Expectation Maximization Explanation

Chunpai Wang

May 20, 2016

Expectation Maximization (EM) has many applications, which can be categorized into two

- the data has missing values, due to problems with or limitations of the observation process.
- optimizing the likelihood function is analytically intractable but when the likelihood function can be simplified by assuming the existence of the values for additional but missing (or hidden) parameters.

1 Maximum Likelihood and Hidden Variable

The probability density function has the form

$$p(x|\theta)$$

where θ is a set of parameters, i.e. which could be mean and variance in Gaussian distribution.

Given a dataset $X = \{x_1, ..., x_n\}$,

$$p(X|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = L(\theta|X)$$

is called the likelihood of the parameters given the data. Data X is observed and fixed, we would like to maximize likelihood $L(\theta|X)$ to find the optimal parameters θ such that fit the data, which is called MLE

$$\theta^* = \arg\max_{\theta} L(\theta|X)$$

However, if the data X is not completed, where there are some data are not observed and hidden, we can express this missing data with hidden variable Y. Then, we have the log-MLE as below

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \log L(\theta|X, Y) \\ &= \arg\max_{\theta} \log p(X, Y|\theta) \\ &= \arg\max_{\theta} \int \log p(X|y)p(y|\theta)dy\end{aligned}$$

It is very hard to maximize the log-likelihood, and sometimes we are not able to find the analytical expression by taking the derivative of the log-likelihood w.r.t θ and setting to zero, since there is summation or integral inside the logarithm. However, EM will be introduced to solve this by approximating the likelihood of observed data X, which has a very good approximation-bound guarantee.

2 EM Algorithm

Given data X and assume it is generated under some distribution, but is incomplete dataset. We introduce $Z = (X, Y)$ to represent the completed dataset and assume a joint density function

$$\begin{aligned}p(Z|\theta) &= p(X, Y|\theta) \\ &= p(Y|X, \theta)p(X|\theta)\end{aligned}$$

Then, we can write the likelihood of the complete dataset as

$$\log L(\theta|Z) = \log L(\theta|X,Y) = \log p(X,Y|\theta)$$

since here Y is unknown, we can treat this likelihood function as a function of random variable Y, and θ and X are constant, we can view this log likelihood as a function of Y with given X and θ. We denote by

$$h_{X,\theta}(Y) = \log p(X,Y|\theta) = \log L(\theta|Z)$$

Note here, Y is unknown and random, but presumably governed by an underlying distribution.

The EM will approximate the optimal solution by two steps iteratively:

E-step:

$$Q(\theta,\theta^{(i-1)}) = \mathbb{E}_Y\Big[\log p(X,Y|\theta^{(i)})|X,\theta^{(i-1)}\Big] = \mathbb{E}\big[h_{X,\theta^{(i)}}(Y)|X,\theta^{(i-1)}\big]$$

E-step is to calculate the conditional expectation of the completed data log-likelihood w.r.t. the unknown data Y given the observed data X and the current parameter estimation of θ. Thus, we need to initialize a reasonable θ. Since we view Y as a random variable, and according to the conditional expectation formula, for example

$$\mathbb{E}\big[h(Y)|X=x\big] = \int_y h(y) f_{y|x}(y|x)dy$$

we can express the

$$\mathbb{E}_Y\Big[\log p(X,Y|\theta)|X,\theta^{(i-1)}\Big] = \int_y \log p(X,Y|\theta)\cdot p_{y|X}(y|X,\theta^{(i-1)})dy$$

M-step:

$$\theta^{(i)} = \arg\max_\theta Q(\theta,\theta^{(i-1)})$$

M-step is to find the optimal θ such that maximize the expectation we derived in E-step, under the assumption that Y is known from E-step.

3 Derivation of the EM Algorithm

EM is an iterative procedures for approximately maximizing the likelihood $L(X|\theta)$ of the incomplete data X. Assume at iteration $i-1$, and the estimate of θ is $\theta^{(i-1)}$, we want to compute an updated estimated θ such that $L(\theta|X) > L(\theta^{(i-1)}|X)$ to achieve maximum likelihood iteratively. Therefore, we need to consider the gap (difference,metric) that

$$L(\theta|X) - L(\theta^{(i-1)}|X)$$

We are going to find the lower bound of the $L(\theta|X)$. We represent function L as log-likelihood, X as observed data, and Y as hidden variable for missing data.

$$\begin{aligned}
& L(\theta|X) - L(\theta^{(i-1)}|X) && (1)\\
&= \log p(X|\theta) - \log p(X|\theta^{(i-1)}) && (2)\\
&= \log \int_y p(X|y,\theta)p(y|\theta)dy - \log p(X|\theta^{(i-1)}) && (3)\\
&= \log \int_y \frac{p(y|X,\theta^{(i-1)})}{p(y|X,\theta^{(i-1)})} p(X|y,\theta)p(y|\theta)dy - \log p(X|\theta^{(i-1)}) && (4)\\
&= \log \int_y p(y|X,\theta^{(i-1)}) \frac{p(X|y,\theta)p(y|\theta)}{p(y|X,\theta^{(i-1)})}dy - \log p(X|\theta^{(i-1)}) && (5)\\
&\geq \int_y p(y|X,\theta^{(i-1)}) \log \frac{p(X|y,\theta)p(y|\theta)}{p(y|X,\theta^{(i-1)})}dy - \log p(X|\theta^{(i-1)}) && (6)\\
&= \int_y p(y|X,\theta^{(i-1)}) \log \frac{p(X|y,\theta)p(y|\theta)}{p(y|X,\theta^{(i-1)})p(X|\theta^{(i-1)})}dy && (7)
\end{aligned}$$

- Step 6, since

$$\int_y p(y|X,\theta^{(i-1)})dy = 1$$

and logarithm is concave function, we can use the Jensen inequality

$$f(\sum_i \lambda_i x_i) \geq \sum_i \lambda_i f(x_i) \quad \text{if f is concave and} \quad \sum_i \lambda_i = 1.$$

$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i) \quad \text{if f is convex and} \quad \sum_i \lambda_i = 1.$$

- Step 7, since the $\log p(X|\theta^{(i-1)})$ does not involve the variable y, we can put it into the integral of y.

Therefore, we have the lower bound of

$$L(\theta|X) \geq L(\theta^{(i-1)}|X) + \int_y p(y|X,\theta^{(i-1)}) \log \frac{p(X|y,\theta)p(y|\theta)}{p(y|X,\theta^{(i-1)})p(X|\theta^{(i-1)})} dy$$

, and we denote this lower bound as $B(\theta|\theta^{(i-1)})$. Now, we can see that, if we find a good θ such that increase or maximize the $B(\theta|\theta^{(i-1)})$, then we can achieve our goal to maximize the likelihood function of observed data $L(\theta|X)$. Thus, next step is to find the optimal θ that maximize the lower bound of $L(\theta|X)$

$$\begin{aligned}
\theta^{(i)} &= \arg\max_\theta B(\theta|\theta^{(i-1)}) && (8)\\
&= \arg\max_\theta L(\theta^{(i-1)}|X) + \int_y p(y|X,\theta^{(i-1)}) \log \frac{p(X|y,\theta)p(y|\theta)}{p(y|X,\theta^{(i-1)})p(X|\theta^{(i-1)})} dy && (9)\\
&= \arg\max_\theta \int_y p(y|X,\theta^{(i-1)}) \log \frac{p(X|y,\theta)p(y|\theta)}{p(y|X,\theta^{(i-1)})p(X|\theta^{(i-1)})} dy && (10)\\
&= \arg\max_\theta \int_y p(y|X,\theta^{(i-1)}) \big[\log\ p(X|y,\theta)p(y|\theta) - \log p(y|X,\theta^{(i-1)})p(X|\theta^{(i-1)})\big] dy && (11)\\
&= \arg\max_\theta \int_y p(y|X,\theta^{(i-1)}) \log\ p(X|y,\theta)p(y|\theta) dy && (12)\\
&= \arg\max_\theta \int_y \big[\log\ p(X,y|\theta)\big] p(y|X,\theta^{(i-1)}) dy && (13)\\
&= \arg\max_\theta \mathbb{E}_{Y|X,\theta^{(i-1)}} \Big[\log(X,Y|\theta^{(i-1)})\Big] && (14)\\
&= \arg\max_\theta Q(\theta,\theta^{(i-1)}) && (15)
\end{aligned}$$

- From step 9 12, we omit the parts which do not involve the parameter θ.
- Step 15, we find that the lower bound $B(\theta,\theta^{(i-1)})$ is actually $Q(\theta,\theta^{(i-1)})$ we defined in EM algorithm before.

In the end, we find the the lower bound of $L(\theta|X)$ is actually the expected value of the completed-data log-likelihood with respect to the unknown data Y given the observed data X and the current parameter estimation $\theta^{(i-1)}$. Therefore, we realize that EM algorithm is to maximize the lower bound iteratively with an initial parameter $\theta^{(0)}$, and it is not guaranteed to find the global optimum.

4 Convergence of EM

Theorem 4.1. *Assume $p(X|\theta)$ is the likelihood of observed data X, and $\theta^{(i-1)}$,where $i = 0, ..., k$, is the parameter estimate at i^{th} iteration, and corresponding likelihood is $p(X|\theta^{(i)})$, then*

$$p(X|\theta^{(i)}) \geq p(X|\theta^{(i-1)})$$

Proof. First, we know that

$$p(X|\theta) = \frac{p(X,Y|\theta)}{p(Y|X,\theta)}$$

, thus we have

$$\log p(X|\theta) = \log \frac{p(X,Y|\theta)}{p(Y|X,\theta)} = \log p(X,Y|\theta) - \log p(Y|X,\theta)$$

Recall that

$$Q(\theta,\theta^{(i-1)}) = \mathbb{E}_Y\Big[\log p(X,Y|\theta^{(i)})|X,\theta^{(i-1)}\Big] = \int_y \big[\log\ p(X,y|\theta)\big]p(y|X,\theta^{(i-1)})dy$$

and we define

$$W(\theta,\theta^{(i-1)}) = \int_y \log p(y|X,\theta)p(y|X,\theta^{(i-1)})dy$$

then we have

$$\begin{aligned}
&Q(\theta,\theta^{(i-1)} - W(\theta,\theta^{(i-1)}) \\
&= \int_y \log\ p(X,y|\theta)p(y|X,\theta^{(i-1)})dy - \int_y \log p(y|X,\theta)p(y|X,\theta^{(i-1)})dy \\
&= \int_y \log \frac{p(X,Y|\theta)}{p(Y|X,\theta)}dy \\
&= \int_y \log p(X|\theta)dy \\
&= \log p(X|\theta)
\end{aligned}$$

Therefore, it follows that

$$\begin{aligned}
&\log p(X|\theta^{(i)}) - \log p(X|\theta^{(i-1)}) \\
&= \Big(Q(\theta^{(i)},\theta^{(i-1)} - W(\theta^{(i)},\theta^{(i-1)})\Big) - \Big(Q(\theta^{(i-1)},\theta^{(i-1)} - W(\theta^{(i-1)},\theta^{(i-1)})\Big) \\
&= \underbrace{\Big(Q(\theta^{(i)},\theta^{(i-1)} - Q(\theta^{(i-1)},\theta^{(i-1)}\Big)}_{1} - \underbrace{\Big(W(\theta^{(i)},\theta^{(i-1)}) - W(\theta^{(i-1)},\theta^{(i-1)})\Big)}_{2} \\
&\geq 0
\end{aligned}$$

1. Since in M-step of the EM algorithm, $\theta^{(i)}$ is chosen as to maximize $Q(\theta,\theta^{(i-1)})$. Thus,

$$Q(\theta^{(i)},\theta^{(i-1)}) \geq Q(\theta^{(i-1)},\theta^{(i-1)})$$

2.

$$\begin{aligned}
&W(\theta^{(i)},\theta^{(i-1)}) - W(\theta^{(i-1)},\theta^{(i-1)}) \\
&= \int_y \log p(y|X,\theta^{(i)})p(y|X,\theta^{(i-1)})dy - \int_y \log p(y|X,\theta^{(i-1)})p(y|X,\theta^{(i-1)})dy \\
&= \int_y \log \frac{p(y|X,\theta^{(i)})}{p(y|X,\theta^{(i-1)})}dy \\
&= \int_y \log p(y|X,\theta^{(i)}) - \log p(y|X,\theta^{(i-1)})dy \\
&\leq \log \int_y p(y|X,\theta^{(i)}) - p(y|X,\theta^{(i-1)})dy \\
&= \log 1 \\
&= 0
\end{aligned}$$

Since logarithm is monotone increasing, we have $p(X|\theta^{(i)}) \geq p(X|\theta^{(i-1)})$. □

From this proof, we can see that to prove the convergence of iterative EM algorithm, we need to show that in i^{th} iteration likelihood of observed data is larger than $(i-1)^{th}$ iterations' estimated likelihood, since we our objective is to maximize the likelihood $p(X|\theta)$.

5 Example 1: Coin Tossing

This example is about hidden variable rather than missing values. Assume 3 coins A, B, C, and probabilities of tossing head are π, p, q respectively. Now there is an experiment, we toss coin A first. If A is head, then toss B; if A is tail, then toss C. We have 10 trails, and observed following results:

$$1, 1, 0, 1, 0, 0, 1, 0, 1, 1$$

Since we don't know if these results of B or C, we need to estimate the parameters π, p, q.

Proof. Let X denotes the observed data, and Y denotes the hidden missing variable, which is the tossing result of A. If there is only one observed data, then we write the likelihood as:

$$\begin{aligned} P(x|\theta) &= \sum_{Y=\{H,T\}} P(x, Y|\theta) && (16)\\ &= \sum_{Y=\{H,T\}} P(x|Y,\theta)P(Y|\theta) && (17)\\ &= P(x, Y = H|\theta) + P(x, Y = T|\theta) && (18)\\ &= \pi p^x(1-p)^x + (1-\pi)q^x(1-q)^x && (19) \end{aligned}$$

- Step 1, θ is $\{\pi, p, q\}$, Y is the hidden variable, which has the tossing result $Head$ or $Tail$ from A.

If A is head, the observation is from B, otherwise C. Now, we have observed data $X = \{x_1, x_2, ..., x_n\}$, then we have the likelihood as

$$P(X|\theta) = \prod_{i=1}^{n} [\pi p^{x_i}(1-p)^{x_i} + (1-\pi)q^{x_i}(1-q)^{x_i}]$$

and our goal is to find the optimal $\hat{\theta}$

$$\begin{aligned} \hat{\theta} &= arg\max_{\theta} L(\theta|X)\\ &= arg\max_{\theta} \log P(X|\theta)\\ &= arg\max_{\theta} \log \prod_{i=1}^{n} \left[\pi p^{x_i}(1-p)^{x_i} + (1-\pi)q^{x_i}(1-q)^{x_i}\right]\\ &= arg\max_{\theta} \sum_{i=1}^{n} \log \left[\pi p^{x_i}(1-p)^{x_i} + (1-\pi)q^{x_i}(1-q)^{x_i}\right] \end{aligned}$$

We cannot find the analytic form, because of the plus in the log function. Therefore, we have to use the EM algorithm.

Assume at $(i-1)$ iteration, we have estimated $\theta^{(i-1)} = \{\pi^{(i-1)}p^{(i-1)}q^{(i-1)}\}$.

E-step: write down the Q function

$$Q(\theta,\theta^{(i-1)}) = E_{Y|X}\big[\log P(X,Y|\theta)|X,\theta^{(i-1)}\big] \tag{20}$$

$$= \sum_{j=1}^{n}\sum_{Y=\{H,T\}} P(Y|x_j,\theta^{(i-1)})\log P(x_j,Y|\theta^{(i-1)}) \tag{21}$$

$$= \sum_{j=1}^{n}\sum_{Y=\{H,T\}} P(Y|x_j,\theta^{(i-1)})\log P(x_j,Y|\pi^{(i-1)}p^{(i-1),q^{(i-1)}}) \tag{22}$$

$$= \sum_{j=1}^{n}\sum_{Y=\{H,T\}} \frac{P(Y,x_j|\theta^{(i-1)})}{P(x_j|\theta^{(i-1)})}\log P(x_j,Y|\theta^{(i-1)}) \tag{23}$$

$$= \sum_{j=1}^{n}\Bigg\{\frac{\pi_{(i-1)}p_{(i-1)}^{x_j}(1-p_{(i-1)})^{1-x_j}}{\pi_{(i-1)}p_{(i-1)}^{x_j}(1-p_{(i-1)})^{1-x_j}+(1-\pi_{(i-1)})q_{(i-1)}^{x_j}(1-q_{(i-1)})^{1-x_j}}\log[\pi p^{x_j}(1-p)^{x_j}] \tag{24}$$

$$+\frac{(1-\pi_{(i-1)})q_{(i-1)}^{x_j}(1-q_{(i-1)})^{1-x_j}}{\pi_{(i-1)}p_{(i-1)}^{x_j}(1-p_{(i-1)})^{1-x_j}+(1-\pi_{(i-1)})q_{(i-1)}^{x_j}(1-q_{(i-1)})^{1-x_j}}\log[(1-\pi)q^{x_j}(1-q)^{x_j}]\Bigg\} \tag{25}$$

M-step: compute new parameter by taking derivative of the Q function. Here we denote

$$u_j = \frac{\pi_{(i-1)}p_{(i-1)}^{x_j}(1-p_{(i-1)})^{1-x_j}}{\pi_{(i-1)}p_{(i-1)}^{x_j}(1-p_{(i-1)})^{1-x_j}+(1-\pi_{(i-1)})q_{(i-1)}^{x_j}(1-q_{(i-1)})^{1-x_j}}$$

and

$$1-u_j = \frac{(1-\pi_{(i-1)})q_{(i-1)}^{x_j}(1-q_{(i-1)})^{1-x_j}}{\pi_{(i-1)}p_{(i-1)}^{x_j}(1-p_{(i-1)})^{1-x_j}+(1-\pi_{(i-1)})q_{(i-1)}^{x_j}(1-q_{(i-1)})^{1-x_j}}$$

Then, we have

$$\frac{\partial Q}{\partial \pi} = \sum_{j=1}^{N}\big(\frac{u_j}{\pi}-\frac{1-u_j}{1-\pi}\big) = \sum_{j=1}^{N}\frac{u_j-\pi}{\pi(1-\pi)} = \frac{\sum_{j=1}^{N}u_j - N\cdot\pi}{\pi(1-\pi)} = 0$$

We can do the same thing to parameter p and q, and we get

$$\pi_{(i)} = \frac{1}{N}\sum_{j=1}^{N}u_j \tag{26}$$

$$p_{(i)} = \frac{\sum_{j=1}^{N}u_jx_j}{\sum_{j=1}^{N}u_j} \tag{27}$$

$$q_{(i)} = \frac{\sum_{j=1}^{N}(1-u_j)x_j}{\sum_{j=1}^{N}(1-u_j)} \tag{28}$$

□

6 Example 2: Gaussian Mixture Model

GMM is a mixture-density parameter estimation problem. Gaussian mixture density distribution can be written as a linear superposition of Gaussian in the form

$$p(X|\theta) = \sum_{k=1}^{K}\alpha_k\phi(X|\theta_k) \tag{29}$$

where

- α_k is coefficient, and $\alpha_k \geq 0$ and $\sum_k^K \alpha_k = 1$
- $\phi(X|\theta_k)$ is Gaussian density function, where $\theta_k = (\mu_k, \sigma_k^2)$ and

$$\phi(X|\theta_k) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp\big(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\big)$$

  is the Gaussian distribution for k^{th} model.
- Note, here we assume that we have K component densities mixed together with K mixing coefficients α_i . We can use any p.d.f. to replace Gaussian with other distribution in the model.

Now, given a observed dataset X, and each sample in X is independent identical distributed. We can write the log likelihood of the observed data as

$$\begin{aligned}
\log(L(\Theta|X)) &= \log \prod_{i=1}^{N} p(X_i|\Theta) \\
&= \log \prod_{i=1}^{N} \big(\sum_{k=1}^{K} \alpha_k \phi(x_i|\Theta_k)\big) \\
&= \sum_{i=1}^{N} \log \big(\sum_{k=1}^{K} \alpha_k \phi(x_i|\Theta_k)\big)
\end{aligned}$$

If we want to maximize the log-likelihood to get the optimal parameters Θ_k, several problems we need to solve

- The log-likelihood contains log of the sum, there is no analytical form of the derivative of the log-sum function.
- Except the observed data X, any other hidden information ? Can we make any assumption on the observed data ?
- Can we replace the probability with others, which can make the log-sum easier to solve. For example, the *log*-sum-exp function is convex.

We may assume the observed data x_i is generated as follows:

according to probability α_k, we choose K^{th} Gaussian model $\phi(X|\Theta_k)$ and generate data x_i

This is common assumption of the generative model. Now, we know that we have the observed $X = \{x_1, ..., x_n\}$, but corresponding K^{th}-model is unobserved (hidden). Therefore, we have the random variable

$$y_{ik} = \begin{cases} 1 & \text{if } i^{th} \text{ data is generated by } k^{th} \text{ model.} \\ 0 & \text{otherwise} \end{cases}$$

Thus, for each data item, we have the complete data is $(x_i, y_{i1}, ..., y_{ik})$, and we can write the complete-data likelihood as

$$\begin{aligned}
L(\Theta|X,Y) &= p(X,Y|\Theta) \\
&= \prod_{i=1}^{N} p(x_i, y_i|\Theta) \\
&= \prod_{i=1}^{N} \prod_{k=1}^{K} \alpha_k^{y_{ik}} \phi(x_i|\Theta_k)^{y_{ik}}
\end{aligned}$$

and hence,

$$\log L(\Theta|X,Y) = \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \big[\log \alpha_k + \log \phi(x_i|\Theta_k)\big]$$

If we have the latent values of $\{y_i\}$, we can maximize the $\log L(\Theta|X,Y)$ by taking derivative w.r.t. $\{\alpha_k, \mu_k, \sigma_k^2\}$ separately. However, we don't have the these latent values, we have to consider the expectation, w.r.t. the posterior distribution of the latent variables, of the complete-data log-likelihood.

Since

$$y_{ik} \in \{0,1\},\ \sum_k y_{ik} = 1$$

and for each x_i,

$$p(y_{ik} = 1) = \alpha_k$$
$$0 \leq \alpha_k \leq 1, \sum_{k=1}^{K} \alpha_k = 1,$$

we can write in the form

$$p(y_i = k) = \prod_{k=1}^{K} \alpha_k^{y_{ik}} = \alpha_k$$

which also denotes the probability that data x_i is generated by k^{th}component. Then we can have the conditional distribution of x_i given a particular value for y_i is a Gaussian,

$$p(x_i|y_{ik} = 1) = \phi(x_i|\Theta_k)$$

which also can be written in the form

$$p(x_i|y_i) = \prod_{k=1}^{K} \phi(x_i|\Theta_k)^{y_{ik}} = \phi(x_i|\Theta_k)$$

Thus, we have

$$p(x_i|y_i) = \sum_{y_i} p(y_i)p(x_i|y_i) = \sum_{k=1}^{K} \alpha_k \phi(x_i|\Theta_k)$$

We find this marginal distribution is same as the Gaussian mixture density function on only observed data at the beginning of the section, whereas this involves an explicit latent variable. Now, we have

$$\begin{aligned} p(y_{ik} = 1|x_i, \Theta_k) &= \frac{p(y_{ik} = 1)p(x_i|y_{ik} = 1, \Theta_k)}{\sum_{j=1}^{K} p(y_{ij} = 1)p(x_i|y_{ij} = 1, \Theta_j)} \\ &= \frac{\alpha_k \phi(x_i|\Theta_k)}{\sum_{j=1}^{K} \alpha_j \phi(x_i|\Theta_j)} \\ &= \mathbb{E}(y_{ik}|x_i, \Theta_k) \\ &= 1 * p(y_{ik} = 1|x_i) + 0 * p(y_{ik} = 0|x_i) \end{aligned}$$

which can viewed as the responsibility that component k takes for explaining the observation of x_i.

Now we can write the expectation of complete data log-likelihood

$$\begin{aligned} Q(\Theta, \Theta^{(n-1)}) &= \mathbb{E}_{Y|X,\Theta^{(n-1)}}\big[\log p(X, Y|\Theta)\big] = \log p(X, Y|\Theta) \cdot p(Y|X, \Theta^{(n-1)}) \\ &= \log \prod_{i=1}^{N} \prod_{k=1}^{K} \alpha_k^{y_{ik}} \phi(x_i|\Theta_k)^{y_{ik}} \prod_{i=1}^{N} \prod_{k=1}^{K} p(y_{ik}|x_i, \Theta_k^{(n-1)})\big\} \\ &= \log \prod_{i=1}^{N} \prod_{k=1}^{K} \alpha_k^{y_{ik}} \phi(x_i|\Theta_k)^{y_{ik}} p(y_{ik}|x_i, \Theta_k^{(n-1)})\big\} \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik}\big[\log \alpha_k + \log \phi(x_i|\Theta_k)\big] p(y_{ik}|x_i, \Theta_k^{(n-1)}) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \cdot p(y_{ik}|x_i, \Theta_k^{(n-1)})\big[\log \alpha_k + \log \phi(x_i|\Theta_k)\big] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{E}\big[y_{ik}|x_i, \Theta_k^{(n-1)}\big] \cdot \big[\log \alpha_k + \log \phi(x_i|\Theta_k)\big] \end{aligned}$$

which is the E-step. And the M-step is just

$$\Theta^{(n)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(n-1)})$$

We repeat the E-step and M-step till convergence of either the log-likelihood or the parameter values.

References