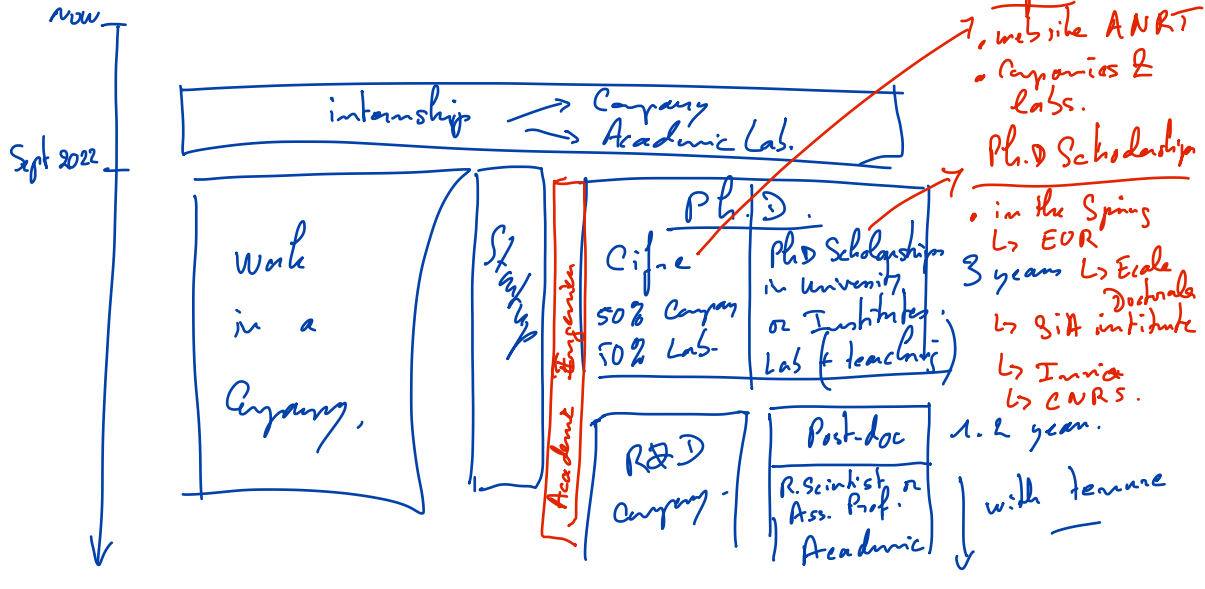
Model-based Statistical Learning

Pr. Charles BOUVEYRON

Professor of Statistics
Chair of the Institut 3IA Côte d'Azur
Université Côte d'Azur & Inria

charles.bouveyron@univ-cotedazur.fr
@cbouveyron

After the MSc :



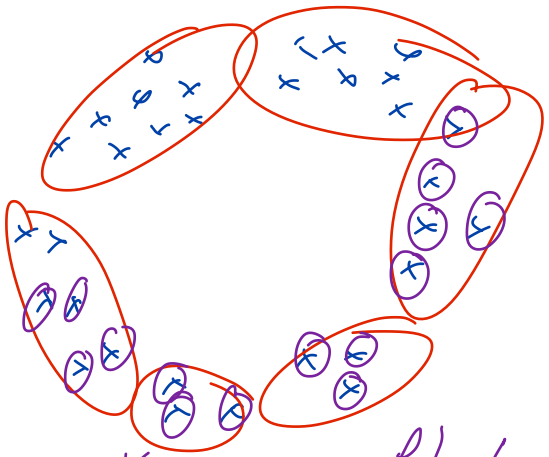
The Gaussian mixture model (GMM)

The GMM is probably the most popular Mixt. model for two main reasons:

- a bad one: it is the MM for which the calculations are the simplest.

- a good one: even though it is a simple model it is flexible enough to fit a large variety of situations.

The GMM: $p(x;\Theta) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \Sigma_k)$



$K = 6$.

$K = 50$

When varying K we can fit to any data distribution. The limit will be to choose K in an appropriate manner.

Model inference in the Gaussian mixture model is not easy because of the specific form of the log-likelihood:

$$\log \mathcal{L}(x;\theta) = \log \left(\prod_{i=1}^{n} \sum_{k=1}^{4} \pi_k \phi(x_i; \mu_k, \Sigma_k) \right)$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{4} \pi_k \phi(\text{———})$$

$\Rightarrow$ the classical solution is to use the EM algo.

The EM algorithm

The idea of the EM algo. is to first revisit the model by introducing a **latent variable** $z \in [0,1]^k$ to encode the class membership.

$$z_{ih} = \begin{cases} 1 & \text{if } x_i \text{ belongs to cluster } h. \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{cases} z|\pi \sim \mathcal{M}(1;\pi) \\ x|z=h \sim \mathcal{N}(x;\mu_h;\Sigma_h) \\ p(z=h) = \pi_h. \end{cases}$$

integrate over z $\Rightarrow$ $p(x|\theta) = \sum_{h=1}^{k} \pi_h \phi(x;\mu_h,\Sigma_h)$

this allows to write the complete likelihood of the couple (X, Z):

$$\log \mathcal{L}(x, z; \theta) = \sum_{i=1}^{n} \left[\log p(z_i | x_i; \theta) + \log p(x_i; \theta) \right]$$

$$= \log \mathcal{L}(x; \theta) + \sum_{i=1}^{n} \log p(z_i | x_i; \theta)$$

$$\Leftrightarrow \log \mathcal{L}(x; \theta) = \log \mathcal{L}(x, z; \theta) - \sum_{i=1}^{n} \log p(z_i | x_i; \theta)$$

there, we demonstrated that the green part is a lower bound of the log likelihood $\Rightarrow$ if we know Z, we can maximize the LB instead of the log likelihood.

The spirit of the EM algorithm is to alternate two steps:

- the Expectation step: knowing a certain value of θ^*, we compute the expectation of the lower bound:

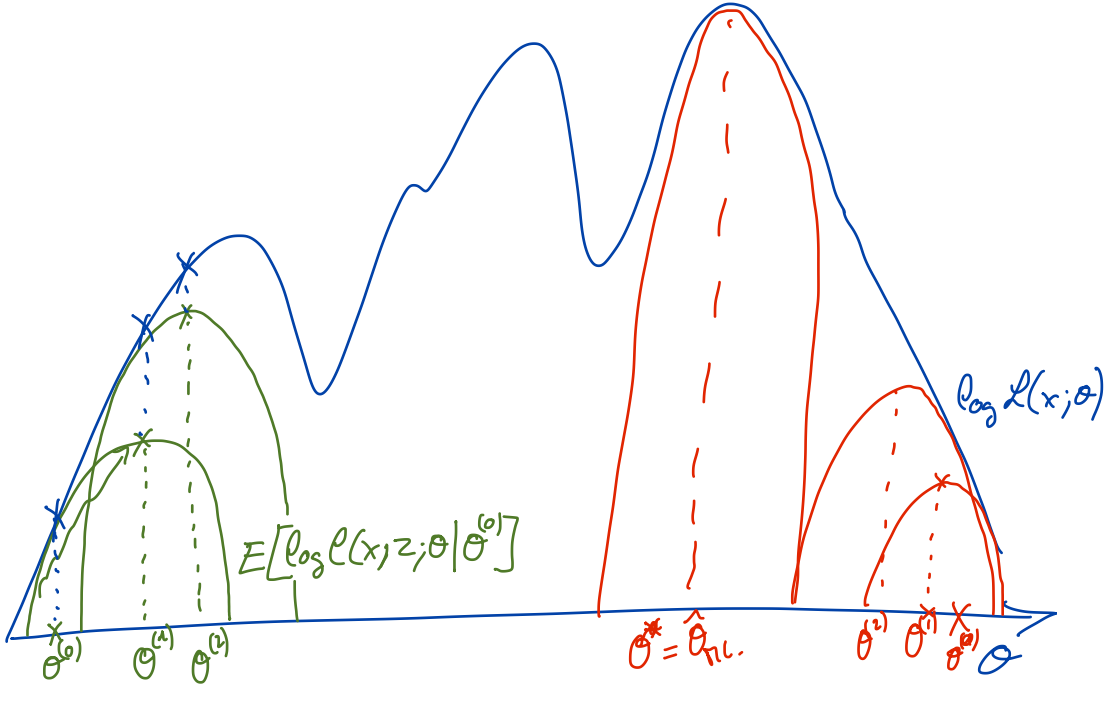
$$E\left[\ell(x, z|\theta)|\theta^*\right] = Q(\theta|\theta^*)$$

- the Maximization step: optimize $Q(\theta|\theta^*)$ over θ to obtain a new estimate of θ^*

The EM algorithm

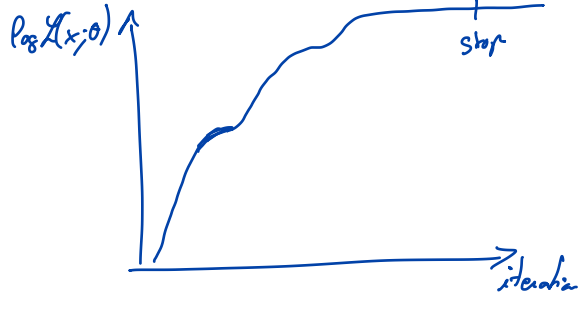
Theorem (Dempster, Laird, Rubin, 1979 + Wu, 1981)

The series of parameters $(\theta^*)_q$ generated by the EM algorithm converges toward a local maximum of the log-likelihood $\log \mathcal{L}(x;\theta)$.


$\log \mathcal{L}(x;\theta)$
$E[\log \mathcal{L}(x,z;\theta|\theta^{(t)}]$
$\theta^{(0)}$
$\theta^{(1)}$
$\theta^{(2)}$
$\theta^* = \hat{\theta}_{ML}$
$\theta^{(2)}$
$\theta^{(1)}$
$\theta^{(0)}$
θ

In practice, to avoid a sensibility to the initialization, we recommend to try several random initialization for $\theta^{(0)}$ and retain at the end the θ^* that leads to the maximum likelihood.

In practice, we also stop the algorithm when a plateau of the likelihood is detected:



The EM algorithm for GMM

The central quantity to calculate in the E step is

$$Q(\theta|\theta^*) = E\left[\ell(x,z|\theta)\,\middle|\,\theta^*\right]$$

and $$\ell(x,z|\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\pi_k\,\phi(x_i;\mu_k,\Sigma_k)\right)$$

and therefore: $$Q(\theta|\theta^*) = \sum_i\sum_k E[z_{ik}|\theta^*]\log\left(\pi_k\,\phi(x_i;\mu_k,\Sigma_k)\right)$$

So, the E step for the GMM reduces to the computation of

$$\gamma_{ik} = E[z_{ik}|\theta^*] \overset{\text{Bayes}}{\propto} P(z_{ik}=1|\theta^*)\,P(x_i|z_{ik}=1,\theta^*)$$
$$\pi_k^*\cdot\phi(x_i;\mu_k^*,\Sigma_k^*)$$

The EM algorithm for GMM

The E step: $\gamma_{ik} \propto \pi_k^* \cdot \phi(x_i; \mu_k^*, \Sigma_k^*)$ $\forall i$ $\forall k$.

The M step: Maximize over π_k, μ_k, Σ_k the function

$$Q(\theta|\theta^*) = \sum_i \sum_k \gamma_{ik} \log\left(\pi_k \phi(x_i; \mu_k, \Sigma_k)\right)$$

where $\phi(x_i; \mu_k, \Sigma_k) = \frac{1}{|\Sigma_k|^{\frac{1}{2}} (2\pi)^d} \exp\left(-\frac{1}{2}(x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k)\right)$

where d is the dimensionality of $x_i \in \mathbb{R}^d$

The EM algorithm for GMM

The update equations for π_k, μ_k and Σ_k can be obtained by simply taking the partial derivatives of $Q(\theta|\theta^*)$ regarding π_k, μ_k and Σ_k respectively and equalling to 0

$$\frac{\partial}{\partial \mu_k} Q(\theta|\theta^*) = 0 \iff \mu_k^* = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} x_i \quad \text{where } n_k = \sum_n \gamma_{ik}.$$

$$\frac{\partial}{\partial \Sigma_k} Q(\theta|\theta^*) = 0 \iff \Sigma_k^* = \frac{1}{n_k} \sum_{i=1}^{n} \gamma_{ik} (x_i - \mu_k^*)^t (x_i - \mu_k^*)$$

$$\frac{\partial}{\partial \pi_k} Q(\,) \text{ under the constraint } \sum_k \pi_k = 1 \; = 0 \iff \pi_k^* = \frac{n_k}{n}$$

The EM algorithm for GMM

Compute $\frac{\partial}{\partial \mu_k} Q(\theta|\theta^*)$ where

$$Q(\theta|\theta^*) = \sum_i \sum_k \gamma_{ik} \log\left(\pi_k \phi(x_i; \mu_k, \Sigma_k)\right)$$

where $\phi(x_i; \mu_k, \Sigma_k) = \frac{1}{|\Sigma_k|^{\frac{1}{2}} (2\pi)^d} \exp\left(-\frac{1}{2}(x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k)\right)$,

so $Q(\theta|\theta^*) = \sum_i \sum_k \gamma_{ik} \left[\log \pi_k + \log \phi(x_i; \mu_k, \Sigma_k)\right]$

$$= \sum_i \sum_k \gamma_{ik} \left[\log \pi_k - d \log(2\pi) - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k)\right]$$

$$\frac{\partial}{\partial \mu_k} Q(\theta|\theta^*) = \sum_i \gamma_{ik} \frac{\partial}{\partial \mu_k} \left[-\frac{1}{2}(x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k)\right]$$

cf. the "book": The Matrix Cook Book. $\frac{\partial}{\partial v} v^t B v = 2Bv$

$$\frac{\partial}{\partial \mu_k} Q = \sum_i \gamma_{ik} \left[-\frac{2}{2} \Sigma_k^{-1} (x_i - \mu_k) \right]$$

$$= -\sum_i \gamma_{ik} \Sigma_k^{-1} (x_i - \mu_k)$$

and now $\frac{\partial}{\partial \mu_k} Q = 0 \iff \sum_i \gamma_{ik} \Sigma_k^{-1} (x_i - \mu_k) = 0$

$$\iff \sum_i \gamma_{ik} (x_i - \mu_k) = 0$$

$$\Rightarrow \mu_k^* = \frac{\sum_i \gamma_{ik} x_i}{\sum_i \gamma_{ik}} = \frac{1}{n_k} \sum_i \gamma_{ik} x_i$$

Home work: do the calculations for Σ_k^* $\quad tr(AB) = tr(BA)$ $\quad tr(\lambda) = \lambda$

$\frac{\partial}{\partial A} \text{trace}(A^{-1}B) = -(A^{-1}BA^{-1})^t$ $\quad \frac{\partial}{\partial A} \log|A| = (A^{-1})^t$

Model selection: how to choose K?

Model selection criteria all penalized the likelihood with a quantity which favors models with a low number of groups.

$$\text{MS Criteria} = \log \mathcal{L}(x;\hat{\theta}) - \text{pen}(K)$$

Two popular criteria:

$$\text{AIC} = \log \mathcal{L}(x;\hat{\theta}) - \nu(\mathcal{M})$$

$$\text{BIC} = \log \mathcal{L}(x;\hat{\theta}) - \frac{1}{2}\nu(\mathcal{M})\log(n)$$

where $\nu(\mathcal{M})$ is the number of free scalar parameters in the model $\mathcal{M}$.

In practice, $\nu(GMM)$ is easy to compute

$$\nu(GMM \text{ with } K) = \text{nb of } \pi_k + \text{nb of } \mu_k + \text{nb of } \Sigma_k$$

$$= (K-1) + Kd + K\frac{d(d+1)}{2}$$

$$Bic(GMM) = \log \mathcal{L}(x;\hat{\theta}) - \frac{(K-1) + Kd + K\frac{d(d+1)}{2}}{2} \log(n)$$