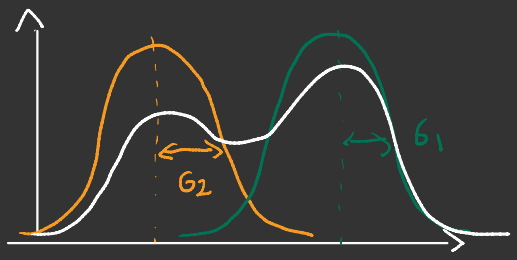
Modèles de Mélange fini

1894 . Karl Pearsen .



$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

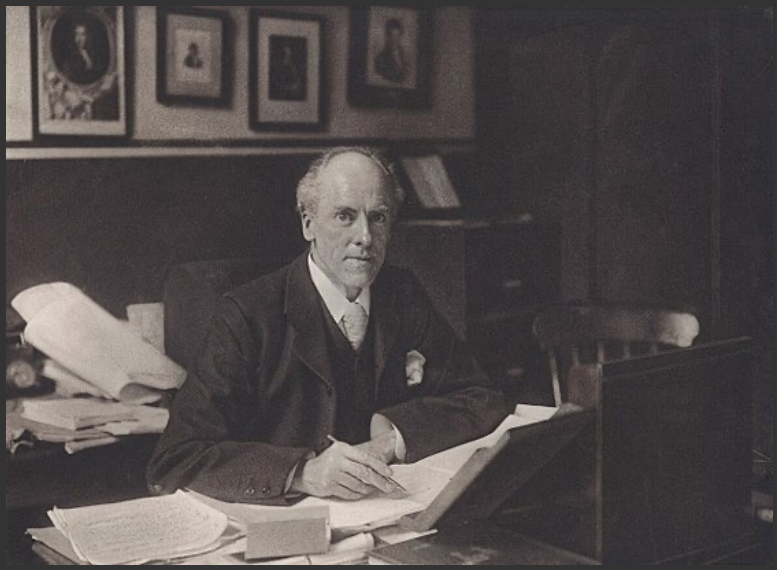
$$\forall k \quad f_{\theta_k}(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp -\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2$$

$$\pi_2 = 1 - \pi_1$$

$$\forall k \quad \pi_k \in]0,1[$$

$$\theta = \{\pi_1, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$$

KARL PEARSON 1894 (Fisher 1912).



KARL PEARSON

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

où les f_k sont des densités paramétriques.

III. "Contributions to the Mathematical Theory of Evolution." By KARL PEARSON, M.A., Professor of Applied Mathematics, University College. Communicated by Professor HENRICI, F.R.S. Received October 18, 1893.

(Abstract.)

1. If a series of measurements, physical, biological, anthropological, or economical, not of the same object, but of a group of objects of the same type or family, be made, and a curve be constructed by plotting up the number of times the measurements fall within a given small unit of range to the range, this curve may be termed a *frequency curve*. As a rule this frequency curve takes the well known form of the curve of errors, and such a curve may be termed a *normal frequency curve*. The latter curve is symmetrical about its maximum ordinate. Occasionally, however, frequency curves do not take the normal form, and are then generally, but not necessarily, asymmetrical. Such abnormal curves arise particularly in biological measurements; they have been found by Professor

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

Observées : $(x_1, \dots, x_n)$

Manquantes : $(z_1, \dots, z_n)$

INITIALISATION :

$$\theta^{(0)} = \left\{ \pi_1^{(0)}, \mu_1^{(0)}, \sigma_1^{2(0)}, \mu_2^{(0)}, \sigma_2^{2(0)} \right\}$$

$\forall k \quad \pi_k^{(0)} = \frac{1}{K}$ ici $\pi_1 = \frac{1}{2}$

$\forall k \quad \mu_k^{(0)} = x_i$ où $i \sim \mathcal{U}\{1, \dots, n\}$

$\forall k \quad \sigma_k = 1$

ITERATION :

$$E : \quad Q(\theta, \theta^q) = \mathbb{E}_{Z|X,\theta^q}\left[\log P(X, Z; \theta) \right]$$

$$M : \quad \theta^{q+1} = \operatorname*{argmax}_{\theta} Q(\theta, \theta^q)$$

$$P(x_1, \dots, x_n, z_1, \dots, z_n; \theta) = \prod_{i=1}^{n} P(x_i, z_i; \theta)$$

$$= \prod_{i=1}^{n} \left(\prod_{k=1}^{K} P(x_i, z_i)^{\Delta\{z_i = k\}} \right)$$

$$\log P(x,z;\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \log P(x_i, z_i=k)$$

$$\log P(z_i=k;\theta)\, P(x_i \mid z_i=k;\theta)$$

$$\pi_k \qquad f_k(x_i;\theta)$$

$$\mathbb{E}_{z|x;\theta^{(q)}}\left[\log P(x,z;\theta)\right] = \sum_i \sum_k \mathbb{E}\left[\mathbb{1}_{\{z_i=k\}}\right] \log \pi_k \, f_k(x_i;\theta)$$

$$P(z_i=k \mid x_i;\theta^{(q)}) = t_{ik}^{(q)}$$

$$T = \begin{bmatrix} \vdots & & \vdots \\ \cdots & t_{ik}^{(q)} & \cdots \\ \vdots & & \vdots \end{bmatrix} \quad \begin{matrix} 1 \\ \vdots \\ n \end{matrix} \qquad \begin{matrix} 1 & \cdots & K \end{matrix}$$

$$\forall i \quad \sum_k t_{ik}^{(q)} = 1$$

$$M: \quad \frac{\partial Q(\theta,\theta^{(q)})}{\partial \mu_k} = 0$$

$$0 = \frac{\partial \left(\sum_i \sum_k t_{ik}^{q} \left(\log \pi_k + \log \left(\frac{1}{\sqrt{2\pi}\,\sigma_k} \exp -\frac{1}{2} \frac{(x_i-\mu_k)^2}{\sigma_k^2} \right) \right) \right)}{\partial \mu_k}$$

$$0 = \frac{\partial \sum_{i=1}^{n} t_{ik}^{(q)} \left(-\frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma_k^2}\right)}{\partial \mu_k} \Rightarrow \sum_{i=1}^{n} t_{ik}^{(q)} (x_i - \mu_k) \cdot (-1) = 0$$

$$\sum_{i=1}^{n} t_{ik}^{(q)} x_i = \sum t_{ik}^{(q)} \mu_k$$

$$\mu_k^{q+1} = \frac{\sum_{i=1}^{n} t_{ik}^{(q)} x_i}{\sum_{i=1}^{n} t_{ik}^{(q)}}$$

$$\frac{\partial Q(\theta, \theta^q)}{\partial \sigma_k^2} = 0 \Rightarrow$$

$$\frac{\partial \left(\sum_{i=1}^{n} t_{ik}^{q} \log \frac{1}{\sqrt{2\pi}\sigma_k} \exp -\frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma_k^2} \right)}{\partial \sigma_k^2} = 0$$

$$\frac{\partial \left(\sum_i t_{ik}^{(q)} \left(-\frac{1}{2}\log \sigma_k^2\right) \right)}{\partial \sigma_k^2} + \frac{\partial}{\partial \sigma_k^2}\left(\sum_i t_{ik}^{(q)} \left(-\frac{1}{2}\frac{(x_i-\mu_k)^2}{\sigma_k^2}\right) \right) = 0$$

$$-\frac{1}{2}\sum_i t_{ik}^{q} \times \frac{1}{\sigma_k^2} - \frac{1}{2}\sum_i t_{ik}^{(q)} \left(-\frac{1}{(\sigma_k^2)^2}(x_i-\mu_k)^2\right) = 0$$

$$\sigma_k^{2\,q+1} = \frac{\sum_{i=1}^{n} t_{ik}^{q} \left(x_i - \mu_k^{q+1} \right)^2}{\sum_{i=1}^{n} t_{ik}^{q}}$$

$$\sum_{k=1}^{K} \pi_k = 1 \qquad \pi_1 + \pi_2 = 1$$

$$\mathcal{L}(\theta, \lambda) = Q(\theta, \theta^q) + \lambda \left(\sum_k \pi_k - 1 \right)$$

$$\begin{cases} \dfrac{\partial \mathcal{L}(\theta, \lambda)}{\partial \theta} = 0 \\ \dfrac{\partial \mathcal{L}(\theta, \lambda)}{\partial \lambda} = 0 \end{cases} \quad \begin{cases} \dfrac{\partial \left(\sum_{i=1}^{n} t_{ik}^{q} \log \pi_k \right)}{\partial \pi_k} + \dfrac{\partial \left(\lambda \sum \pi_k - 1 \right)}{\partial \pi_k} = 0 \\ \sum \pi_k - 1 = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} t_{ik}^{(q)} \cdot \dfrac{1}{\pi_k} - \lambda = 0 \\ \sum \pi_k - 1 = 0 \end{cases} \quad \begin{cases} \pi_k^{(q+1)} = \dfrac{\sum_{i=1}^{n} t_{ik}^{q}}{\lambda} \\ \sum \pi_k = 1 \end{cases}$$

$$\lambda = \sum_{i=1}^{n} \left(\sum_{k=1}^{K} t_{ik}^{q} \right) = \sum_{i=1}^{n} 1 = n$$

$$\pi_k^{q+1} = \frac{\sum_{i=1}^{n} t_{ik}^{q}}{n}$$

CONSEILS DE PROGRAMMATION

$$P(X=x) = \sum_{k=1}^{K} \pi_k \, P_k(X=x)$$

$$P_k(X=x) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad \mathbb{E}[X \mid Z=k] = \lambda_k$$

$$\text{Init}\left(\underbrace{x_1, \dots, x_n}_{\underline{X}}, K\right) \longrightarrow \theta^{(0)} = \left\{\pi_1^{(0)}, \dots, \pi_{K-1}^{(0)}, \lambda_1^{(0)}, \dots \lambda_K^{(0)}\right\}$$

$$\begin{cases} \text{E\_step}\left(\theta^{(q)}, x_1, \dots, x_n\right) \longrightarrow T^{(q)} = \left(t_{ik}\right)_{\substack{i=1..n \\ k=1..K}} \\ \text{M\_step}\left(T, x_1, \dots, x_n\right) \longrightarrow \theta^{(q+1)} \end{cases}$$

CRITÈRE DE CONVERGENCE

$$\frac{\left\| \theta^q - \theta^{q+1} \right\|_2^2}{\left\| \theta^q \right\|_2^2} < \varepsilon \quad \left(\text{ex } \varepsilon = 10^{-6}\right)$$

ALGORITHME EM POUR LES MODÈLES DE MÉLANGE FINI

Dempster, Laird, Rubin (1977)

OBSERVATIONS $(x_1, \dots, x_n)$

MANQUANTES $(z_1, \dots, z_n)$

COMPLÈTES $((x_1, z_1), \dots, (x_n, z_n))$

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$$

π_k : proportion ; f_k : densité de composante

$$p(\theta ; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k f_k(x_i) \right)$$

$$\log p(\theta ; \underbrace{x_1, \dots, x_n}_{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k(x_i) \right)$$

$$P((x_1, z_1), \dots (x_n, z_n)) = P(X, Z) = P(X) \times P(Z|X)$$

$$\log \underbrace{p(\theta ; X)}_{P(X)} = \log P(X, Z ; \theta) - \log P(Z|X ; \theta)$$

$$\underbrace{\mathbb{E}_{Z|X;\vartheta^q}\left[\log P(X;\vartheta)\right]}_{\log P(X;\vartheta) = Q(\vartheta,\vartheta^q) - H(\vartheta,\vartheta^q)} = \mathbb{E}_{Z|X;\vartheta^q}\left[\log P(X,Z;\vartheta)\right] - \mathbb{E}_{Z|X;\vartheta^q}\left[\log P(Z|X;\vartheta)\right]$$

$Q(\vartheta,\vartheta^q)$

$H(\vartheta,\vartheta^q)$

INIT : $\vartheta^{(0)}$

REPETER JUSQU'À CONV. [E(xpectation) : CALCUL DE $Q(\vartheta,\vartheta^q)$; M(aximisation) : $\vartheta^{q+1} = \underset{\vartheta}{\operatorname{argmax}}\, Q(\vartheta,\vartheta^q)$]

MONTRONS QUE CHAQUE ITERATION DE EM AUGMENTE $\ell(\theta ; X)$

$$Q(\theta^{q+1}, \theta^q) \geq Q(\theta^q, \theta^q)$$

$$\times \quad \mathbb{E}_{Z|X;\theta^q}\left[\log P(X,Z;\theta^{q+1}) - \log P(X,Z;\theta^q) \right] \geq 0$$

$$\int P(Z|X;\theta^q) \log \frac{P(X,Z;\theta^{q+1})}{P(X,Z;\theta^q)} \, dZ \geq 0$$

$$\log \mathbb{E}_{Z|X;\theta^q}\left[\frac{P(X,Z;\theta^{q+1})}{P(X,Z;\theta^q)} \right] \geq 0$$

JENSEN

$$\log \int P(Z|X;\theta^q) \frac{P(X,Z;\theta^{q+1})}{P(X,Z;\theta^q)} \, dZ \geq 0$$

$$\log \int P(Z|X;\theta^q) \frac{P(X;\theta^{q+1})}{P(X;\theta^q)} \frac{P(Z|X;\theta^{q+1})}{P(Z|X;\theta^q)} \, dZ \geq 0$$

$$\log \frac{P(X;\theta^{q+1})}{P(X;\theta^q)} \int P(Z|X;\theta^{q+1}) \, dZ \geq 0 \iff \frac{P(X;\theta^{q+1})}{P(X;\theta^q)} \geq 1$$