Author: Peter Sarvari (based on Notes by Dr Stefanos Zafeiriou)

Derivation of Maximum Likelihood for Markov chain with 5 states

We have N observations of length T. The joint distribution of the observations is: (For simplicity we do not write the dependency on the model variables, $\theta = \{\boldsymbol{\pi}, \boldsymbol{A}\}$. The vector $\boldsymbol{\pi}$ has dimension 5*1 and contains the probabilities of starting the sequence in a given state. **A** is the row stochastic transition matrix of dimension 5*5. The elements of matrix **A** is denoted by 'a' (e.g. $a_{j,k}$ is the element in the jth row and kth column). x^i is the union of observations in the ith sequence. x_t^i is a scalar and refers to the value that is recorded as the tth state in the ith sequence. $x_{t,j}^i$ is one, if x_t^i takes the jth possible value and zero otherwise. X is the union of x^i.)

$$p(X) = \prod_{i=1}^{N} p(x^i) = \prod_{i=1}^{N} p(x_1^i) * \prod_{t=2}^{T} p(x_t^i|\, x_{t-1}^i) \quad (1)$$

The probability of starting with a state

$$p(x_1^i) = \prod_{j=1}^{5} \pi_j^{x_{1j}^i} \quad (2)$$

Note that from the diagram we actually know $x_{1j}^i = 0\ for\ j = \{2,3,4,5\}\ and\ x_{1j}^i = 0\ for\ j = 1$, however since the exercise asked for a Maximum Likelihood algorithm, we will disregard the information provided by the diagram.

$$p(x_t^i|x_{t-1}^i) = \prod_{k=1}^{5}\prod_{j=1}^{5} a_{j,k}^{x_{t-1,j}^i * x_{t,k}^i} \quad (3)$$

Putting (2) and (3) into (1)

$$p(X) = \prod_{i=1}^{N}\prod_{j=1}^{5} \pi_j^{x_{1j}^i} * \prod_{t=2}^{T}\prod_{k=1}^{5}\prod_{j=1}^{5} a_{j,k}^{x_{t-1,j}^i * x_{t,k}^i} \quad (4)$$

Taking the log of (4)

$$\ln\big(P(X)\big) = \sum_{i=1}^{N}\sum_{j=1}^{5} x_{1j}^i * \ln(\pi_j) + \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{k=1}^{5}\sum_{j=1}^{5} x_{t-1,j}^i * x_{t,k}^i * \ln(a_{j,k})$$

Writing the Lagrangian for π_j

$$L(\pi_j) = \sum_{i=1}^{N}\sum_{j=1}^{5} x_{1j}^i * \ln(\pi_j) - \lambda * (\sum_{j} \pi_j - 1)$$

Optimality condition

$$\frac{\delta L(\pi_j)}{\delta \pi_j} = \sum_{i=1}^{N} x_{1j}^i * \frac{1}{\pi_j} - \lambda = 0 \ \rightarrow$$

$$\sum_{j} \lambda * \pi_j = \lambda = \sum_{i=1}^{N}\sum_{j} x_{1j}^i = N \rightarrow$$

$$\pi_j = \frac{\sum_{i=1}^{N} x_{1j}^i}{N} \quad (5)$$

Writing the Lagrangian for $a_{j,k}$

$$L(a_{j,k}) = \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{k=1}^{5}\sum_{j=1}^{5} x_{t-1,j}^i * x_{t,k}^i * \ln(a_{j,k}) - \lambda * (\sum_k a_{j,k} - 1)$$

Optimality condition:

$$\frac{\delta L(a_{j,k})}{\delta a_{j,k}} = \sum_{i=1}^{N}\sum_{t=2}^{T} x_{t-1,j}^i * x_{t,k}^i * \frac{1}{a_{j,k}} - \lambda = 0 \rightarrow$$

$$\sum_k \lambda * a_{j,k} = \lambda = \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_k x_{t-1,j}^i * x_{t,k}^i = \sum_{i=1}^{N}\sum_{t=2}^{T} x_{t-1,j}^i * \sum_k x_{t,k}^i = \sum_{i=1}^{N}\sum_{t=2}^{T} x_{t-1,j}^i \rightarrow$$

$$a_{j,k} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} x_{t-1,j}^i * x_{t,k}^i}{\sum_{i=1}^{N}\sum_{t=2}^{T} x_{t-1,j}^i} \quad (6)$$

Equation (5) basically says that the best estimate of the starting probability of a given state is counting the number of times a sequence started in that state and divide this by the number of sequences.

Equation (6) states that the best estimate of the transition probability from j to k is counting the number of times (across all sequences) when j is followed by k divided by the number of times j occurred.

<u>Derivation of Expectation-Maximization for Hidden Markov Models with discrete observations taking 5 values</u>

We have N observations of length T. The joint distribution of the observations and the latent variables: (For simplicity we do not write the dependency on the model variables, $\theta = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$. The vector $\boldsymbol{\pi}$ has dimension K*1 and contains the probabilities of starting the sequence in a given state. **A** is the row stochastic transition matrix of dimension K*K and matrix **B** (K*5) gives the emission probabilities, the rows corresponding to one latent variable and different columns corresponding to different possible observations. The elements of matrix **A** is denoted by 'a' and the elements of matrix **B** is denoted by 'b' (e.g. $b_{k,j}$ is the element in the kth row and jth column). x^i is the union of observations in the ith sequence. x_t^i is a scalar and refers to the value that is recorded as the tth state in the ith sequence. $x_{t,j}^i$ is one, if x_t^i takes the jth possible value and zero otherwise. X is the union of x^i. The same notation applies to the latent variable, z.)

$$p(X,Z) = p(X|Z) * p(Z) = \prod_{i=1}^{N} p(x^i|z^i) * p(z^i) \quad (1)$$

The emission probabilities for a sequence:

$$p(x^i|z^i) = \prod_{t=1}^{T}\prod_{j=1}^{5}\prod_{k}^{K} b_{k,j}^{x_{t,j}^i * z_{t,k}^i} \quad (2)$$

Note that j goes up to 5, because the observations can take 5 possible values. K is the number of latent states (unknown).

The probabilities of a latent sequence:

$$p(z^i) = p(z_1^i) * \prod_{t=2}^{T} p(z_t^i | z_{t-1}^i) \quad (3)$$

The probability of a given first latent state in a sequence:

$$p(z_1^i) = \prod_{k=1}^{K} \pi_k^{z_{1,k}^i} \quad (4)$$

The probability of transitioning from one latent state to another in a sequence:

$$p(z_t^i | z_{t-1}^i) = \prod_{k=1}^{K} \prod_{j=1}^{K} a_{j,k}^{z_{t-1,j}^i * z_{t,k}^i} \quad (5)$$

Putting (5) and (4) back into (3):

$$p(z^i) = \prod_{k=1}^{K} \pi_k^{z_{1,k}^i} * \prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{j=1}^{K} a_{j,k}^{z_{t-1,j}^i * z_{t,k}^i} \quad (6)$$

Putting (6) and (2) back into (1):

$$p(X, Z) = \prod_{i=1}^{N} \prod_{t=1}^{T} \prod_{j=1}^{5} \prod_{k}^{K} b_{k,j}^{x_{t,j}^i * z_{t,k}^i} * \prod_{k=1}^{K} \pi_k^{z_{1,k}^i} * \prod_{t=2}^{T} \prod_{k=1}^{K} \prod_{j=1}^{K} a_{j,k}^{z_{t-1,j}^i * z_{t,k}^i} \quad (7)$$

Taking the log of (7):

$$\ln(p(X, Z)) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{5} \sum_{k=1}^{K} x_{t,j}^i * z_{t,k}^i * \ln(b_{k,j}) + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{1,k}^i * \ln(\pi_k) + \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} z_{t-1,j}^i * z_{t,k}^i * \ln(a_{j,k})$$

Since we do not know the value of the latent variables, we can only use their expectation in the optimization procedure. The expectation is taken over the posterior ($p(Z|X)$), since we calculate the expectations of the latent variables using the data. The way to calculate the expectations will be detailed later.

$$E(\ln(p(X, Z))) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{5} \sum_{k=1}^{K} x_{t,j}^i * E(z_{t,k}^i) * \ln(b_{k,j}) + \sum_{i=1}^{N} \sum_{k=1}^{K} E(z_{1,k}^i) * \ln(\pi_k) + \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} E(z_{t-1,j}^i * z_{t,k}^i) * \ln(a_{j,k})$$

Maximization part: Assume we have $E(z_{t,k}^i) \; and \; E(z_{t-1,j}^i * z_{t,k}^i)$

Lagrangian for $b_{k,j}$

$$L\left(b_{k,j}\right) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j=1}^{5}\sum_{k=1}^{K} x_{t,j}^{i} * E(z_{t,k}^{i}) * \ln(\, b_{k,j}) - \lambda * (\sum_{j} b_{k,j} - 1)$$

Optimality condition:

$$\frac{\delta L\left(b_{k,j}\right)}{\delta b_{k,j}} = \sum_{i=1}^{N}\sum_{t=1}^{T} x_{t,j}^{i} * E(z_{t,k}^{i}) * \frac{1}{b_{k,j}} - \lambda = 0 \rightarrow$$

$$\sum_{j} b_{k,j} * \lambda = \lambda = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{j} x_{t,j}^{i} * E(z_{t,k}^{i}) = \sum_{i=1}^{N}\sum_{t=1}^{T} E(z_{t,k}^{i}) \rightarrow$$

$$b_{k,j} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{t,j}^{i} * E(z_{t,k}^{i})}{\sum_{i=1}^{N}\sum_{t=1}^{T} E(z_{t,k}^{i})} \quad (8)$$

Lagrangian for $a_{j,k}$

$$L\left(a_{j,k}\right) = \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{j=1}^{K} E(z_{t-1,j}^{i} * z_{t,k}^{i}) * \ln(a_{j,k}) - \lambda * (\sum_{k} a_{j,k} - 1)$$

Optimality condition:

$$\frac{\delta L\left(a_{j,k}\right)}{\delta a_{j,k}} = \sum_{i=1}^{N}\sum_{t=2}^{T} E(z_{t-1,j}^{i} * z_{t,k}^{i}) * \frac{1}{a_{j,k}} - \lambda = 0 \rightarrow$$

$$\sum_{k} \lambda * a_{j,k} = \lambda = \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{k} E(z_{t-1,j}^{i} * z_{t,k}^{i}) \rightarrow$$

$$a_{j,k} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} E(z_{t-1,j}^{i} * z_{t,k}^{i})}{\sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{k} E(z_{t-1,j}^{i} * z_{t,k}^{i})} \quad (9)$$

Lagrangian for π_j

$$L\left(\pi_j\right) = \sum_{i=1}^{N}\sum_{j=1}^{5} E(z_{1j}^{i}) * \ln(\pi_j) - \lambda * (\sum_{j} \pi_j - 1)$$

Optimality condition:

$$\frac{\delta L\left(\pi_j\right)}{\delta \pi_j} = \sum_{i=1}^{N} E(z_{1j}^{i}) * \frac{1}{\pi_j} - \lambda = 0 \;\rightarrow$$

$$\sum_{j} \lambda * \pi_j = \lambda = \sum_{i=1}^{N}\sum_{j} E(z_{1j}^{i}) \rightarrow$$

$$\pi_j = \frac{\sum_{i=1}^{N} E(z_{1j}^{i})}{\sum_{i=1}^{N}\sum_{j} E(z_{1j}^{i})} \quad (10)$$

If $E(z_{1j}^{i})$ is normalized, then $\sum_{i=1}^{N}\sum_{j} E(z_{1j}^{i}) = N$ and we can write:

$$\pi_j = \frac{\sum_{i=1}^{N} E(z_{1j}^i)}{N} \quad (11)$$

Equation (8), (9) and (10) give the update rule for the parameters in our model. We assumed that we have $E(z_{t,k}^i)\ and\ E(z_{t-1,j}^i * z_{t,k}^i)$. Let's see how to calculate these! For simplicity, we will drop the sequence index i in the following derivations. Note that we introduce a change of notation here: from now on z_t is a vector of K dimension, and the index of the activated element is the state (latent variable) at time t.

$$E(z_{t,k}) = \sum_k z_{t,k} * p\left(z_{t,k} \middle| x_1 \dots x_T\right) = p(z_{t,k} = 1 | x_1 \dots x_T)$$

As usual, we denote $p\left(z_{t,k} = 1 \middle| x_1 \dots x_T\right)$ by $p\left(z_{t,k} \middle| x_1 \dots x_T\right)$.

$$p\left(z_{t,k} \middle| x_1 \dots x_T\right) = \frac{p\left(x_1 \dots x_t, z_{t,k}\right) * p(x_{t+1} \dots x_T | z_{t,k})}{p(x_1 \dots x_T)} \stackrel{\text{def}}{=} \frac{\alpha\left(z_{t,k}\right) * \beta\left(z_{t,k}\right)}{p(x_1 \dots x_T)} = \hat{\alpha}\left(z_{t,k}\right) * \hat{\beta}\left(z_{t,k}\right)$$

$$\stackrel{\text{def}}{=} p\left(z_{t,k} \middle| x_1 \dots x_t\right) * \frac{p\left(x_{t+1} \dots x_T \middle| z_{t,k}\right)}{p(x_{t+1} \dots x_T | x_1 \dots x_t)} \quad (12.1)$$

Where $\boldsymbol{\alpha}(\boldsymbol{z_t})$ and $\boldsymbol{\beta}(\boldsymbol{z_t})$ are column vectors of dimension K and $\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_t})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{z_t})$ are their normalized versions. This means that the elements in $\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_t})$ sum to one and using the definition used by Kevin Murphy, also the elements of $\widehat{\boldsymbol{\beta}}(\boldsymbol{z_t})$ sum to one. However, then the expectation probabilities will not be normalized. To make the expectation probabilities normalized, we define $\widehat{\boldsymbol{\beta}}(\boldsymbol{z_t})$ otherwise (as introduced in the lectures) and then the elements do not sum to one. This is fine, since $\widehat{\boldsymbol{\beta}}(\boldsymbol{z_t})$ is not a probability distribution over the states (K. Murphy, page 610). Note that both approaches lead to valid EM algorithms, since we do not require the expectations to be normalized, we only care about how the elements compare to each other (how big they are *relative* to each other). Here, we detail the method discussed in the lectures, but in the first part of the coursework, Kevin Murphy's implementation is used (mainly because the normalization constant, c_t determined in 'ForwardFiltering.m' was not passed to the 'BackwardFiltering.m' function skeleton given to us).

Define c_t as

$$c_t \stackrel{\text{def}}{=} p(x_t | x_1 \dots x_{t-1}) \quad (12.2)$$

Note that then,

$$\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_t}) * \prod_{m=1}^{t} c_m = \boldsymbol{\alpha}(\boldsymbol{z_t}) \quad (12.3)$$

$$\prod_{m=1}^{T} c_m = p(x_1 \dots x_T) \quad (12.4)$$

Also,

$$\boldsymbol{\alpha}(\boldsymbol{z_t}) = \sum_{\boldsymbol{z_{t-1}}} \boldsymbol{\alpha}(\boldsymbol{z_{t-1}}) * p(\boldsymbol{z_t} | \boldsymbol{z_{t-1}}) * p(x_t | \boldsymbol{z_t}) = \boldsymbol{B}_{:,x_t} .* \left(\boldsymbol{A}' * \boldsymbol{\alpha}(\boldsymbol{z_{t-1}})\right) \quad (13.1)$$

$$c_t * \widehat{\boldsymbol{\alpha}}(\boldsymbol{z_t}) = \sum_{\boldsymbol{z_{t-1}}} \widehat{\boldsymbol{\alpha}}(\boldsymbol{z_{t-1}}) * p(\boldsymbol{z_t} | \boldsymbol{z_{t-1}}) * p(x_t | \boldsymbol{z_t}) = \boldsymbol{B}_{:,x_t} .* \left(\boldsymbol{A}' * \widehat{\boldsymbol{\alpha}}(\boldsymbol{z_{t-1}})\right) \quad (13.2)$$

And

$$\boldsymbol{\beta}(\boldsymbol{z_t}) = \sum_{z_{t+1}} \boldsymbol{\beta}(\boldsymbol{z_{t+1}}) * p(\boldsymbol{z_{t+1}}|\boldsymbol{z_t}) * p(x_{t+1}|\boldsymbol{z_{t+1}}) = \boldsymbol{A} * \left(\boldsymbol{\beta}(\boldsymbol{z_{t+1}}).* \boldsymbol{B}_{:,x_{t+1}}\right) \quad (14.1)$$

$$c_{t+1} * \widehat{\boldsymbol{\beta}}(\boldsymbol{z_t}) = \sum_{z_{t+1}} \widehat{\boldsymbol{\beta}}(\boldsymbol{z_{t+1}}) * p(\boldsymbol{z_{t+1}}|\boldsymbol{z_t}) * p(x_{t+1}|\boldsymbol{z_{t+1}}) = \boldsymbol{A} * \left(\widehat{\boldsymbol{\beta}}(\boldsymbol{z_{t+1}}).* \boldsymbol{B}_{:,x_{t+1}}\right) \quad (14.2)$$

Note that we denoted the Hadamard product as '.*'. Equations (13) and (14) have a recursive fashion. To be able to use them, we need to define $\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_1})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{z_T})$, respectively.

$$\boldsymbol{\alpha}(\boldsymbol{z_1}) = p(x_1, \boldsymbol{z_1})$$

$$\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_1}) = p(\boldsymbol{z_1}|x_1) = \frac{p(x_1|\boldsymbol{z_1}) * p(\boldsymbol{z_1})}{p(x_1)}$$

Note that $p(x_1) = \sum_{\boldsymbol{z_1}} p(x_1, \boldsymbol{z_1}) = \sum_{\boldsymbol{z_1}} \boldsymbol{\alpha}(\boldsymbol{z_1})$, so $\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_1}) = \frac{\boldsymbol{\alpha}(\boldsymbol{z_1})}{\sum_{\boldsymbol{z_1}} \boldsymbol{\alpha}(\boldsymbol{z_1})}$. So the normalization constant, c$_1$ is simply $\sum_{\boldsymbol{z_1}} \boldsymbol{\alpha}(\boldsymbol{z_1})$. Then (13.2), we have

$$c_2 * \widehat{\boldsymbol{\alpha}}(\boldsymbol{z_2}) = \sum_{z_{t-1}} \widehat{\boldsymbol{\alpha}}(\boldsymbol{z_1}) * p(\boldsymbol{z_2}|\boldsymbol{z_1}) * p(x_2|\boldsymbol{z_2}) = \boldsymbol{B}_{:,x_2}.* \left(\boldsymbol{A}' * \widehat{\boldsymbol{\alpha}}(\boldsymbol{z_1})\right) \quad (15)$$

Since everything on the right hand side of (15) is normalized, and $\widehat{\boldsymbol{\alpha}}(\boldsymbol{z_2})$ is normalized by definition, c_2 must be the normalization constant. This means that $c_2 = \sum_{\boldsymbol{z_2}} \boldsymbol{\alpha}(\boldsymbol{z_2})$. Continuing the reasoning in the same fashion, we realize that

$$c_t = \sum_{\boldsymbol{z_t}} \boldsymbol{\alpha}(\boldsymbol{z_t}) \quad (16)$$

To calculate $\widehat{\beta}(z_T)$, consider the following:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{z_T}) = p(x_{T+1} \ldots x_T|\boldsymbol{z_T}) = p(\square|\boldsymbol{z_T}) = \boldsymbol{1}$$

Where 1 is a column vector of ones of dimension K. This because the probability of no event is one. (K. Murphy, Page 611)

A more elegant derivation is provided by Bishop (page 622). He considers that since (see equation 12.1) $p(z_{T,k}|x_1 \ldots x_T) = \hat{\alpha}(z_{T,k}) * \hat{\beta}(z_{T,k}) = p(z_{T,k}|x_1 \ldots x_T) * \hat{\beta}(z_{T,k})$, $\hat{\beta}(z_{T,k})$ must be 1 for all k, otherwise we have inconsistency.

We also need to find $E(z_{t-1,j} * z_{t,k})$!

$$\begin{aligned} E(z_{t-1,j} * z_{t,k}) &= \sum_{j,k} z_{t-1,j} * z_{t,k} * p(z_{t-1}, z_t|x_1 \ldots x_T) = p(z_{t-1,j} = 1, z_{t,k} = 1|x_1 \ldots x_T) \\ &= p(z_{t-1,j}, z_{t,k}|x_1 \ldots x_T) \end{aligned}$$

$$\begin{aligned} p(z_{t-1,j}, z_{t,k}|x_1 \ldots x_T) &= \frac{p(z_{t-1,j}, x_1 \ldots x_{t-1}) * p(z_{t,k}|z_{t-1,j}) * p(x_t|z_{t,k}) * p(x_{t+1} \ldots x_T|z_{t,k})}{p(x_1 \ldots x_T)} \\ &= \frac{\alpha(z_{t-1,j}) * p(z_{t,k}|z_{t-1,j}) * p(x_t|z_{t,k}) * \beta(z_{t,k})}{p(x_1 \ldots x_T)} \\ &= c_t^{-1} * \hat{\alpha}(z_{t-1,j}) * p(z_{t,k}|z_{t-1,j}) * p(x_t|z_{t,k}) * \hat{\beta}(z_{t,k}) \quad (17) \end{aligned}$$

Finally, we are done with the expectation step. As a summary, the expectations can be calculated as follows:

$$E\left(z_{t,k}\right) = \hat{\alpha}\left(z_{t,k}\right) * \hat{\beta}\left(z_{t,k}\right) \quad (12.1)$$

$$E\left(z_{t-1,j} * z_{t,k}\right) = c_t^{-1} * \frac{\hat{\alpha}\left(z_{t-1,j}\right) * p\left(z_{t,k} \middle| z_{t-1,j}\right) * p\left(x_t \middle| z_{t,k}\right) * \hat{\beta}\left(z_{t,k}\right)}{p(x_1 \ldots x_T)} \qquad (17)$$