# EM algorithm for mixture of Gaussian distributions

julien.chiquet@gmail.com

M2 ISG - summer 2016 -
network modelling - practical 2

**Instructions.** Each student must send a small report generated with `R` markdown answering the questions and including comments associated with the `R` code.

## 1 Context

We consider a collection of random variables $(X_1, \ldots, X_n)$ associated with $n$ individuals drawn from $Q$ populations. The label of each individual describes the population (or class) to which it belongs and is unobserved. The $Q$ classes have *a priori* distribution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$ with $\alpha_q = \mathbb{P}(i \in q)$. The hidden random indicator variables $(Z_{iq})_{i \in \mathcal{P}, q \in \mathcal{Q}}$ describe the label of each individuals, that is,

$$\alpha_q = \mathbb{P}(Z_{iq} = 1) = \mathbb{P}(i \in q), \qquad \text{such that} \sum_{q=1}^{Q} \alpha_q = 1. \tag{1}$$

Remark that we have $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iQ}) \sim \mathcal{M}(1, \boldsymbol{\alpha})$. The distribution of $X_i$ conditional on the label of $i$ is assumed to be a univariate gaussian distribution with unknown parameters, that is, $X_i | Z_{iq} = 1 \sim \mathcal{N}(\mu_q, \sigma_q^2)$.

## 2 Questions

1. *Likelihood.* Write the model complete-data loglikelihood.
2. *E-step.* For fixed values of $\hat{\mu}_q$, $\hat{\sigma}_q^2$ and $\hat{\alpha}_q$, give the expression of the estimates of the posterior probabilities $\tau_{iq} = \mathbb{P}(Z_{iq} = 1 | X_i)$.
3. *M-step.* For fixed values of $\hat{\tau}_{iq}$, show that the maximization step leads to the following estimator for the model parameters :

$$\hat{\alpha}_q = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{iq}, \quad \hat{\mu}_q = \frac{\sum_i \hat{\tau}_{iq} x_i}{\sum_i \hat{\tau}_{iq}}, \quad \hat{\sigma}_q^2 = \frac{\sum_i \hat{\tau}_{iq} (x_i - \hat{\mu}_q)^2}{\sum_i \hat{\tau}_{iq}} \tag{2}$$

4. *Implementation.* Test your EM algorithm on simulated data. Try different values for $\mu_q$, $\sigma_q$. Also consider different initialization.
5. *Optional.* Compute the ICL criterion and test it on your simulated data.

$$ICL(Q) = -2 \log L(X, \hat{Z}; \hat{\alpha}, \hat{\mu} \hat{\sigma^2}) + \log(n) \mathrm{df}(Q). \tag{3}$$