Exercise 1: Discrete distributions

Let $K \in \mathbb{N}^*$, $\mathcal{X} = \{v_1, \dots, v_K\}$ K real numbers. Let $(\pi_k)_{k \in [\![1, K]\!]}$ s.t

$$\forall k \in [\![1, K]\!], \quad \pi_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \pi_k = 1$$

1) We want to generate a random variable Y on $\mathcal{X}$ that follow the given distribution.

$$\forall k \in [\![1, K]\!], \qquad \mathbb{P}(Y = v_k) = \pi_k$$

So, if we denote Z a random variable taking value in $[\![1, K]\!]$ and s.t $\mathbb{P}(Z = k) = \pi_k$, we can write:

$$Y = \sum_{k=1}^{K} v_k \, \mathbb{1}_{\{Z = k\}} \quad \text{a.s (almost surely)}$$

We want to generate the random variable Z to be able to generate Y.

We will use the inverse function method to do so.

$\forall t \in [\![1, K]\!]$, $F_Z(t) := \mathbb{P}(Z \leq t) = \mathbb{P}(Z \leq \lfloor t \rfloor)$

$$= \sum_{k=1}^{\lfloor t \rfloor} \pi_k$$

and $\begin{cases} \forall t > K, \ F_Z(t) = 1 \\ \forall t < 1, \ F_Z(t) = 0 \end{cases}$

So, the inverse CDF is: $\forall y \in [0, 1]$,

$$F_Z^{-1}(y) := \min_{\substack{k \in [\![1, K]\!] \\ F_Z(k) \geq y}} k$$

Finally, if we sample $U \sim \mathcal{U}([0, 1])$, then:

$F_Z^{-1}(U)$ follows the same law as Z.

and $Y = \sum_{k=1}^{K} v_k \mathbb{1}_{\{Z = k\}} = \sum_{k=1}^{K} v_k \mathbb{1}_{\{F_Z(k-1) < U \leq F_Z(k)\}}$ a.s

with convention $F_Z(0) := 0$.

Exercise 2: GMM and the EM algorithm.

1) Let's denote $\alpha = (\alpha_j)_{j \in [\![1, m]\!]}$, $\mu = (\mu_j)_{j \in [\![1, m]\!]}$ and $\Sigma = (\Sigma_j)_{j \in [\![1, m]\!]}$ with $m \in \mathbb{N}^*$ fixed.

Then, $\Theta = (\alpha, \mu, \Sigma) \in \mathbb{R}^m \times \mathbb{R}^{md} \times \mathbb{R}^{md^2}$

with d the dimension of the Gaussians.

Let $(x_1, \dots, x_n)$ the observations of n iid samples $(X_1, \dots, X_n)$ of our model. We have:

$\forall \Theta$,

$$\mathcal{L}(x_1, \dots, x_n ; \Theta) := \prod_{i=1}^{n} f_\Theta(x_i) = \prod_{i=1}^{n} \sum_{j=1}^{m} f_\Theta(x_i, Z_i = j)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{m} f_\Theta(x_i | Z_i = j)\, \mathbb{P}(Z_i = j)$$

$$\boxed{= \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_j \frac{e^{-\frac{1}{2}(x_i - \mu_j) \Sigma_j^{-1} (x_i - \mu_j)^T}}{(2\pi)^{d/2} |\Sigma_j|^{1/2}}}$$

3) Our goal is to find θ maximizing the following log-likelihood:

$$l_{(x_1,\dots,x_n)}(\theta) := \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} \alpha_j f_\theta(x_i \mid z_i = j)\right)$$

This is hard to achieve, so we use EM.

First, the complete data log-likelihood is given by:

$$l\left((x_i)_{i\in[\![1,n]\!]}, (z_i)_{i\in[\![1,n]\!]}; \theta\right) = \log\left(\prod_{i=1}^{n} f(x_i, z_i \mid \theta)\right)$$

$$\overset{(x_i,z_i)_i \perp\!\!\!\perp}{=} \sum_{i=1}^{n} \log\left(f_\theta(x_i \mid z_i, \theta) f(z_i)\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log\left(f_\theta(x_i, z_i = j, \theta_j) \alpha_j\right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log\left(\mathcal{N}(x_i; \mu_j, \Sigma_j) \alpha_j\right)$$

So,

$$\mathbb{E}\left[l((x_i)_i, (z_i)_i; \theta) \mid \theta^{(c)}\right] = \sum_{i=1}^{n} \sum_{j=1}^{m} \underbrace{\mathbb{E}[z_{ij} \mid (x_i)_i, \theta^{(c)}]}_{:= \tau_{ij}} \log\left(\mathcal{N}(x_i; \mu_j^{(c)}, \Sigma_j^{(c)}) \alpha_j\right)$$

τ_{ij}: probability of observation i to be in cluster j given x_i and the parameters at step c.