Expectation Maximization for Two Stage Function

Julia Ericson

July 2021

1 Aim

The aim of the study is to identify transitions between states during working memory training, where the learning process within one state differs from the learning process in another state. This model can be though of as a Hidden Markov Model (HMM) where the performance y during training is observed but the states x are hidden. It is the state of each of the hidden variables x that the HMM estimate.

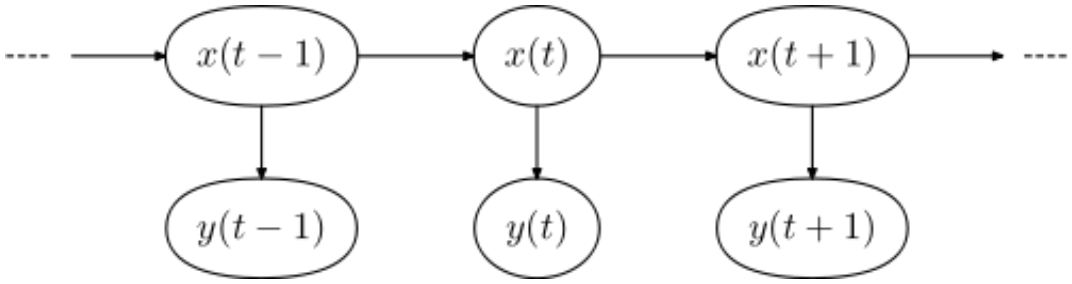


Figure 1: A graphical representation of a Hidden Markov Model

In the example which is described below, there are two states and the performance y is modelled to grow linearly within each state such that $y = a_1 + b_1 t + w$ during state one, and $y = a_2 + b_2 t + w$ during state two. w is a white noise, $w \sim N(0, \sigma)$, where the σ is assumed to be identical in both states, and t is the time. As seen in figure 1, the probability of x_i is only dependent on the state of the previous hidden variable x_{t-1} such that:

$$p(x_i|x_{i-1}, \dots x_1) = p(x_i|x_{i-1}) = Ax_{i-1}. \tag{1}$$

x_{i-1} is a 2 x 1 probability vector where $x_{i-1,1}$ is the probability of being in state one during observation $i-1$ and $x_{i-1,2}$ is the probability of being in state two during observation $i-1$. Matrix A is called a transition matrix. In this example of cognitive training, the model is assumed to be able to transition from state one to state two but not back into state one again. Therefore the transition matrix becomes:

$$A = \begin{bmatrix} 1-\pi & 0 \\ \pi & 1 \end{bmatrix} \tag{2}$$

where π is the probability to transition from state one to state two.

In order to calculate the most likely state for each hidden variable, the parameters a_1, a_2, b_1, b_2 and π, have to be optimized. However, in order to fit the parameters, the hidden variables have to be known. To solve this problem, Expectation Maximization (EM) is used. In the EM algorithm, the parameters are optimized by iterating over the two following steps:

1. Calculating the probability of the hidden variables x given observations y and the current guess of the parameters θ. During the first iteration, θ is usually randomly initialized.
2. Optimizing the parameters θ using the current guess of x such that the likelihood $p(y|\theta)$ is maximized.

2 EM algorithm theory

The aim of the algorithm is to find the parameters which maximize the likelihood of the data, $\max_\theta\ p(y|\theta)$. Since $\max_\theta\ p(y|\theta) = \max_\theta \log p(y|\theta)$, the log likelihood is used instead as the expression is less challenging from a numerical perspective. Furthermore, since the probability of y depends on the hidden states x, these have to marginalize out in the following way:

$$\log p(y|\theta) = \log \int_x p(x, y|\theta) dx \tag{3}$$

However, the above expression is difficult to optimize and therefore, a lower bound to the expression is used formulated instead:

$$\begin{aligned} \log \int_x p(x, y|\theta) dx &= \log \int_x q(x) \frac{p(x, y|\theta)}{q(x)} dx \\ &\geq \int_x q(x) \log \frac{p(x, y|\theta)}{q(x)} dx \\ &= \int_x q(x) \log p(x, y|\theta) dx - \int_x q(x) \log q(x) dx \\ &= \mathcal{L}(q, \theta) \end{aligned} \tag{4}$$

The inequality above can be generalized as the inequality $f(\mathrm{E}[X]) \geq \mathrm{E}[f(X)]$ for a concave function $f(X)$. This is known as Jensen's inequality, and it has an intuitive graphical interpretation for those who wish to look it up. $\mathcal{L}(q, \theta)$ is called the Evidence Lower Bound (ELBO) function, since the true value of the log likelihood will never be smaller than $\mathcal{L}(q, \theta)$ for any given function q. The first step of the EM-algorithm is to maximize the ELBO function with respect to the distribution $q(x)$ such that for iteration k:

$$q_{k+1} = \arg\max_q\ \mathcal{L}(q, \theta_k). \tag{5}$$

The function q which maximizes $\mathcal{L}(q, \theta_k)$ is $p(x|y, \theta_k)$ since

$$\begin{aligned} \int_x p(x|y, \theta_k) \log\ p(xy|\theta_k) dx - \int_x p(x|y, \theta_k) \log\ p(x|y, \theta_k) dx &= \int_x p(x|y, \theta_k) \log\ p(x|y, \theta_k) p(y|\theta_k) dx - \\ &\quad - \int_x p(x|y, \theta_k) \log\ p(x|y, \theta_k) dx \\ &= \int_x p(x|y, \theta) \log\ p(x|y, \theta) dx + \int_x p(x|y, \theta) \log\ p(y|\theta) dx - \\ &\quad - \int_x p(x|y, \theta) \log\ p(x|y, \theta) dx \\ &= \int_x p(x|y, \theta) \log\ p(y|\theta) dx \\ &= \log\ p(y|\theta). \end{aligned} \tag{6}$$

The second step is to maximize $\mathcal{L}(q_{k+1}, \theta)$ with respect to θ

$$\theta_{k+1} = \arg\max_\theta\ \mathcal{L}(q_{k+2}, \theta). \tag{7}$$

In this step, what is really maximized is the expected value $\mathrm{E}_{p(x|y,\theta_k)}[\log\ p(y, x|\theta)]$ since

$$\int_x p(x|y,\theta_k)\log p(x,y|\theta)dx - \int_x p(x|y,\theta_k)\log p(x|y,\theta_k)dx = \mathrm{E}_{p(x|y,\theta_k)}[\log\, p(y,x|\theta)] - \mathrm{E}_{p(x|y,\theta_k)}[\log\, p(y|\theta_k)] \tag{8}$$

where the second term does not depend on θ.

To summarize, the idea of the EM algorithm is to first estimate the hidden states of a model using the current parameter estimation and thereafter use this estimation to calculate new optimal parameter values. Since the expectation step does not change any of the parameter values and the ELBO function equals the log-likelihood at the beginning of each maximization step, it can be guaranteed that the log-likelihood does not decrease during a parameter.

More information on the EM theory can be found in the paper by Roweis & Ghahramani (1999), and an explanation of Jensen's inequality and the ELBO function is found in Bishop's *Pattern Recognition and Machine Learning* (p. 55 - 58).

3 EM algorithm for the Two Stage Linear Model

3.1 Maximizing the likelihood

This section derives the expression for the parameter update of the two state linear model of cognitive training. For an HMM, each iteration of the EM algorithm solves the following optimization problem:

$$\begin{aligned}\theta_{k+1} &= \arg\max_{\theta} \mathrm{E}_{p(x|y,\theta_k)}[\log\, p(y,x|\theta)] \\ &= \arg\max_{\theta} \mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=1}^{N} \log\, p(y_m|x_m,\theta) + \log\, p(x_m|x_{m-1},\ldots,x_1,\theta)\Big].\end{aligned} \tag{9}$$

To begin with, we optimize the first term in the equation above. As defined in section 1, $p(y_m|x_m = i) \sim N(a_i + b_i t_m, \sigma)$ and therefore:

$$\begin{aligned}a_1,b_1,a_2,b_2 &= \arg\max_{a_1,b_1,a_2,b_2} \mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=1}^{N} \log\, p(y_m|x_m,a_1,b_1,a_2,b_2)\Big] \\ &= \arg\max_{a_1,b_1,a_2,b_2} \mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=1}^{N} \log\, \frac{1}{\sqrt{2\pi\sigma}}\exp\Big(-\frac{1}{2\sigma^2}(y_m - a_{x_m} - b_{x_m}t_m)^2\Big)\Big] \\ &= \arg\max_{a_1,b_1,a_2,b_2} \mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=1}^{N} -\frac{1}{2\sigma^2}(y_m - a_{x_m} - b_{x_m}t_m)^2\Big]\end{aligned} \tag{10}$$

Furthermore, if a_1 and b_1 are optimized separately from a_2 and b_2

$$a_1,b_1 = \arg\min_{a_1,b_1} \sum_{m=1}^{N} p(x_m = 1|y)\frac{1}{2\sigma^2}(y_m - a_1 - b_1 t_m)^2 \tag{11}$$

and

$$a_2,b_2 = \arg\min_{a_2,b_2} \sum_{m=1}^{N} p(x_m = 2|y)\frac{1}{2\sigma^2}(y_m - a_2 - b_2 t_m)^2, \tag{12}$$

where the equations have been multiplied by minus one and are minimized instead of maximized. These two equations are objective functions for weighted linear regression where $p(x_m = i|y)$ is the weight. By setting the derivative to zero and solving for a and b we get

$$
\begin{aligned}
a_i &= \frac{\sum_m p(x_m = i|y) y_m - b_i \sum_m p(x_m = i|y) x_m}{\sum_m p(x_m = i|y)} \\
b_i &= \frac{\sum_m p(x_m = i|y)\big(x_m - \frac{\sum_m p(x_m=i|y)x_i}{\sum_m p(x_m=i|y)}\big)\big(y_m - \frac{\sum_m p(x_m=i|y)y_i}{\sum_m p(x_m=i|y)}\big)}{\sum_m p(x_m = i|y)\big(x_m - \frac{\sum_m p(x_m=i|y)x_m}{\sum_m p(x_m=i|y)}\big)^2}.
\end{aligned} \tag{13}
$$

The second part of equation 9 is thereafter used to derive the transition probability π:

$$
\pi = \arg\max_{\pi} \mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=1}^{N} \log\, p(x_m|x_{m-1},\ldots,x_1,\pi)\Big]. \tag{14}
$$

The expected value can be rewritten in the following way:

$$
\begin{aligned}
\mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=1}^{N} \log\, p(x_m|x_{m-1},\ldots,x_1,\pi)\Big] =& \mathrm{E}_{p(x|y,\theta_k)}\Big[\sum_{m=2}^{N} \log\, p(x_m|x_{m-1},\pi) + \log\, p(x_1)\Big] \\
&= \sum_{m=2}^{N} \Big(p(x_m = 1, x_{m-1} = 1|y,\pi_k)\log(1-\pi) + \\
&\quad + p(x_m = 2, x_{m-1} = 2|y,\pi_k)\log(1) + \\
&\quad + p(x_m = 2, x_{m-1} = 1|y,\pi_k)\log(\pi)\Big)
\end{aligned} \tag{15}
$$

In the above equation, $p(x_m = 2, x_{m-1} = 1|y,\pi_k)$ is the probability that the transition happened between input $m-1$ and m, while $p(x_m = 1, x_{m-1} = 1|y,\pi_k)$ and $p(x_m = 2, x_{m-1} = 1|y,\pi_k)$ are the probabilities that the transition happens after m or before $m-1$ respectively. Furthermore, notice that $\log(1) = 0$, and the second part of the equation is therefore zero. To maximize the equation with respect to π, the derivative is set to zero and solved for π. The solution to the equation is:

$$
\pi_{k+1} = \frac{\sum_{m=2}^{N} p(x_m = 2, x_{m-1} = 1|y,\pi_k)}{\sum_{m=2}^{N} p(x_m = 2, x_{m-1} = 1|y,\pi_k) + \sum_{m=2}^{N} p(x_m = 1, x_{m-1} = 1|y,\pi_k)}. \tag{16}
$$

3.2 Defining the conditional probability of the hidden states

Equations 13 and 16 require an expression for the probability $p(x|y,\theta)$. Since $p(x|y,\theta) = \frac{p(x,y|\theta)}{\sum_x p(y,x|\theta)}$, it is enough to find an expression for $p(x,y|\theta)$. Similarly to Tenison & Anderson (2016), $P(j)$ is defined as the combined probability of the observed variables and the hidden variables transforming into state two between observations j and $j+1$. The expression for $P(j)$ is

$$
P(j) = (1-\pi)^{j-1}\pi \prod_{m=1}^{j} p(y_m|x_m = 1) \prod_{m=j+1}^{N} p(y_m|x_m = 2) \tag{17}
$$

for $j \in \{1,\ldots,N-1\}$. Furthermore, the probability that the last observation is still in state one is defined as

$$
P(N) = (1-\pi)^{N-1} \prod_{m=1}^{N} p(y_m|x_m = 1). \tag{18}
$$

Therefore, the probability that x_m is in state two given the observations and the parameters θ is

$$p(x_m = 2|y, \theta) = \frac{\sum_{j=1}^{m-1} P(j)}{\sum_{j=1}^{N} P(j)} \tag{19}$$

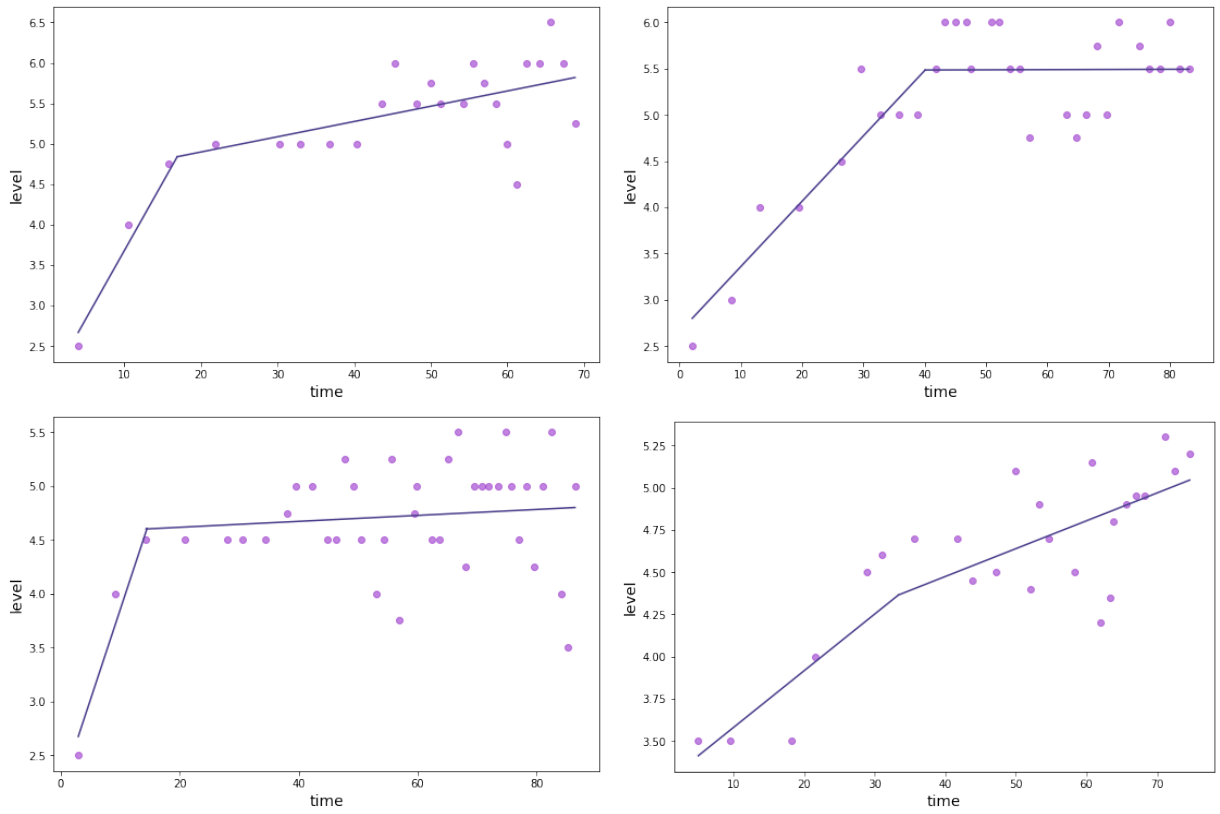
for $j \in \{2, \ldots, N\}$. Since the hidden variables can only be in two states this further means that

$$p(x_m = 1|y, \theta) = 1 - p(x_m = 2|y, \theta). \tag{20}$$

Finally, the sequence will be forced to start in state one which means that $p(x_1 = 2|y, \theta) = 0$ and $p(x_1 = 1|y, \theta) = 1$.

4 Practical Example

In this section the algorithm is applied on data from Working Memory Grid training on Vektor. The median level on each day is used to represent the measured performance y. The time t is the total time spent in minutes on the exercise Working Memory Grid. Below are four examples of training data fitted by the algorithm.



$$\begin{aligned} y &= \alpha + \beta_{strat} t + \beta_{cap} t \\ y &= \alpha + \beta_{strat} t_\tau + \beta_{cap} t \end{aligned} \tag{21}$$

5 References

C. Bishop. 2006. *Pattern Recognition and Machine Learning.*

S. Roweis & Z. Ghahramani. 1999. A Unifying Review of Linear Gaussian Models. *Neural Computation.*

C. Tenison & J. Anderson. 2016. Modelling the Distinct Phases of Skill Acquisition. *Journal of Experimental Psychology: Learning, Memory and Cognition.*