# EM algorithm

January 25, 2018

## 1 Context

We have observed data $Y$ and unobserved data $Z$. The goal is to compute the likelihood of the data, $p_\theta(Y)$.

$$\log(p_\theta(Y)) = \log(p_\theta(Y,Z)) - \log(p_\theta(Z|Y)).$$

The advantage of this is to link $p_\theta(Y)$ with $p_\theta(Y,Z)$ which is easier to compute in general. We now take the expectation, conditioned on the data $Y$ :

$$\log(p_\theta(Y)) = \mathbb{E}_\theta\left(\log(p_\theta(Y,Z))|Y\right) \underbrace{-\mathbb{E}_\theta\left(\log(p_\theta(Y|Z))|Y\right)}_{\mathcal{H}(p_\theta(Y|Z))}$$

**E step :** Data $Y$ is considered fixed, leading the entropy term to be fixed as well. This step is dedicated to the computation of $\mathbb{E}_\theta\left(\log(p_\theta(Y,Z))|Y\right)$, which is the conditional expectation of the complete log-likelihood and where only the hidden part $Z$ is varying.

**M step :** We consider that $\theta$ is varying and we want to maximise the expectation with respect to these parameters. This step generally uses the value computed in the previous E step.

Repeating these steps, we get in then end optimised values of the parameters, which can give us some information about the hidden variable $Z$. We are also able to compute the likelihood of the model, but it is generally not the first interest and use of the EM algorithm.

## 2 Example of Gaussian mixture models

The data Y is an array of dimension $n \times d$, being for example $n$ samples of $d$ different species. Let $Y_i$ be the $i^{th}$ row (i.e. sample) of Y. We then assume that data from the species follow a mixture of K multivariate Gaussians :

$$\forall k \in \{1,..K\}, f_k(Y_i) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_k)}} \exp\left(-\frac{1}{2}(Y_i-\mu_k)^T\Sigma_k^{-1}(Y_i-\mu_k)\right)$$

With $d$ being the size of both $Y_i$ and $\mu_k$. The covariance matrix $\Sigma_k$ has size $d \times d$.

**E step :**

$$\log(p_\theta(Y,Z)) = \sum_{i,k} \mathbb{1}_{\{Z_i=k\}} \times \log(\pi_k f_k(Y_i)|Y)$$

$$\mathbb{E}_\theta(\log(p_\theta(Y,Z))|Y) = \sum_{i,k} \mathbb{E}_\theta\left(\mathbb{1}_{\{Z_i=k\}}|Y_i\right)\left[\log(\pi_k) + \log(f_k(Y_i))\right]$$

We can estimate the expectation with $\tau_{ik} = \frac{\pi_k f_k(Y_i)}{\sum_l \pi_l f_l(Y_i)}$ :

$$
\begin{aligned}
&= \sum_{i,k} \tau_{ik} [\log(\pi_k) + \log(f_k(Y_i))] \\
&= \sum_{i,k} \tau_{ik} \left[ \log(\pi_k) - \frac{1}{2} \log\left((2\pi)^d \det(\Sigma_k)\right) - \frac{1}{2}(Y_i - \mu_k)^T \Sigma_k^{-1} (Y_i - \mu_k) \right]
\end{aligned}
$$

**M step :** Maximising the last expression, we get after some algebraic manipulations :

- $\hat{\mu}_k = \frac{\sum_i \tau_{ik} y_i}{\sum_i \tau_{ik}}$
- $\hat{\Sigma}_k = \frac{\sum_i \tau_{ik} (y_i - \mu_k)^T (y_i - \mu_k)}{\sum_i \tau_{ik}}$
- $\hat{\pi}_k = \frac{1}{n} \sum_i \tau_{ik}$

## 3 Example of mixtures of Gaussian Dependence Trees

Let $T$ be a standard gaussian dependence tree : all means are null and all variances are equal to 1. We are considering a mixture of hidden trees, $k$ an $l$ are nodes of the trees ( i.e. variables or species).

$$
\mathbb{P}(T) = \frac{1}{B} \prod_{k,l \in T} \beta_{kl} \text{ , with } B = \sum_T \prod_{k,l \in T} \beta_{kl}
$$

$$
\begin{aligned}
\mathbb{P}(Y = y_i | T) &= \mathbb{P}(y_i^1 | T) \prod_{j=2}^{d} \frac{\mathbb{P}(y_i^j, y_i^{a_j} | T)}{\mathbb{P}(y_i^{a_j} | T)} \\
&= \underbrace{\prod_{j=1}^{d} \mathbb{P}(y_i^j | T)}_{\text{A}} \prod_{(k,l) \in T} \underbrace{\frac{\mathbb{P}(y_i^k, y_i^l | T)}{\mathbb{P}(y_i^k | T) \times \mathbb{P}(y_i^l | T)}}_{\psi_{kl}(Y_i)} \\
&= A \prod_{k,l \in T} \psi_{kl}(Y_i)
\end{aligned}
$$

Replacing the standard gaussian maximum likelihood estimates for the parameters, we know (cf. Chow gaussian document) that :

$$
\log(\hat{A}) = \sum_{j=1}^{d} -\frac{1}{2} \left( \log(2\pi \hat{\sigma}_j^2) + 1 \right),
$$

which is independant from the tree structure. We also know the explicit form of $\log(\psi_{kl})$ :

$$
\log(\psi_{kl}(Y_i)) = \frac{-1}{2} \left( \log\left(1 - \frac{\sigma_{kl}^2}{\sigma_k^2 \sigma_l^2}\right) + \frac{((y_i^k \sigma_l)2 + (y_i^l \sigma_k)^2 - 2\sigma_{kl} y_i^k y_i^k l)}{\det(\Sigma_{kl})} - \left( \left(\frac{y_i^k}{\sigma_k}\right)^2 + \left(\frac{y_i^l}{\sigma_l}^2\right) \right) \right)
$$

Remembering that we work with standard normal distributions, the last two expressions are greatly simplified. The correlation $\rho_{kl}$ between $y_k$ and $y_l$ is now their covariance too, and after some algebraic manipulations:

$$
\begin{aligned}
\log(\hat{A}) &= \sum_{i=1}^{n} -\frac{1}{2} (\log(2\pi) + 1) \\
\log(\psi_{kl}(Y_i)) &= \log\left(\frac{1}{\sqrt{1 - \rho_{kl}^2}}\right) + \frac{\rho_{kl}}{1 - \rho_{kl}^2} \cdot y_i^k y_i^l - \frac{\rho_{kl}^2}{1 - \rho_{kl}^2} \cdot \frac{(y_i^k)^2 + (y_i^l)^2}{2}
\end{aligned}
$$

### 3.1 E step :

$$\mathbb{P}(Y,T) = \mathbb{P}(T) \times \mathbb{P}(Y|T)$$

$$\begin{aligned}
\log(\mathbb{P}(Y,T)) &= \sum_{i=1}^{n} \left( \sum_{(k,l)\in T} \log(\beta_{kl}) + \log(\psi_{kl}(Y_i)) - \log(B) + \log(A) \right) \\
&= \sum_{i=1}^{n} \left[ \sum_{k,l} \mathbb{1}_{\{(k,l)\in T\}} \left(\log(\beta_{kl}) + \log(\psi_{kl}(Y_i))\right) \right] - n\log(B) + n\log(A)
\end{aligned}$$

Conditional expectation :

$$\mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y] = \sum_{i=1}^{n}\sum_{k,l} \mathbb{P}((k,l)\in T|Y_i) \times \left[\log(\beta_{kl}) + \sum_{i=1}^{n} \log(\psi_{kl}(Y_i))\right] - n\log(B) + n\log(A)$$

Computation of conditional probability : using Bayes, we specially consider the proportion of trees which contain an edge between the nodes $k$ and $l$.

$$\begin{aligned}
\mathbb{P}((k,l)\in T|Y_i) &= \sum_{T\in\mathcal{T}:(k,l)\in T} \mathbb{P}(T|Y_i) \\
&= \frac{\sum_{(k,l)\in T} \mathbb{P}(T)\mathbb{P}(Y_i|T)}{\sum_T \mathbb{P}(T)\mathbb{P}(Y_i|T)} \\
&= \frac{\sum_{(k,l)\in T} \prod_{uv} \beta_{uv}\psi_{uv}(Y_i)}{\sum_T \prod_{uv} \beta_{uv}\psi_{uv}(Y_i)}
\end{aligned}$$

We define the Laplacian matrix as the following symmetric matrix :

$$\mathcal{Q}_{uv}(W_\beta) = \begin{cases} -\beta_{uv} & 1 \le u < v \le n \\ \sum_{w=1}^{n} \beta_{wv} & 1 \le u = v \le n. \end{cases}$$

Lets $\mathcal{Q}^*$ be the first $(n-1)$ rows and columns of $\mathcal{Q}$. The Matrix Tree Theorem (MTT) of West [**?**] says that for any adjacence matrix $A$ of a multigraph G, $|\mathcal{Q}^*(A)|$ is the number of spanning trees of G, where $|\cdot|$ is the determinant. Meila *et al.* [**?**] demonstrate the generalization of the MTT (GMTT)for a real-valued matrix, so that we now get :

$$\mathbb{P}((k,l)\in T|Y) = 1 - \frac{|\mathcal{Q}^*(W_\beta^{-kl} \bigodot \psi)|}{|\mathcal{Q}^*(W_\beta \bigodot \psi)|}$$

Where the notation $W_\beta^{-kl}$ means that the entry at the $k^{\text{th}}$ line and $l^{\text{th}}$ column has been set to zero (we concretely erased the edge between nodes $k$ and $l$). This last quantity will be computed using the Kirshner theorem, allowing for a great gain in computation time.

### 3.2 M step :

Moving to the M step, the quantity $\tau_i^{kl} = \mathbb{P}((k,l)\in T|Y_i)$ has been computed and is now considered as fixed. We maximise the conditional expectation with respect to parameters $\beta_{kl}$.

$$\arg\max_{\beta_{kl}} \left\{ \sum_{i=1}^{n}\sum_{k,l} \tau_i^{kl} \times [\log(\beta_{kl}) + \log(\psi_{kl}(Y))] - n\log(B) + n\log(A) \right\}$$

We derive with respect to $\beta_{kl}$:

$$\frac{\partial \mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y]}{\partial \beta_{kl}} = \frac{1}{\beta_{kl}} \sum_{i=1}^{n} \tau_i^{kl} - \frac{n}{B}\frac{\partial B}{\partial \beta_{kl}} \tag{1}$$

Meila *et al.* give a formula for the derivative of $B$, using the GMTT. Lets define the $M(W_\beta)$ symmetric matrix with 0 diagonal such that :

$$\begin{cases} M_{uv} = [\mathcal{Q}^{*-1}]_{uu} + [\mathcal{Q}^{*-1}]_{vv} - 2[\mathcal{Q}^{*-1}]_{uv} & u, v < n \\ M_{nv} = M_{vn} = [\mathcal{Q}^{*-1}]_{vv} & v < n \\ M_{vv} = 0. \end{cases}$$

Meila *et al.* then demonstrate that

$$\begin{aligned} \frac{\partial B}{\partial \beta_{kl}} &= M_{kl} |\mathcal{Q}^*(W_\beta)| \\ &= M_{kl} \times B \end{aligned}$$

The last equality comes from the GMTT : $B = |\mathcal{Q}^*(W_\beta)|$. Replacing in equation 1 and setting the expression to 0 we get :

$$\boxed{\hat{\beta}_{kl} = \frac{1}{M_{kl}} \times \frac{1}{n} \sum_{i=1}^{n} \tau_i^{kl}}$$

# References