A variational interpretation of classification EM

Yves Auffray

A variational interpretation of classification EM

Yves Auffray*†

September 27, 2018

Abstract

EM and CEM algorithms which respectively implement mixture approach and classification approach of clustering problem are shown to be instances of the same variational scheme.

Keywords— Classification, Clustering, Mixtures, EM, Variational Methods

*Dassault Aviation
†INRIA Projet SELECT

1 Motivation

Model-based clustering approaches of partitioning problem have two possible forms, mixture approach and classification approach. The former leads to the Expectation-Maximisation algorithm (EM), while the latter is implemented by the Classification EM (CEM) procedure which is an iterative algorithm that maximizes a criterion called CML for Classification Maximum Likelihood (see [2] and [3]).
In this paper, both EM and CEM are shown to be intances of a more general variational algorithm scheme called AEM for (Approximated EM).

2 CML and CEM

Given a set E, we aim at partitioning a n-uplets $(x_1, \cdots, x_n) \in E^n$ into K classes.
Celeux & Govaert [2] and Govaert & Nadif [4] quote two CML-typed criteria:

$$C_1(\boldsymbol{\theta}, \mathbf{z}) = \sum_{i=1}^{n} \log(f_{\theta_{z_i}}(x_i)) = \sum_{k=1}^{K} \sum_{i:z_i=k} \log(f_{\theta_k}(x_i)) \tag{1}$$

$$C_2(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{z}) = \sum_{i=1}^{n} \log(\pi_{z_i} f_{\theta_{z_i}}(x_i)) = \sum_{k=1}^{K} \sum_{i:z_i=k} \log(\pi_k f_{\theta_k}(x_i)) \tag{2}$$

where the following notation has been used:

- $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ is a probability distribution on $\{1, \cdots, K\}$
- $(f_\theta)_{\theta \in \boldsymbol{\Theta}}$ is a family of densities with respect to a positive measure μ on E, whose parameter is $\theta \in \boldsymbol{\Theta}$
- $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K) \in \boldsymbol{\Theta}^K$
- $\mathbf{z} = (z_1, \cdots, z_n) \in \{1, \cdots, K\}^n$ specifies a partition of $(x_1, \cdots, x_n)$ into K classes.

$C_1(\boldsymbol{\theta}, \mathbf{z})$ is the log-likelihood of the parameter $(\boldsymbol{\theta}, \mathbf{z})$ when considering $(x_1, \cdots, x_n)$ as a realisation of a random n-uplet $(X_1, \cdots, X_n)$ whose density with respect to $\mu^{\otimes n}$ is

$$p^{(\boldsymbol{\theta}, \mathbf{z})}_{X_1, \cdots, X_n}(\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n) = \prod_{i=1}^{n} f_{\theta_{z_i}}(\boldsymbol{\xi}_i).$$

$C_2(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{z})$ is linked to $C_1(\boldsymbol{\theta}, \mathbf{z})$ by

$$C_2(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{z}) = C_1(\boldsymbol{\theta}, \mathbf{z}) + \sum_{j=1}^{K} \mathbf{Card}(\{i : z_i = j\}) \log(\pi_j).$$

Now, CEM implements the following pseudo-code:
Algorithm CEM

1. **Inputs:** $x_1, \cdots, x_n$
2. **Initialization:** $\boldsymbol{\pi}^{(0)}, \boldsymbol{\theta}^{(0)} \in \boldsymbol{\Theta}^K, t = 0$
3. **Loop:** $t = t + 1$

(a) **(E)** $f_i^{(t)} : j \in \{1, \cdots, K\} \mapsto \frac{\pi_j^{(t-1)} f_{\theta_j^{(t-1)}}(x_i)}{\sum_{m=1}^K \pi_m^{(t-1)} f_{\theta_m^{(t-1)}}(x_i)}, \quad i = 1, \cdots, N$

(b) **(C)** $z_i^{(t)} = \text{argmax}_j f_i^{(t)}(j)$

(c) **(M)** $(\boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)}) = \text{argmax}_{\boldsymbol{\pi},\boldsymbol{\theta}} C_2(\boldsymbol{\pi}, \boldsymbol{\theta}, \mathbf{z}^t) = \text{argmax}_{\boldsymbol{\pi},\boldsymbol{\theta}} \sum_{i=1}^n \log(\pi_{z_i^{(t)}} f_{\theta_{z_i^{(t)}}}(x_i))$

(d) **Convergence test**: $C_2(\boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)})$ still increasing ?

4. **Output:** $(\widetilde{\mathbf{z}}, \widetilde{\boldsymbol{\pi}}, \widetilde{\boldsymbol{\theta}}) = (\mathbf{z}^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)})$

$C_2(\boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)})$ increases with t. As the number of assignments $\mathbf{z} = (z_1, \cdots, z_n) \in \{1, \cdots, K\}^n$ of the $x_i, i = 1, \cdots, n$ to the K classes is finite, CEM necessary reaches a fixed point in a finite number of iterations.

3 Another view point

3.1 The key remark

We firstly recall a useful elementary fact.

Let (Y_1, Y_2) be a couple of random variables in $E_1 \times E_2$ with density p_{Y_1,Y_2} with respect to the product $\mu_1 \otimes \mu_2$ of two positive measures μ_1 on E_1 and μ_2 on E_2.
The following remark gives a useful expression of the log-evidence $\log(p_{Y_1}(\mathbf{y}_1))$

Remark. *Let f be a probability density on E_2 with respect to μ_2.*
For all $\mathbf{y}_1 \in E_1$ we have:

$$\log(p_{Y_1}(\mathbf{y}_1)) = \int_{E_2} f(\mathbf{y}_2) \log(p_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2)) \mu_2(d\mathbf{y}_2) + H(f) + \mathcal{K}(f, p_{Y_2|Y_1=\mathbf{y}_1}) \tag{3}$$

where $H(f)$ is the entropy of f and $\mathcal{K}(f, p_{Y_2|Y_1=\mathbf{y}_1})$ the Kullback divergence of f and $p_{Y_2|Y_1=\mathbf{y}_1}$.

Proof

$$\begin{aligned}
\log(p_{Y_1}(\mathbf{y}_1)) &= \int_{E_2} f(\mathbf{y}_2) \log(p_{Y_1}(\mathbf{y}_1)) \mu_2(d\mathbf{y}_2) \\
&= \int_{E_2} f(\mathbf{y}_2) \log(\frac{p_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2)}{p_{Y_2|Y_1=\mathbf{y}_1}(\mathbf{y}_2)}) \mu_2(d\mathbf{y}_2) \\
&= \int_{E_2} f(\mathbf{y}_2) \log(p_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2)) \mu_2(d\mathbf{y}_2) - \int_{E_2} f(\mathbf{y}_2) \log(p_{Y_2|Y_1=\mathbf{y}_1}(\mathbf{y}_2)) \mu_2(d\mathbf{y}_2) \\
&= \int_{E_2} f(\mathbf{y}_2) \log(p_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2)) \mu_2(d\mathbf{y}_2) - \int_{E_2} f(\mathbf{y}_2) \log(f(\mathbf{y}_2)) \mu_2(d\mathbf{y}_2) \\
&\quad + \int_{E_2} f(\mathbf{y}_2) \log(\frac{f(\mathbf{y}_2)}{p_{Y_2|Y_1=\mathbf{y}_1}(\mathbf{y}_2)}) \mu_2(d\mathbf{y}_2)
\end{aligned}$$

which gives (3) since

$$H(f) = -\int_{E_2} f(\mathbf{y}_2) \log(f(\mathbf{y}_2)) \mu_2(d\mathbf{y}_2)$$

and

$$\mathcal{K}(f, p_{Y_2|Y_1=\mathbf{y}_1}) = \int_{E_2} f(\mathbf{y}_2) \log(\frac{f(\mathbf{y}_2)}{p_{Y_2|Y_1=\mathbf{y}_1}(\mathbf{y}_2)}) \mu_2(d\mathbf{y}_2).$$

□
Now let us invoke what is commonly called free energy:

$$\mathcal{F}(\mathbf{y}_1, f) = \int_{E_2} f(\mathbf{y}_2) \log(p_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2))\mu_2(d\mathbf{y}_2) + H(f) \tag{4}$$

for $\mathbf{y}_1 \in E_1$ and a density f on E_2.
As an immediate consequence of (3) we deduce
Consequence : Let $\mathbf{y}_1 \in E_1$ and $\mathcal{D}$ a set of densities on E_2 which contains $p_{Y_2|Y_1=\mathbf{y}_1}$.
Then

$$\log(p_{Y_1}(\mathbf{y}_1)) = \max_{f\in\mathcal{D}} \mathcal{F}(\mathbf{y}_1, f) \tag{5}$$

and

$$p_{Y_2|Y_1=\mathbf{y}_1} =_{\mu_2-\text{a.s.}} \text{argmax}_{f\in\mathcal{D}}\mathcal{F}(\mathbf{y}_1, f). \tag{6}$$

So, the important fact lies in equation (5) which reformulates the computation of the log-evidence $\log(p_{Y_1}(\mathbf{y}_1))$ as a variational problem.

3.2 Application

3.2.1 AEM algorithm scheme

We are now in a parametric estimation context:

- $p^{\boldsymbol{\lambda}}_{(Y_1,Y_2)}$, the density of (Y_1, Y_2) with respect to $\mu_1 \otimes \mu_2$, depends on the parameter $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$
- from a realisation of (Y_1, Y_2), we only observe $\mathbf{y}_1$, its E_1 component.
- we want to calculate the maximum likelihood estimator $\widehat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$:

$$\widehat{\boldsymbol{\lambda}} = \text{argmax}_{\boldsymbol{\lambda}}\ \log(p^{\boldsymbol{\lambda}}_{Y_1}(\mathbf{y}_1))$$

In this context, if $\mathcal{D}$ is a set of density which contains $p^{\boldsymbol{\lambda}}_{Y_2|Y_1=\mathbf{y}_1}$, (5) reads

$$\log(p^{\boldsymbol{\lambda}}_{Y_1}(\mathbf{y}_1)) = \max_{f\in\mathcal{D}} \mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}, f).$$

Hence

$$\max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \log(p^{\boldsymbol{\lambda}}_{Y_1}(\mathbf{y}_1)) = \max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \max_{f\in\mathcal{D}} \mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}, f). \tag{7}$$

This equality suggests that in order to maximise $\log(p^{\boldsymbol{\lambda}}_{Y_1}(\mathbf{y}_1))$ in $\boldsymbol{\lambda}$, $\mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}, f)$ could be maximized alternatively in $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ and in $f \in \mathcal{D}$.
Moreover, if we relax the condition on $\mathcal{D}$ containing $p^{\boldsymbol{\lambda}}_{Y_2|Y_1=\mathbf{y}_1}$, $\widehat{\boldsymbol{\lambda}}$ will be only approximated.
It is what the following AEM (Approximated EM) algorithm achieves.
$\mathcal{D}$ being any set of densities on E_2:
Algorithm AEM

1. **Input:** $\mathbf{y}_1$
2. **Initialization:** $\boldsymbol{\lambda}^{(0)} \in \boldsymbol{\Lambda}, t = 0$
3. **Loop:** $t = t + 1$
   (a) **(AE)** $f^{(t)} = \text{argmax}_{f\in\mathcal{D}}\mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}^{(t-1)}, f)$

(b) (AM)

$$\boldsymbol{\lambda}^{(t)} = \text{argmax}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}}\mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}, f^{(t)}) = \text{argmax}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \int_{E_2} f^{(t)}(\mathbf{y}_2) \log(p^{\boldsymbol{\lambda}}_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2))\mu_2(d\mathbf{y}_2)$$

(c) Convergence test[1]**:** $\mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}^{(t)}, f^{(t)}) - \mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}^{(t-1)}, f^{(t-1)}) \leq \epsilon$ **?**

4. **Output:** $(\widetilde{\boldsymbol{\lambda}}, \widetilde{f}) = (\boldsymbol{\lambda}^{(t)}, f^{(t)})$.

When $\mathcal{D}$ contains $\{p^{\boldsymbol{\lambda}}_{Y_2|Y_1=\mathbf{y}_1} : \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\}$ we recover EM:
Algorithm EM

1. **Input:** $\mathbf{y}_1$
2. **Initialization:** $\boldsymbol{\lambda}^{(0)} \in \boldsymbol{\Lambda}, t = 0$
3. **Loop:** $t = t + 1$
   (a) (E) $f^{(t)} = p^{\boldsymbol{\lambda}^{(t-1)}}_{Y_2|Y_1=\mathbf{y}_1}$
   (b) (M) $\boldsymbol{\lambda}^{(t)} = \text{argmax}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \int_{E_2} f^{(t)}(\mathbf{y}_2) \log(p^{\boldsymbol{\lambda}}_{Y_1,Y_2}(\mathbf{y}_1, \mathbf{y}_2))\mu_2(d\mathbf{y}_2)$
   (c) Convergence Test
4. **Output:** $(\widetilde{\boldsymbol{\lambda}}, \widetilde{f}) = (\boldsymbol{\lambda}^{(t)}, f^{(t)})$.

If E_2 is finite and μ_2 is the counting measure, another instance of AEM is obtained, by choosing

$$\mathcal{D} = \{\delta_{\mathbf{y}_2} : \mathbf{y}_2 \in E_2\}$$

where $\delta_{\mathbf{y}_2}$ is a Dirac measure.
We immediatly get

$$\mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}, \delta_{\mathbf{y}_2}) = \log(p^{\boldsymbol{\lambda}}_{(Y_1,Y_2)}(\mathbf{y}_1, \mathbf{y}_2)) \tag{8}$$

and AEM becomes:
Algorithm δEM

1. **Input:** $\mathbf{y}_1$
2. **Initialization:** $\boldsymbol{\lambda}^{(0)} \in \boldsymbol{\Lambda}, t = 0$
3. **Loop:** $t = t + 1$
   (a) (δE) $\mathbf{y}_2^{(t)} = \text{argmax}_{\mathbf{y}_2\in E_2} \log(p^{\boldsymbol{\lambda}^{(t-1)}}_{(Y_1,Y_2)}(\mathbf{y}_1, \mathbf{y}_2))$
   (b) (δM) $\boldsymbol{\lambda}^{(t)} = \text{argmax}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \log(p^{\boldsymbol{\lambda}}_{(Y_1,Y_2)}(\mathbf{y}_1, \mathbf{y}_2^{(t)}))$
   (c) Convergence test
4. **Output:** $(\widetilde{\boldsymbol{\lambda}}, \widetilde{\mathbf{y}_2}) = (\boldsymbol{\lambda}^{(t)}, \mathbf{y}_2^{(t)})$.

[1] $t \in \mathbb{N} \mapsto \mathcal{F}(\mathbf{y}_1, \boldsymbol{\lambda}^{(t)}, f^{(t)})$ is an increasing function.

4 Application to clustering

4.1 Context

We now have data $(x_1, \cdots, x_n) \in E^n$ being a set E with a positive measure μ. We want to partition these data among K classes. The data are considered as the observable part of a realisation

$$((x_1, z_1), \cdots, (x_n, z_n)) \in (E \times \{1, \cdots, K\})^n$$

of a n-sample

$$((X_1, Z_1), \cdots, (X_n, Z_n))$$

of a couple of random variables (X, Z) whose density with respect[2] to $\mu \otimes \nu_K$ has the following form:

$$p_{X,Z}^{(\boldsymbol{\pi},\boldsymbol{\theta})} : (\xi, \zeta) \in E \times \{1, \cdots, K\} \mapsto \pi_\zeta f_{\theta_\zeta}(\xi) \tag{9}$$

where

- $(f_\theta)_{\theta \in \boldsymbol{\Theta}}$ is a family of densities on (E, μ)
- $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ belongs to $\boldsymbol{\Theta}^K$
- $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ is a probability measure on $\{1, \cdots, K\}$.

Whence

- X is a mixture: $p_X^{(\boldsymbol{\pi},\boldsymbol{\theta})}(\xi) = \sum_{k=1}^K \pi_k f_{\theta_k}(\xi)$
- $p_{Z|X=x}^{(\boldsymbol{\pi},\boldsymbol{\theta})}(\zeta) = \frac{\pi_\zeta f_{\theta_\zeta}(x)}{\sum_{k=1}^K \pi_k f_{\theta_k}(x)}$

And the density of $(X_1, \cdots, X_n, Z_1, \cdots, Z_n)$ with respect to $(\mu \otimes \nu_K)^n$ reads

$$p_{X_1,\cdots,X_n,Z_1,\cdots,Z_n}^{(\boldsymbol{\pi},\boldsymbol{\theta})} : (\xi_1, \cdots, \xi_n, \zeta_1, \cdots, \zeta_n) \in E^n \times \{1, \cdots, K\}^n \mapsto \prod_{i=1}^n \pi_{\zeta_i} f_{\theta_{\zeta_i}}(\xi_i) \tag{10}$$

while

$$p_{Z_1,\cdots,Z_n|X_1=x_1,\cdots,X_n=x_n,}^{(\boldsymbol{\pi},\boldsymbol{\theta})}(\zeta_1, \cdots, \zeta_n) = \prod_{i=1}^n p_{Z_i|X_i=x_i}^{(\boldsymbol{\pi},\boldsymbol{\theta})}(\zeta_i) = \prod_{i=1}^n \frac{\pi_{\zeta_i} f_{\theta_{\zeta_i}}(x_i)}{\sum_{k=1}^K \pi_k f_{\theta_k}(x_i)}. \tag{11}$$

Thus we are precisely in the situation of section 3.1 with:

- $E_1 = E^n, \mu_1 = \mu^{\otimes n},\ E_2 = \{1, \cdots, K\}^n, \mu_2 = \nu_K^{\otimes n}$
- $Y_1 = (X_1, \cdots, X_n), Y_2 = (Z_1, \cdots, Z_n)$
- $\mathbf{y}_1 = (x_1, \cdots, x_n)$, $\boldsymbol{\lambda} = (\boldsymbol{\pi}, \boldsymbol{\theta})$.y

From (3), the log-likelihood $\log(p_{X_1,\cdots,X_n}^{(\boldsymbol{\pi},\boldsymbol{\theta}))}(x_1, \cdots, x_n)) = \sum_{i=1}^n \log(\sum_{k=1}^K \pi_k f_{\theta_k}(x_i))$ takes the form:

$$\log(p_{X_1,\cdots,X_n}^{(\boldsymbol{\pi},\boldsymbol{\theta}))}(x_1, \cdots, x_n)) = \mathcal{F}(x_1, \cdots, x_n, (\boldsymbol{\pi}, \boldsymbol{\theta}), f) + \mathcal{K}(f, p_{Z_1,\cdots,Z_n|X_1=x_1,\cdots,X_n=x_n})$$

[2] ν_K is the counting measure on $\{1, \cdots, K\}$

with

$$\mathcal{F}(x_1,\cdots,x_n,(\boldsymbol{\pi},\boldsymbol{\theta}),f)=\sum_{(\zeta_1,\cdots,\zeta_n)\in\{1,\cdots,K\}^n} f(\zeta_1,\cdots,\zeta_n)\log(\prod_{i=1}^{n}\pi_{\zeta_i}f_{\theta_{\zeta_i}}(x_i))+H(f)$$

for any probability measure f on $\{1,\cdots,K\}^n$.
Futhermore, if $f:(\zeta_1,\cdots,\zeta_n)\mapsto f_1(\zeta_1)\cdots f_n(\zeta_n)$ is a tensorial product, as $p^{(\boldsymbol{\pi},\boldsymbol{\theta})}_{Z_1,\cdots,Z_n|X_1=x_1,\cdots,X_n=x_n}$ according to (11), we get

$$\mathcal{F}(x_1,\cdots,x_n,(\boldsymbol{\pi},\boldsymbol{\theta}),f)=\sum_{i=1}^{n}\sum_{k=1}^{K} f_i(k)\log(\pi_k f_{\theta_k}(x_i))+H(f). \tag{12}$$

4.2 EM and $\boldsymbol{\delta}$EM for clustering

In this context EM reads:
Algorithm EM (clustering)

1. **Input:** $x_1,\cdots,x_n$
2. **Initialization:** $\boldsymbol{\pi}^{(0)},\boldsymbol{\theta}^{(0)},t=0$
3. **Loop:** $t=t+1$
   (a) **(E)** $f^{(t)}=p^{(\boldsymbol{\pi}^{(t-1)},\boldsymbol{\theta}^{(t-1)})}_{Z_1,\cdots,Z_n|X_1=x_1,\cdots,X_n=x_n,}=\otimes_{i=1}^{n}p^{(\boldsymbol{\pi}^{(t-1)},\boldsymbol{\theta}^{(t-1)})}_{Z_i|X_i=x_i}$
   where $p^{(\boldsymbol{\pi}^{(t-1)},\boldsymbol{\theta}^{(t-1)})}_{Z_i|X_i=x_i}(\zeta_i)=\frac{\pi^{(t-1)}_{\zeta_i}f_{\theta^{(t-1)}(x_i)}}{\sum_{k=1}^{K}\pi^{(t-1)}_k f_{\theta^{(t-1)}_k}(x_i)}$
   (b) **(M)** $(\boldsymbol{\pi}^{(t)},\boldsymbol{\theta}^{(t)})=\text{argmax}_{(\boldsymbol{\pi},\boldsymbol{\theta})}\sum_{i=1}^{n}\sum_{k=1}^{K}p^{(\boldsymbol{\pi}^{(t-1)},\boldsymbol{\theta}^{(t-1)})}_{Z_i|X_i=x_i}(k)\log(\pi_k f_{\theta^{(t-1)}_k}(x_i))$
   (c) **Convergence test**
4. **Ouput:** $(\widetilde{\boldsymbol{\pi}},\widetilde{\boldsymbol{\theta}})=(\boldsymbol{\pi}^{(t)},\boldsymbol{\theta}^{(t)})$

The assignments $(z_1,\cdots,z_n)\in\{1,\cdots,K\}^n$ of the $x_1,\cdots,x_n$ to the classes are obtained by

$$z_i=\text{argmax}_{j\in\{1,\cdots,K\}}\widetilde{\pi}_j f_{\widetilde{\theta_j}}(x_i)$$

where $\widetilde{\boldsymbol{\pi}}=(\widetilde{\pi_1},\cdots,\widetilde{\pi_K})$ and $\widetilde{\boldsymbol{\theta}}=(\widetilde{\theta_1},\cdots,\widetilde{\theta_K})$ are the EM outputs.

And what happens to $\boldsymbol{\delta}$**EM** in this context ?
Let us first note that if $f=\delta_{(z_1,\cdots,z_n)}$, we have

$$\mathcal{F}(x_1,\cdots,x_n,(\boldsymbol{\pi},\boldsymbol{\theta}),\delta_{(z_1,\cdots,z_n)})=\sum_{i=1}^{n}\log(\pi_{z_i}f_{\theta_{z_i}}(x_i))$$

which is exactly $C_2(\boldsymbol{\pi},\boldsymbol{\theta},\mathbf{z})$ (see (2)).
As for $\boldsymbol{\delta}$**EM**, it reads:
Algorithm $\boldsymbol{\delta}$EM(clustering-C2)

1. **Input:** $x_1,\cdots,x_n$

2. **Initialization:** $\boldsymbol{\pi}^{(0)}, \boldsymbol{\theta}^{(0)}, t = 0$

3. **Loop:** $t = t + 1$

   (a) **(δE)** $\mathbf{z}^{(t)} = (z_1^{(t)}, \cdots, z_n^{(t)}) = \text{argmax}_{\boldsymbol{\zeta} \in \{1,\cdots,K\}^n} \sum_{i=1}^n \log(\pi_{\zeta_i}^{(t-1)} f_{\theta_{\zeta_i}^{(t-1)}}(x_i))$

   i.e $z_i^{(t)} = \text{argmax}_{\zeta \in \{1,\cdots,K\}} \log(\pi_\zeta^{(t-1)} f_{\theta_\zeta^{(t-1)}}(x_i))$

   (b) **(δM)** $(\boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)}) = \text{argmax}_{(\boldsymbol{\pi},\boldsymbol{\theta})} \sum_{i=1}^n \log(\pi_{z_i^{(t)}} f_{\theta_{z_i^{(t)}}}(x_i))$

   (c) **Convergence test**

4. **Output:** $\begin{cases} (\widetilde{\boldsymbol{\pi}}, \widetilde{\boldsymbol{\theta}}) = (\boldsymbol{\pi}^{(t)}, \boldsymbol{\theta}^{(t)}) \\ \widetilde{\mathbf{z}} = \mathbf{z}^{(t)} \end{cases}$.

And **δEM(clustering-C2)** is nothing else than CEM.

4.3 Where we refind $C_1(\boldsymbol{\theta}, \mathbf{z})$

If we specialize model (9) and use

$$p_{X,Z}^{(\boldsymbol{\theta})} : (\xi, \zeta) \in E \times \{1, \cdots, K\} \mapsto \frac{1}{K} f_{\theta_\zeta}(\xi)$$

for the density of (X, Z), we get this log-likelihood

$$\log(p_{X_1,\cdots,X_n}^{(\boldsymbol{\theta})})(x_1, \cdots, x_n)) = \mathcal{F}(x_1, \cdots, x_n, \boldsymbol{\theta}, f) + \mathcal{K}(f, p_{Z_1,\cdots,Z_n|X_1=x_1,\cdots,X_n=x_n})$$

where (12) gives

$$\mathcal{F}(x_1, \cdots, x_n, \boldsymbol{\theta}, f) = \sum_{i=1}^n \sum_{k=1}^K f_i(k) \log(\frac{1}{K} f_{\theta_k}(x_i)) + H(f)$$

if $f = f_1 \otimes \cdots \otimes f_n$ is a tensorial product on $\{1, \cdots, K\}$.
If we apply **δEM** in this context, we first remark that

$$\mathcal{F}(x_1, \cdots, x_n, \boldsymbol{\theta}, \delta_{z_1,\cdots,z_n}) = \sum_{i=1}^n \log(\frac{1}{K} f_{\theta_{z_i}}(x_i)) = C_1(\boldsymbol{\theta}, \mathbf{z}) - n \log(K)$$

which leads to this algorithm:
Algorithm δEM(clustering-C1)

1. **Input:** $x_1, \cdots, x_n$

2. **Initialization:** $\boldsymbol{\theta}^{(0)}, t = 0$

3. **Loop:** $t = t + 1$

   (a) **(δE)** $\mathbf{z}^{(t)} = (z_1^{(t)}, \cdots, z_n^{(t)}) = \text{argmax}_{\boldsymbol{\zeta} \in \{1,\cdots,K\}^n} \sum_{i=1}^n \log(f_{\theta_{\zeta_i}^{(t-1)}}(x_i))$

   i.e $z_i^{(t)} = \text{argmax}_{\zeta \in \{1,\cdots,K\}} \log(f_{\theta_\zeta^{(t-1)}}(x_i))$

   (b) **(δM)** $\boldsymbol{\theta}^{(t)} = \text{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \log(f_{\theta_{z_i^{(t)}}}(x_i))$

   (c) **Convergence test**

4. **Ouput:** $(\widetilde{\mathbf{z}}, \widetilde{\boldsymbol{\theta}}) = (\mathbf{z}^{(t)}, \boldsymbol{\theta}^{(t)})$.

Particularly, when $E = \mathbb{R}^d$ and f_θ is the normal density with mean $\theta \in \mathbb{R}^d$ and variance-covariance matrix $\text{Id}_{\mathbb{R}^d}$, (**δEM(clustering-C1)**) is the K-means algorithm.

5 Discussion

This paper shows how CEM algorithm can be seen, like EM, as solving the variational problem which naturally results from clustering problem modelisation.
Due to restrictions on argument of this variational problem, unlike that of EM, the solution of CEM is an approximation. However, both are instances of a general scheme, that we prefer to call Approximated EM (AEM) instead of Variational EM as it is usually named: indeed EM is already a variational algorithm, and the real difference between EM and AEM is that the latter calculate an approximation.
We have taken the greatest care to recall the key argument of this type of variational technique in its extreme and remarkable simplicity. Of course we find this argument in many authors (Bishop [1] for example), but, as a matter of fact, many other authors obfuscate this simplicity by parasitic considerations or simply ignore the argument itself.

References

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 47:315-332, 1992.

[3] McLachlan G.J. The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistic*.

[4] G. Govaert and M. Nadif. *Co-Clustering, Models, Algorithms and Applications*. Wiley, 2014.