z_T)$ is a set of component indicator vectors such that $z_{rt} = 1$ if the observation is drawn from the r th mixture component, $z_{rt} = 0$ otherwise. From this we can calculate

$$Q(\lambda_{fmm}, \lambda_{fmm}^{(k)}) = \sum_{r=1}^R \sum_{t=1}^T \tau_{rt}^{(k)} \log w_r b_r(O_t) \tag{1.52}$$

where

$$\tau_{rt} = P(Z_{rt} = 1 | O_t, \lambda_{fmm}) \quad t = 1, \dots, T \tag{1.53}$$

and Z is a probabilistic component indicator variable analogous to z . We can calculate $\tau_{rt}^{(k)}$ via Bayes' Rule:

$$\tau_{rt}^{(k)} = \frac{w_r^{(k)} b_r^{(k)}(O_t)}{\sum_{r=1}^R w_r^{(k)} b_r^{(k)}(O_t)}. \tag{1.54}$$

We choose updates of w_r and b_r that maximize Q . We can find the update for w_r using our solution to (1.19):

$$w_r = \frac{\sum_{t=1}^T \tau_{rt}^{(k)}}{\sum_{t=1}^T \sum_{r=1}^R \tau_{rt}^{(k)}} = \frac{1}{T} \sum_{t=1}^T \tau_{rt}^{(k)}. \quad (1.55)$$

To find the update rule for b_r we find the maximum directly via the derivative, solving

$$\frac{\partial L_F}{\partial \theta_{rm}} = \sum_{t=1}^T \tau_{rt}^{(k)} \frac{\partial}{\partial \theta_{rm}} \log b_r(O_t, \theta_{rm}) = 0, \quad (1.56)$$

which has no general analytical solution. As in the HMM case, for certain forms of the output distribution an analytic solution is available.

The hidden Markov model can be seen as a special case of finite mixture model, one in which there is a single observation O and N^T mixture components, each corresponding to a different state sequence Q . In this view we have

$$w_Q = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}, \quad (1.57)$$

$$b_Q(O) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \quad (1.58)$$

We can also construct a hidden Markov model whose state outputs are themselves finite mixture models. For example, if each finite state of the model has R mixture components, then the model is $\lambda = (\pi, A, w, B)$ where $w = (w_{11}, \dots, w_{NR})$, $B = (b_{11}, \dots, b_{NR})$ and π and A retain their original meanings. Let $W = (w_{1r_1}, \dots, w_{Nr_N})$ be some choice of mixture components for each model state. Then for this model we have

$$P(O|\lambda) = \sum_{\text{all } W, Q} \pi_{q_1} w_{q_1 r_{q_1}} b_{q_1 r_{q_1}}(O_1) a_{q_1 q_2} w_{q_2 r_{q_2}} b_{q_2 r_{q_2}}(O_2) \cdots \\ \cdots a_{q_{T-1} q_T} w_{q_T r_{q_T}} b_{q_T r_{q_T}}(O_T). \quad (1.59)$$

Calculation of the forward and backward parameters proceeds as follows:

1. Initialization:

$$\alpha_t(i) = \pi_i \sum_{r=1}^R w_{ir} b_{ir}(O_1), \quad i = 1, \dots, N. \quad (1.60)$$

$$\beta_T(i) = 1, \quad i = 1, \dots, N. \quad (1.61)$$

2. Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] \sum_{r=1}^R w_{jr} b_{jr}(O_{t+1}), \quad t = 1, \dots, T-1, \\ j = 1, \dots, N. \quad (1.62)$$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} \left(\sum_{r=1}^R w_{jr} b_{jr}(O_{t+1}) \right) \beta_{t+1}(j), \quad t = T-1, \dots, 1, \\ i = 1, \dots, N. \quad (1.63)$$

Derivation of this procedure follows that of the forward-backward procedure for the standard hidden Markov model. Once the forward and backward variables have been calculated, we can derive τ_{it} and τ_{ijt} according to (1.45) and (1.47), with the difference that

$$\tau_{ijt} = \frac{\alpha_t(i)a_{ij} \left(\sum_{r=1}^R w_{jr}b_{jr}(O_{t+1}) \right) \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}. \quad (1.64)$$

This allows us to re-estimate at π and A at each iteration using (1.23) and (1.24). Now we have

$$\begin{aligned} \tau_{irt} &= P(Z_{irt} = 1|O, \lambda) \\ &= P(Z_{irt} = 1|Z_{it} = 1, O, \lambda)P(Z_{it}|O, \lambda) \\ &= \frac{w_{ir}b_{ir}(O_t)}{\sum_{r=1}^R w_{ir}b_{ir}(O_t)}\tau_{it} \end{aligned} \quad (1.65)$$

We can re-estimate the mixture weights according to:

$$w_{ir} = \frac{\sum_{t=1}^T \tau_{irt}^{(k)}}{\sum_{r=1}^R \sum_{t=1}^T \tau_{irt}^{(k)}} = \frac{\sum_{t=1}^T \tau_{irt}^{(k)}}{\sum_{t=1}^T \tau_{it}^{(k)}}. \quad (1.66)$$

Again there is no general form for the output distributions, but in the special case of Gaussian outputs for each mixture model component we have,

$$\mu_{ir} = \frac{\sum_{t=1}^T \tau_{irt}^{(k)} O_t}{\sum_{t=1}^T \tau_{irt}^{(k)}}, \quad (1.67)$$

$$\Sigma_{ir} = \frac{\sum_{t=1}^T \tau_{irt}^{(k)} (O_t - \mu_i^{(k+1)})(O_t - \mu_i^{(k+1)})^T}{\sum_{t=1}^T \tau_{irt}^{(k)}}. \quad (1.68)$$

It is worth noting that this model reduces to a simple finite mixture model in the case that the hidden Markov model has but one state. Therefore, hidden Markov models and finite mixture models can each be seen as special cases of the other.