

Semaine 4 : Apprentissage non-supervisé et réduction de dimension

5 : Maximum de vraisemblance

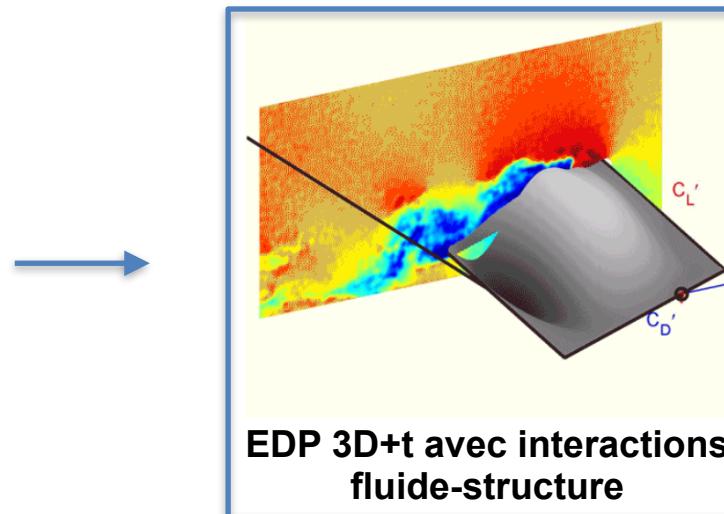
Laurent Risser

Ingénieur de Recherche CNRS
Institut de Mathématiques de Toulouse / 3IA ANITI

lrisser@math.univ-toulouse.fr

1) Notations essentielles en probabilités

Variable aléatoire Une *variable aléatoire* (v.a.) X est une application définie sur l'ensemble des résultats possibles d'une expérience aléatoire. Dans le cadre de ce cours ses résultats possibles seront toujours dans \mathbb{R} ou un sous-ensemble de \mathbb{R} . On distinguera en particulier le *cas continu*, par exemple si X représente l'incertitude sur une estimation de la température et le *cas discret*, par exemple $X \in \{0, 1\}$ pour modéliser le résultat lorsque l'on joue à pile ou face.



ou

$$\mathbb{P}(X = \text{pile}) = p$$

(avec p estimé à partir de n tirages)

1) Notations essentielles en probabilités

Loi de probabilité La *loi de probabilité* d'une v.a. décrit la probabilité d'obtenir les différents résultats de cette variable.

Loi de probabilité discrète Par exemple si l'on joue à pile ou face avec une pièce parfaitement équilibrée, on a $\mathbb{P}(X = 0) = 1 - p = 0.5$ et $\mathbb{P}(X = 1) = p = 0.5$. On remarquera que la somme des probabilités de tous les résultats possibles dans le cas discret est toujours 1.

Loi de probabilité continue Dans le cas continu, écrire $\mathbb{P}(X = x)$ n'a aucun sens puisque la probabilité d'une valeur exacte est infinitésimale. On pourra par contre utiliser la *fonction de répartition* $F_X(x) = \mathbb{P}(X \leq x)$ pour représenter comment se répartissent les probabilités des différents résultats de X . Il sera alors possible de quantifier les chances que X soit sur une certaine gamme de valeurs $\mathbb{P}(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$. Naturellement, on aura toujours $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$. De manière purement équivalente à la fonction de répartition $p_X(x)$, la *densité de probabilité* pourra de même représenter la loi de probabilité d'une v.a. X suivant :

$$p_X(x) = \frac{\partial F_X}{\partial x}(x)$$

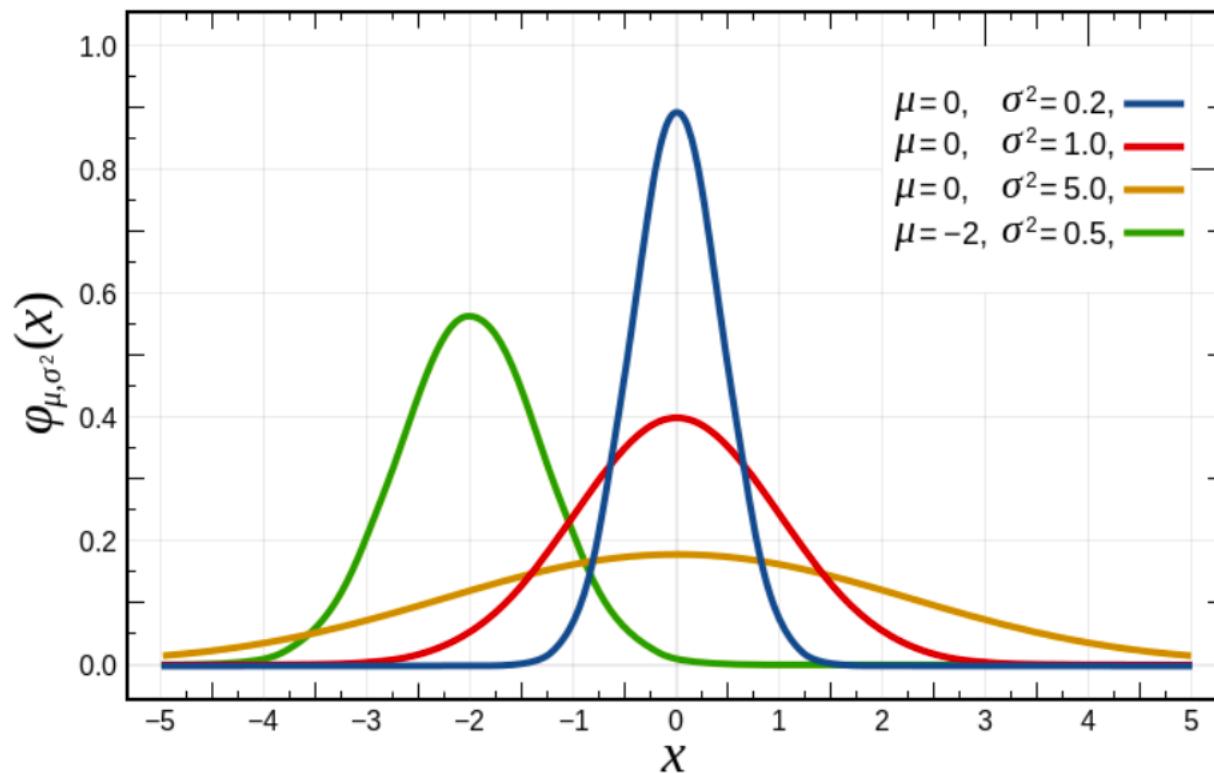
En utilisant les densités de probabilités, les chances que X tombe sur une gamme de valeurs $[x_1, x_2]$ sera alors

$$\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x)dx.$$

1) Notations essentielles en probabilités

La densité de probabilité de la loi normale de moyenne μ et d'écart type σ s'écrit :

$$\varphi_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

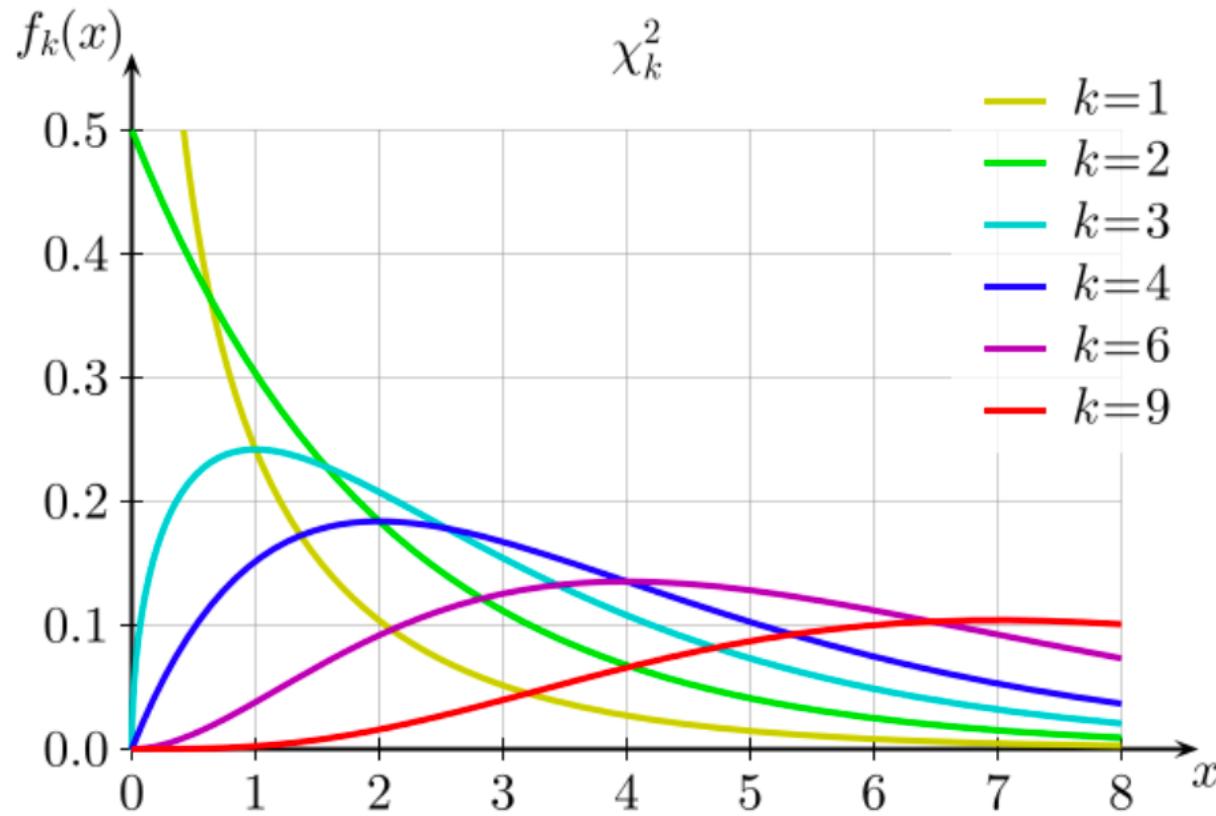


Si la variable aléatoire X suit une loi normale de moyenne μ et d'écart type σ , on écrit $X \sim \mathcal{N}(\mu, \sigma^2)$.

1) Notations essentielles en probabilités

La densité de probabilité de la loi du χ^2 à k degrés de libertés :

$$f_k(x) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2}$$



Si la variable aléatoire X suit une loi du χ^2 à k degrés de libertés, on écrit
 $X \sim \chi^2(k)$

Supposons que n variables aléatoires X_1, X_2, \dots, X_n indépendantes mais suivant une même loi de probabilité soient tirées. L'espérance (ou moyenne) m et l'écart type s de leur loi est connue. Le nombre d'observations n est aussi supposé grand (typiquement $n > 30$). Alors, la somme des X_i peut être approchée par une loi normal de moyenne nm et d'écart type $s\sqrt{n}$, i.e. :

$$\sum_{i=1}^n X_i \sim \mathcal{N}(nm, s^2 n),$$

où la *densité de probabilité* de la loi normale $\mathcal{N}(\mu, \sigma^2)$ est (voir aussi appendice A) :

$$f_{\theta=\{\mu, \sigma\}}(X_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

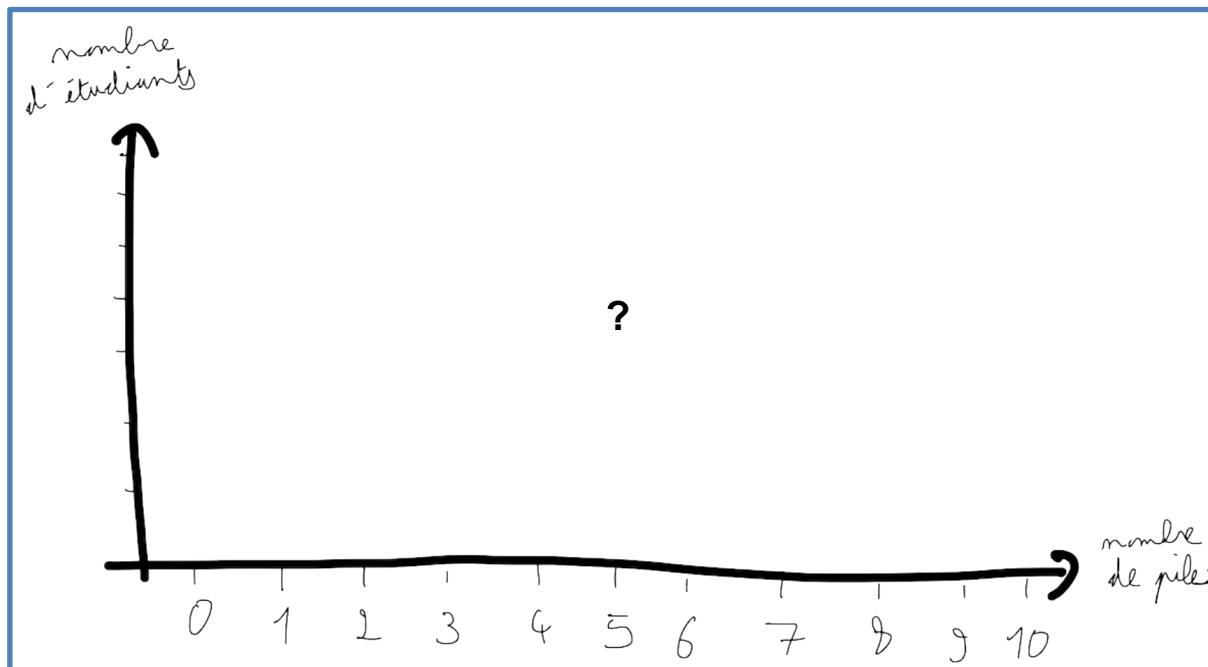
Exemple :

- $X \sim \text{Bern}(0.5)$, c'est à dire qu'on a une probabilité de 0.5 que chaque X_i tiré vaille 1 et une probabilité de 0.5 qu'il vaille 0 (comme à pile ou face).
- La somme de 100 tirages sera alors approchée par : $\sum_{i=1}^{100} X_i \sim \mathcal{N}(50, 25)$

1) Notations essentielles en probabilités - Illustration avec le Théorème Central Limite

Vérifions empiriquement cette loi

- Chaque étudiant de la classe tire $n = 10$ fois une pièce à pile ou face avec et compte le nombre de fois que la pièce est tombée sur pile. Pile correspond alors à $X_i = 1$ et face à $X_i = 0$.
- On suppose que $\mathbb{P}(X = 1) = 0.5$ et $\mathbb{P}(X = 0) = 0.5$, ce qui est sans doute très proche de la réalité. Ainsi l'espérance (moyenne) de X est $m = 0.5$ et son écart type est $s = 0.5$.

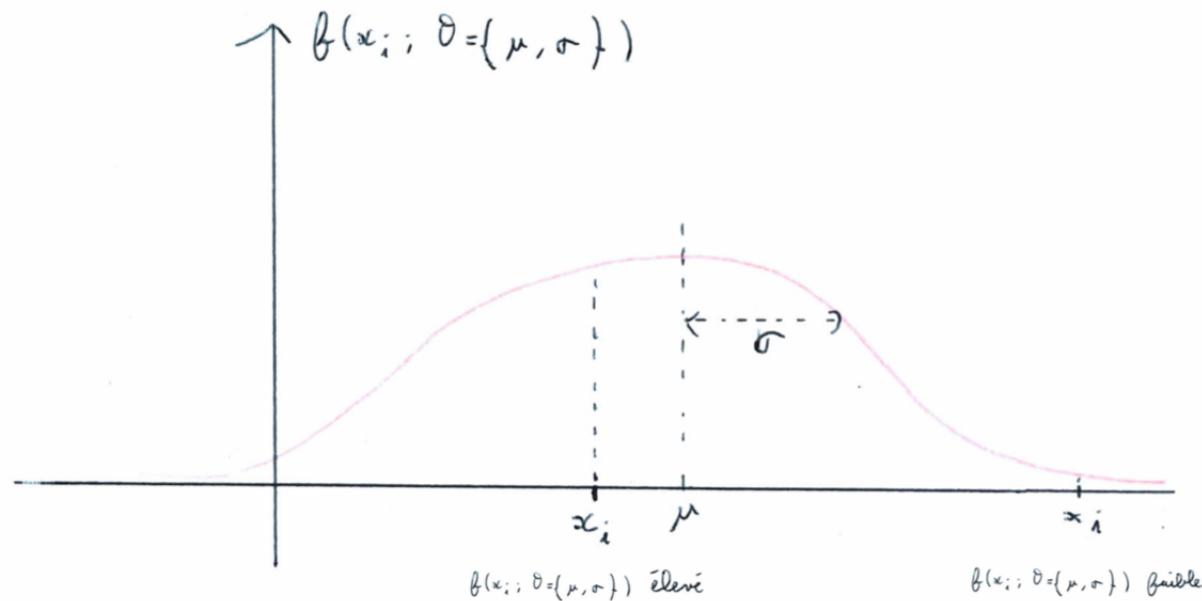


2) Maximum de vraisemblance

On dénote X une variable aléatoire (v.a.) supposée suivre une loi discrète (e.g. Bernoulli) ou continue (e.g. Normale) de paramètres θ . On note aussi $x_1, \dots, x_i, \dots, x_n$ les observations de X .

Pour une observation x_i donnée, on modélise alors la loi de X avec la fonction $f(x_i; \theta)$. Cette fonction vaut $f(x_i; \theta) = \mathbb{P}_\theta(X = x_i)$ si X est une v.a. discrète et $f(x_i; \theta) = f_\theta(x_i)$ si X est continue, où $f_\theta(x_i)$ est la densité de la loi en fonction de ses paramètres θ .

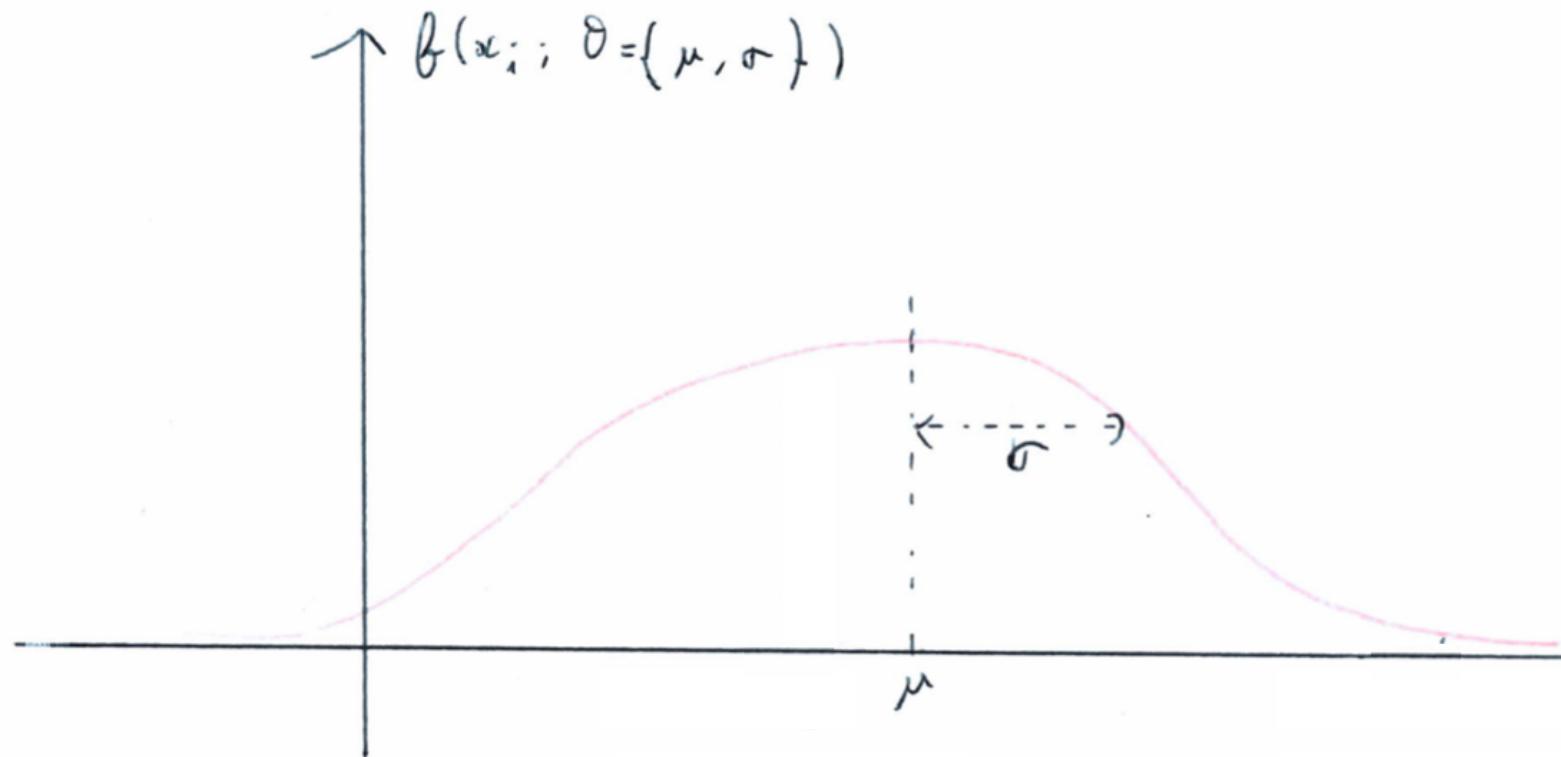
Pour des paramètres θ donnés (ex : moyenne et écart type d'une loi normale), $f(x_i; \theta)$ sera alors d'autant plus élevée que x_i a des chances d'être tirée en fonction des θ .



2) Maximum de vraisemblance

La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

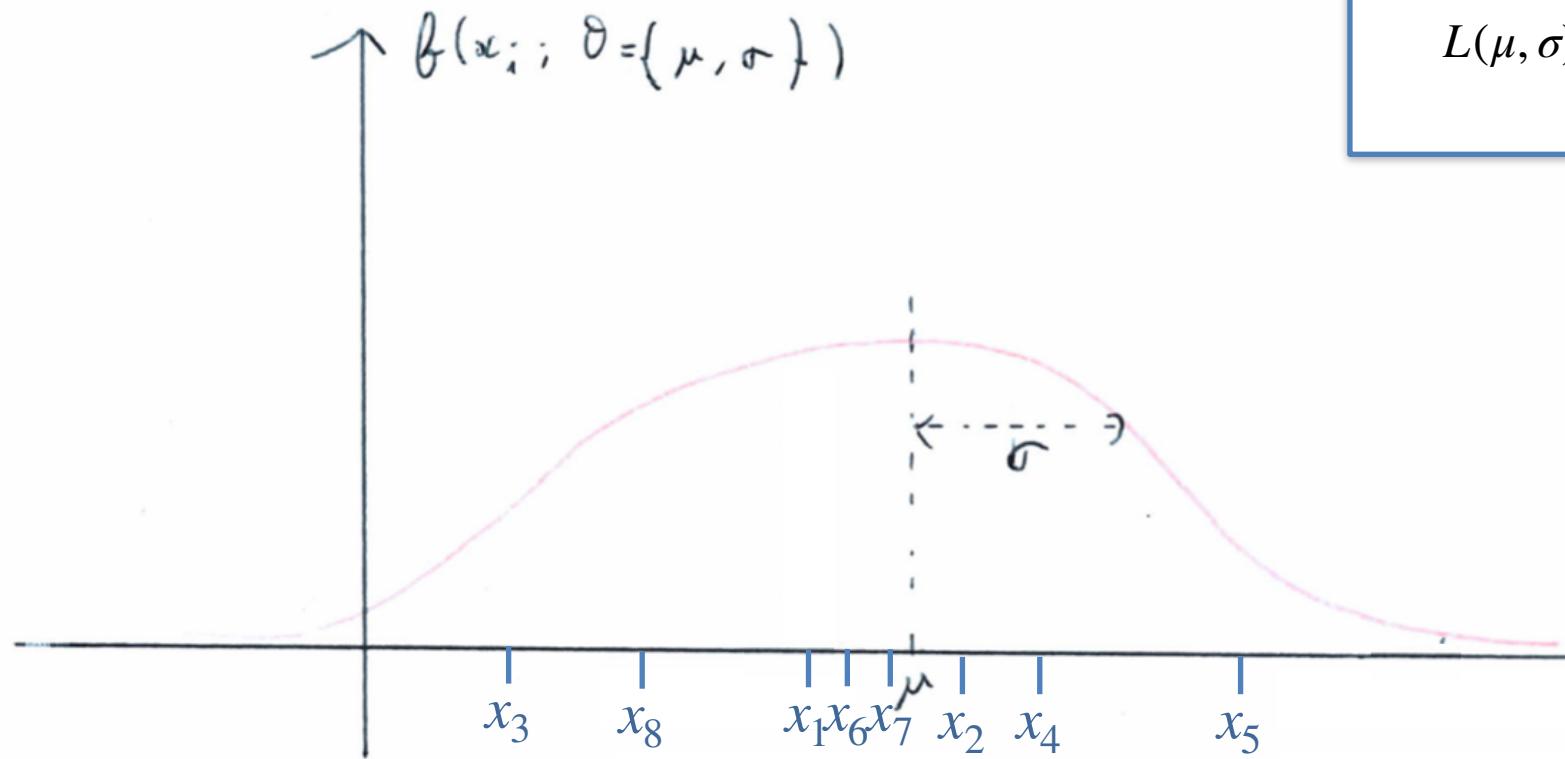
$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



2) Maximum de vraisemblance

La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

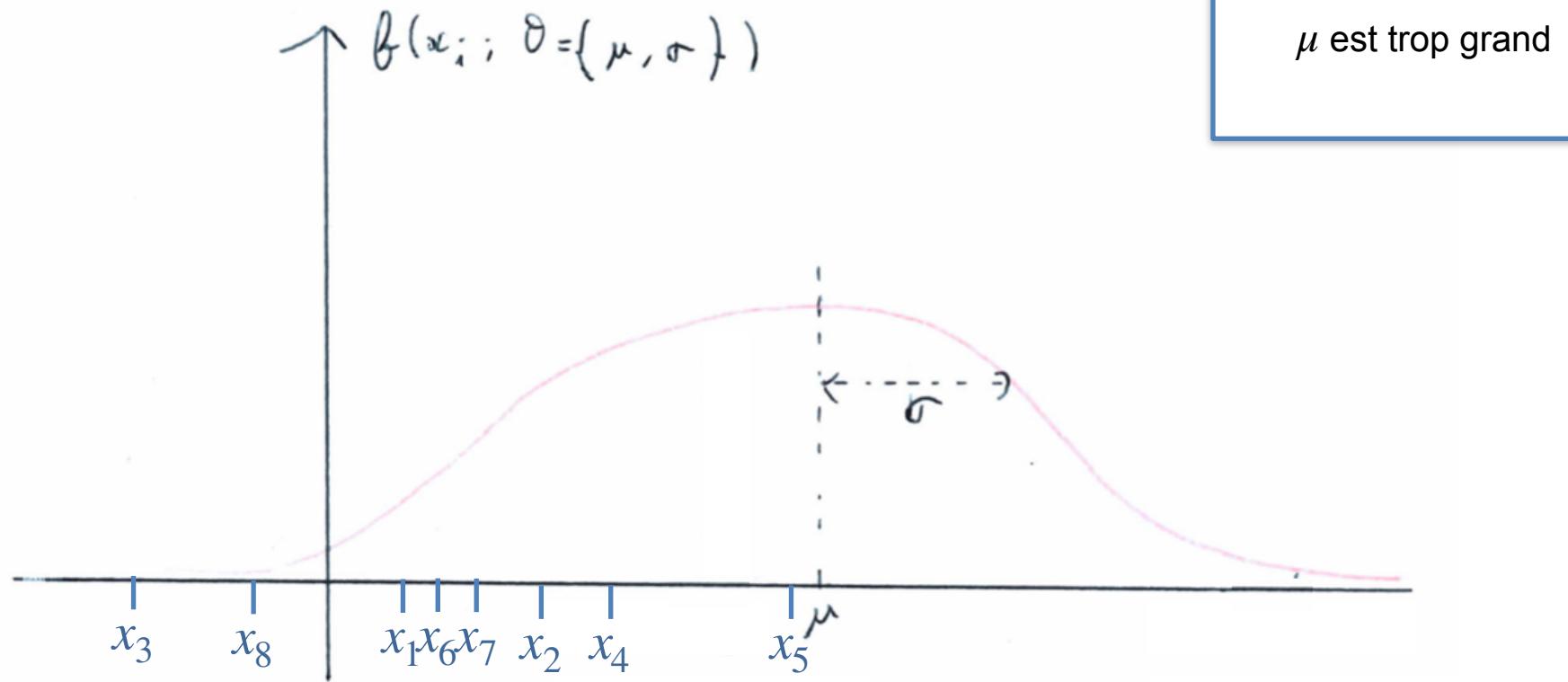
$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



2) Maximum de vraisemblance

La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

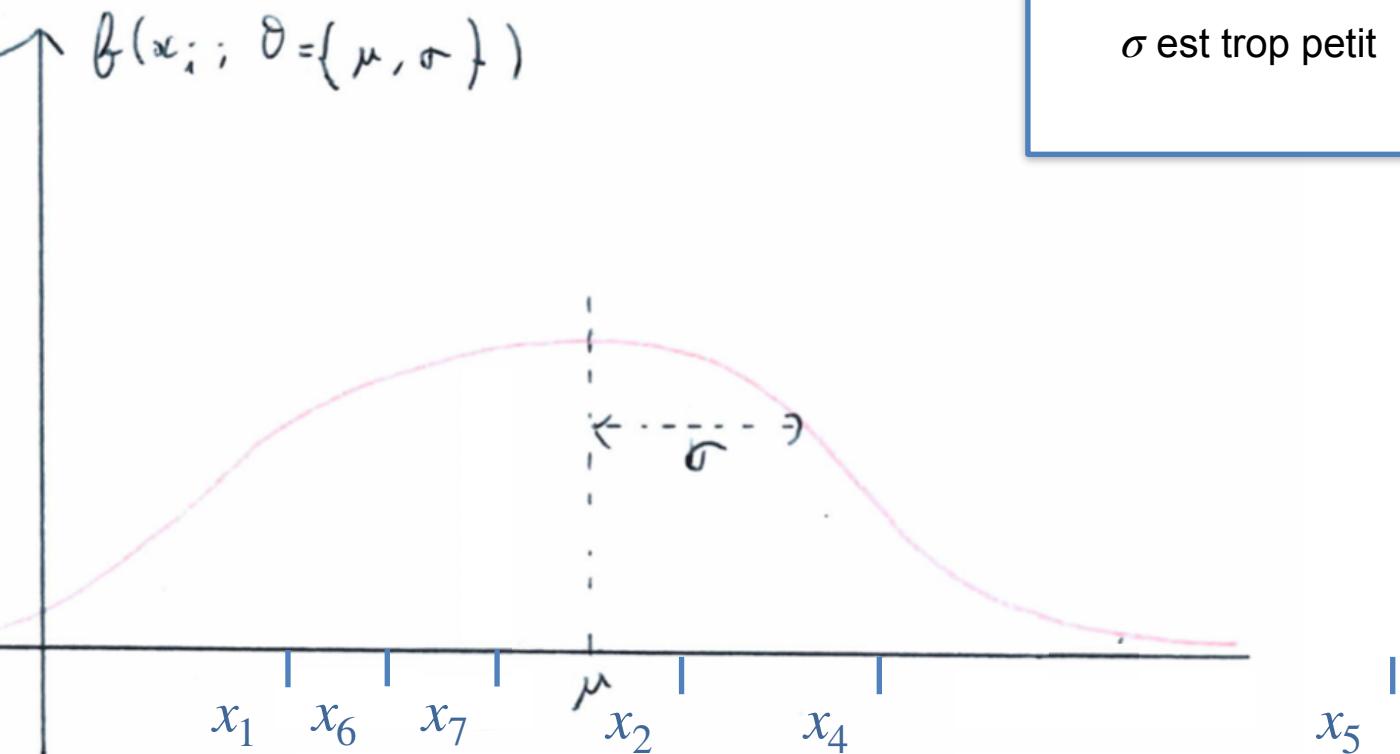
$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



2) Maximum de vraisemblance

La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

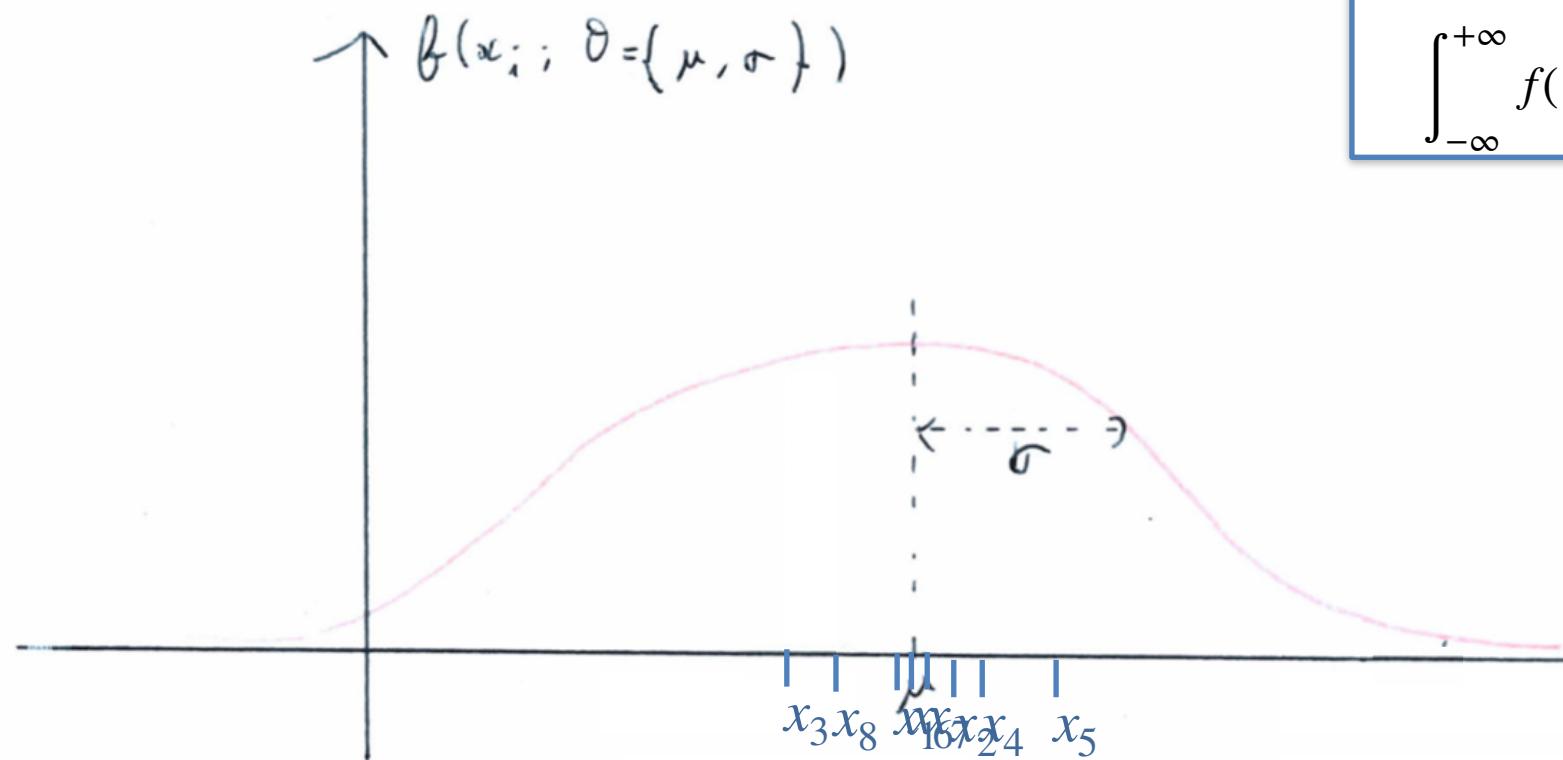
$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



2) Maximum de vraisemblance

La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$



2) Maximum de vraisemblance

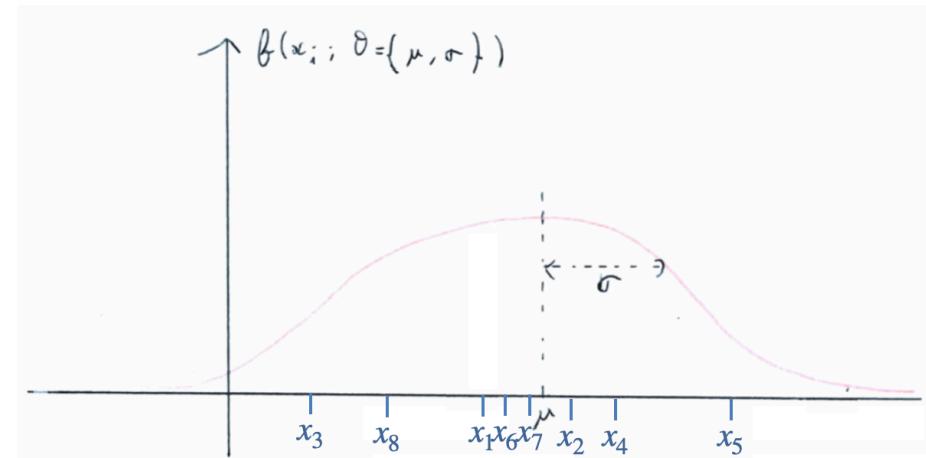
La vraisemblance des paramètres θ en fonction des observations x_1, \dots, x_n est alors :

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

Afin de trouver les paramètres d'une loi les plus vraisemblables, une fois les $\{x_i\}_{i=1,\dots,n}$ connus, on calculera le *maximum de vraisemblance* :

$$\hat{\theta} = \arg \max_{\theta} L(\theta),$$

qui renverra les paramètres les plus vraisemblables en fonction des observations et de la loi choisie.



2) Maximum de vraisemblance

Pour des raisons numériques, il est aussi bien pratique de maximiser la log-vraisemblance au lieu de la vraisemblance brute :

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \log(L(\theta)) \\ &= \arg \max_{\theta} \log \left(\prod_{i=1}^n f(x_i; \theta) \right) \\ &= \arg \max_{\theta} \sum_{i=1}^n \log f(x_i; \theta)\end{aligned}$$

Vu que la fonction \log est strictement croissante les paramètres optimum $\hat{\theta}$ seront les mêmes avec la log-vraisemblance ou la vraisemblance.

2) Maximum de vraisemblance

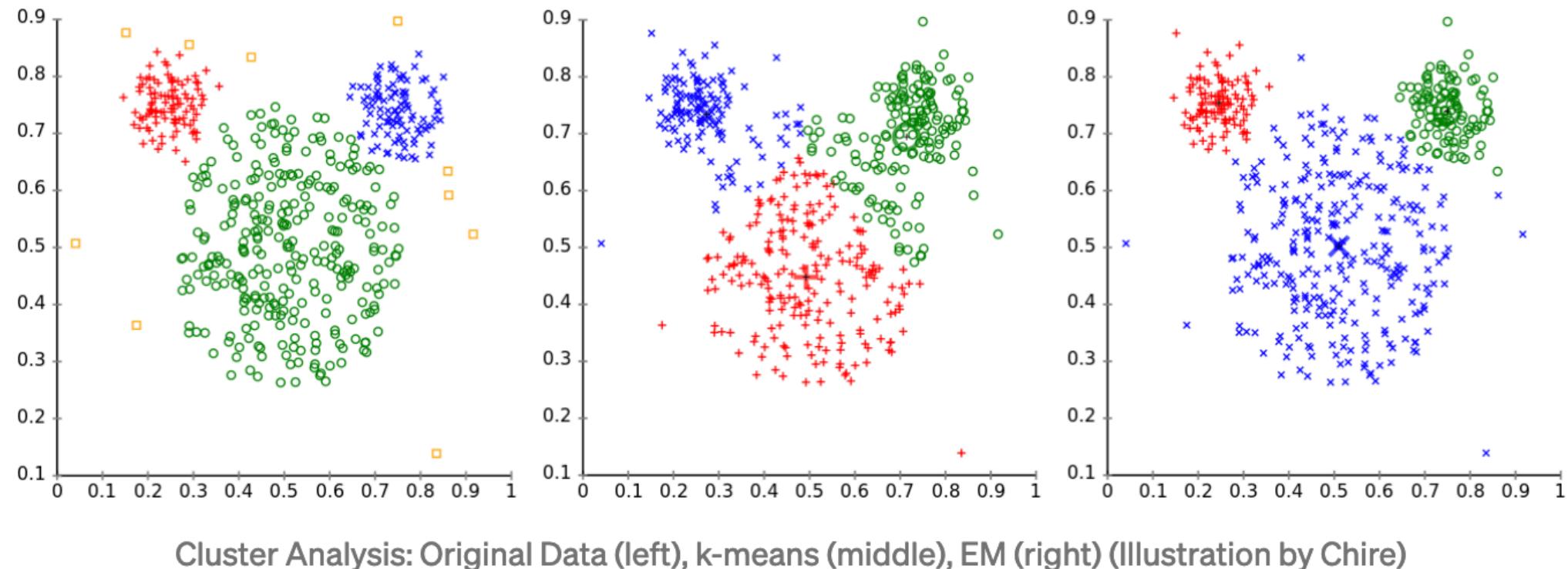
Remarque à propos de l'apprentissage supervisé où on connaît $\{x_i, y_i\}_{i=1,\dots,n}$ et où on cherche les paramètres θ qui minimisent $\sum_{i=1}^n (f_\theta(x_i) - y_i)^2$:

Faisons l'hypothèse que les erreurs d'approximation du modèle $e_i = y_i - f_\theta(x_i)$ suivent une loi normale centrée, i.e. $e_i \sim \mathcal{N}(0, \sigma)$. Ce choix par défaut est commun et semble raisonnable quand f_θ est bien calibré. Nous pouvons alors utiliser le principe de maximum de vraisemblance pour estimer les paramètres θ du modèle f_θ .

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{e_i^2}{2\sigma^2}\right) \\ &= \arg \max_{\theta} \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n e_i^2\right) \\ &= \arg \min_{\theta} \sum_{i=1}^n e_i^2 \\ &= \arg \min_{\theta} \sum_{i=1}^n (y_i - f_\theta(x_i))^2\end{aligned}$$

3) Modèle de mélange de Gaussiennes

On a vu précédemment que l'utilisation de la vraisemblance peut être pertinente pour classifier des données échantillonnées suivant des loi et/ou des paramètres différents !



On va :

- Attribuer une densité de probabilité $\mathcal{N}(\mu_k, \Sigma_k)$ sur chaque groupe $k = 1, \dots, K$.
- Estimer à partir des données les moyennes μ_k et covariances Σ_k de chaque groupe k .
- Attribuer à chaque observation son label le plus pertinent en fonction des $\{\mu_k, \Sigma_k\}_{k=1, \dots, K}$.

3) Modèle de mélange de Gaussiennes

Partie basée sur le cours de F. Santos (U. Bordeaux) sur E.M.

Définition (Densité de mélange). — On appelle densité mélange, ou loi mélange, une fonction de densité qui est une combinaison linéaire convexe de plusieurs fonctions de densité. Autrement dit, f est une densité mélange s'il existe $K \in \mathbb{N}$, des densités f_1, \dots, f_K et des réels p_1, \dots, p_K sommant à 1, tels que :

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

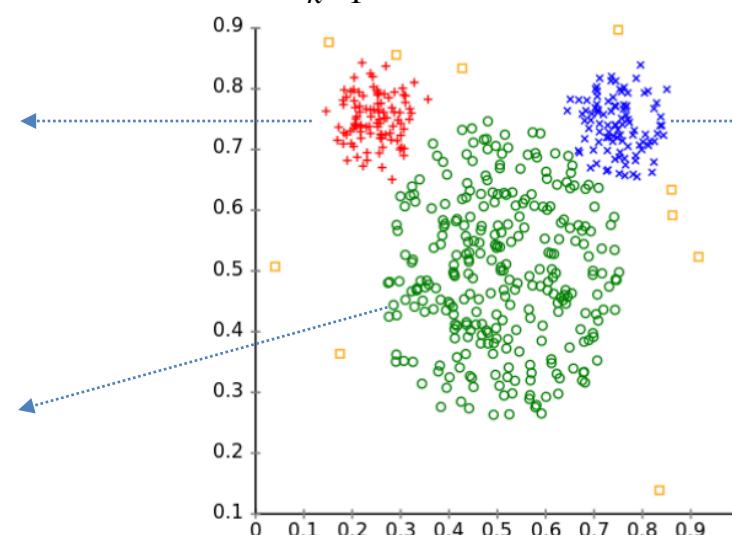
En écrivant les p_k comme des $\mathbb{P}\{Z=k\}$ avec Z variable aléatoire discrète à valeurs dans $\llbracket 1, K \rrbracket$, on a alors :

$$f(x) = \sum_{k=1}^K \mathbb{1}_{\{Z=k\}} \mathbb{P}(Z=k) f_k(x)$$

- $Z = 2$
- $f_2 \rightarrow$ densité $\mathcal{N}(\mu_2, \Sigma_2)$

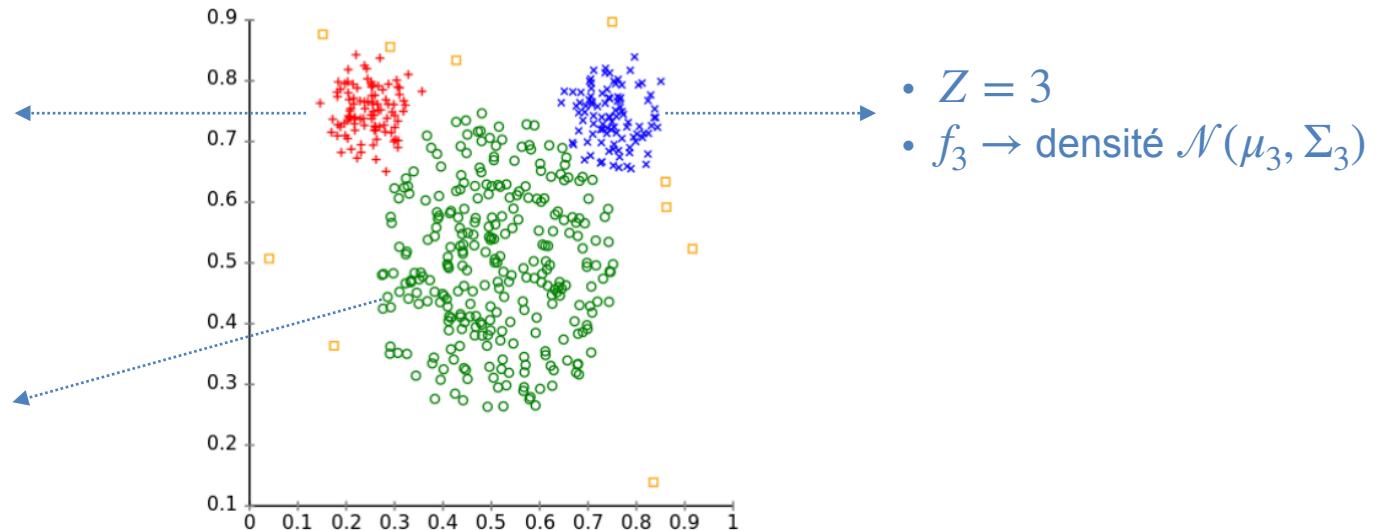
- $Z = 1$
- $f_1 \rightarrow$ densité $\mathcal{N}(\mu_1, \Sigma_1)$

- $Z = 3$
- $f_3 \rightarrow$ densité $\mathcal{N}(\mu_3, \Sigma_3)$



3) Modèle de mélange de Gaussiennes

- $Z = 2$
- $f_2 \rightarrow \text{densité } \mathcal{N}(\mu_2, \Sigma_2)$



- $Z = 1$
- $f_1 \rightarrow \text{densité } \mathcal{N}(\mu_1, \Sigma_1)$

Optimisation de la vraisemblance : $L_n(X, Z | \theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \left[\sum_{k=1}^K 1_{\{Z_i=k\}} \mathbb{P}(Z = k) f_k(x) \right]$

Paramètres :

$$\rightarrow \mu_1 = (\mu_1^1, \mu_1^2) , \quad \Sigma_1 = [[\sigma_1^{1,1}, \sigma_1^{1,2}], [\sigma_1^{2,1}, \sigma_1^{2,2}]] , \quad p_1$$

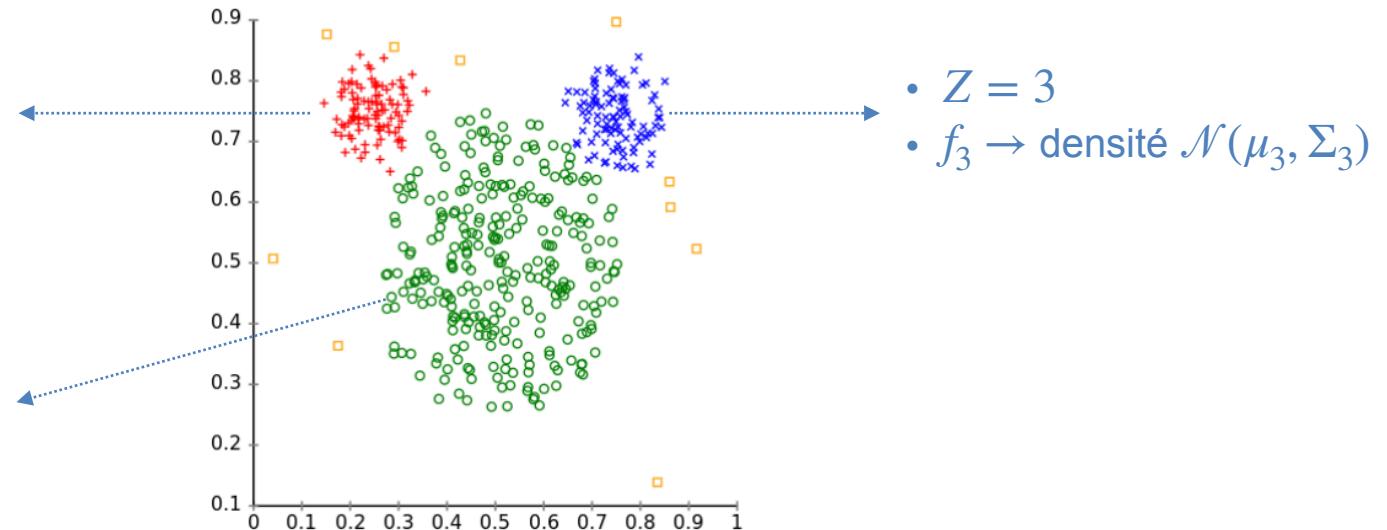
$$\rightarrow \mu_2 = (\mu_2^1, \mu_2^2) , \quad \Sigma_2 = [[\sigma_2^{1,1}, \sigma_2^{1,2}], [\sigma_2^{2,1}, \sigma_2^{2,2}]] , \quad p_2$$

$$\rightarrow \mu_3 = (\mu_3^1, \mu_3^2) , \quad \Sigma_3 = [[\sigma_3^{1,1}, \sigma_3^{1,2}], [\sigma_3^{2,1}, \sigma_3^{2,2}]] , \quad p_3$$

3) Modèle de mélange de Gaussiennes

- $Z = 2$
- $f_2 \rightarrow \text{densité } \mathcal{N}(\mu_2, \Sigma_2)$

- $Z = 1$
- $f_1 \rightarrow \text{densité } \mathcal{N}(\mu_1, \Sigma_1)$



Optimisation de la vraisemblance : $L_n(X, Z | \theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \left[\sum_{k=1}^K 1_{\{Z_i=k\}} \mathbb{P}(Z = k) f_k(x) \right]$

Paramètres :

$$\rightarrow \mu_1 = (\mu_1^1, \mu_1^2) , \quad \Sigma_1 = [[\sigma_1^{1,1}, \sigma_1^{1,2}], [\sigma_1^{2,1}, \sigma_1^{2,2}]] , \quad p_1$$

$$\rightarrow \mu_2 = (\mu_2^1, \mu_2^2) , \quad \Sigma_2 = [[\sigma_2^{1,1}, \sigma_2^{1,2}], [\sigma_2^{2,1}, \sigma_2^{2,2}]] , \quad p_2$$

$$\rightarrow \mu_3 = (\mu_3^1, \mu_3^2) , \quad \Sigma_3 = [[\sigma_3^{1,1}, \sigma_3^{1,2}], [\sigma_3^{2,1}, \sigma_3^{2,2}]] , \quad p_3$$

... mais aussi $1_{\{Z_i=k\}}$, c'est à dire le label k attribué à chaque observation X_i , $i = 1, \dots, n$

3) Algorithme E.M.

L'algorithme EM tire son nom du fait qu'à chaque itération il opère deux étapes distinctes :

- la phase « Expectation », souvent désignée comme « l'étape E », procède comme son nom le laisse supposer à l'estimation des données inconnues, sachant les données observées et la valeur des paramètres déterminée à l'itération précédente ;
- la phase « Maximisation », ou « étape M », procède donc à la maximisation de la vraisemblance, rendue désormais possible en utilisant l'estimation des données inconnues effectuée à l'étape précédente, et met à jour la valeur du ou des paramètre(s) pour la prochaine itération.

Dans le problème du mélange de Gaussiennes nous aurons :

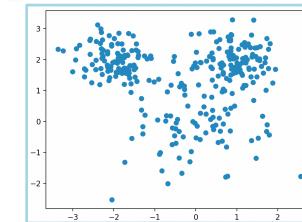
- Expectation : Estimation des probabilités d'avoir le label Z_i attribués à chaque X_i en fonction de l'approximation courante des $\theta = (\mu_1^1, \dots, p_1, \mu_2^1, \dots, p_2, \mu_3^1, \dots, p_3)$.
- Maximisation : Optimisation du maximum de vraisemblance des $(\mu_1^1, \dots, \dots, \dots, p_3)$ avec les probabilités de labels $Z = (Z_1, \dots, Z_n)$ estimés.

3) Algorithme E.M.

Plus formellement

- nous disposons d'observations i.i.d. $\mathbf{X} = (X_1, \dots, X_n)$ de vraisemblance notée $P(\mathbf{X}|\theta)$;
- maximiser $\log P(\mathbf{X}|\theta)$ est impossible ;

$$f(x) = \sum_{k=1}^3 1_{\{Z_i=k\}} \mathbb{P}(Z = k) f_k(x)$$



- on considère des données cachées $\mathbf{Z} = (Z_1, \dots, Z_n)$ dont la connaissance rendrait possible la maximisation de la « vraisemblance des données complètes », $\log P(\mathbf{X}, \mathbf{Z}|\theta)$;

- comme on ne connaît pas ces données \mathbf{Z} , on estime la vraisemblance des données complètes en prenant en compte toutes les informations connues : l'estimateur est naturellement $\mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta_m} [\log P(\mathbf{X}, \mathbf{z}|\theta)]$ (« étape E » de l'algorithme) ;
- et on maximise enfin cette vraisemblance estimée pour déterminer la nouvelle valeur du paramètre (« étape M » de l'algorithme).

Ainsi, le passage de l'itération m à l'itération $m + 1$ de l'algorithme consiste à déterminer :

$$\theta_{m+1} = \arg \max_{\theta} \left\{ \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta_m} [\log P(\mathbf{X}, \mathbf{z}|\theta)] \right\}$$

3) Algorithme E.M.

- comme on ne connaît pas ces données \mathbf{Z} , on estime la vraisemblance des données complètes en prenant en compte toutes les informations connues : l'estimateur est naturellement $\mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta_m} [\log P(\mathbf{X}, \mathbf{z}|\theta)]$ (« étape E » de l'algorithme) ;

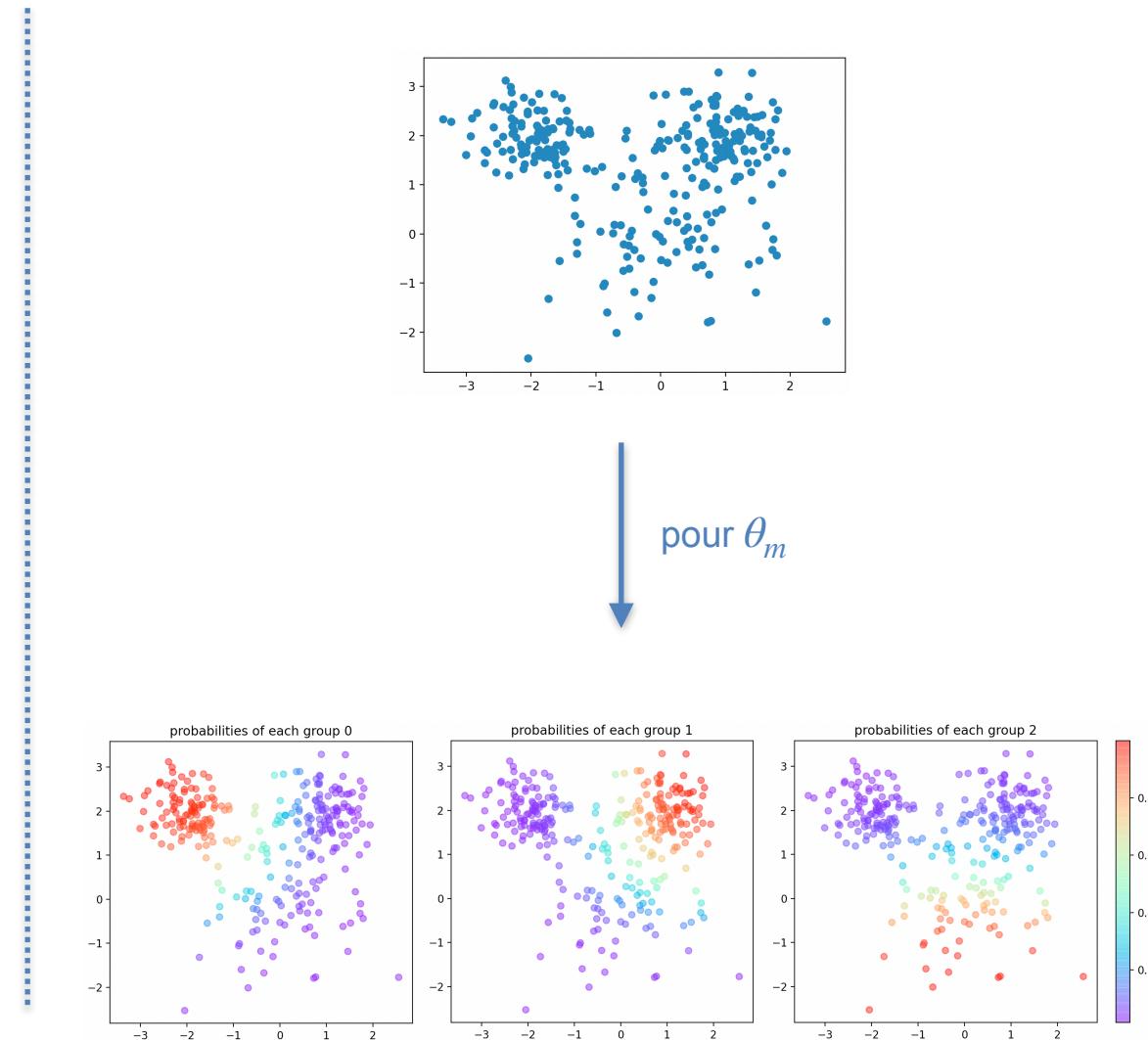
$$\tilde{p}_{k,i} = \mathbb{P}\{Z_i = k | X_i, \theta_m\}$$

$$= \frac{\mathbb{P}\{X_i | Z_i = k, \theta_m\} \mathbb{P}\{Z_i = k\}}{\sum_{l=1}^3 \mathbb{P}\{X_i, Z_i = l | \theta_m\}}$$

$$= \frac{\mathbb{P}\{X_i | Z_i = k, \theta_m\}}{\mathbb{P}\{X | \theta_m\}}$$

$$= \frac{p_k \mathbb{P}\{X_i | Z_i = k, \theta_m\}}{\sum_{l=1}^3 p_l \mathbb{P}\{X_i | Z_i = k, \theta_m\}}$$

avec θ_m l'estimation courante des paramètres et $\mathbb{P}\{X_i | Z_i = k, \theta_m\} = f_k(X_i)$



3) Algorithme E.M.

- comme on ne connaît pas ces données \mathbf{Z} , on estime la vraisemblance des données complètes en prenant en compte toutes les informations connues : l'estimateur est naturellement $\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m} [\log P(\mathbf{X}, \mathbf{z}|\theta)]$ (« étape E » de l'algorithme) ;

Pour les données complètes, on aurait :

$$\log(L_n(X, Z | \theta)) = \sum_{i=1}^n \left[\sum_{k=1}^K 1_{\{Z_i=k\}} \left(\log(p_k) - \log(2\pi) - \frac{1}{2} \log(\det(\Sigma_k)) - \frac{1}{2}(x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right) \right]$$

Les $1_{\{Z_i=k\}}$ ne sont pas connus. On va calculer leur espérance à l'aide des $\tilde{p}_{k,i}$:

$$\mathbb{E}_{Z|X,\theta_m} [\log(L_n(X, Z | \theta))] = \sum_{i=1}^n \sum_{k=1}^K \tilde{p}_{k,i} \left(\log(p_k) - \log(2\pi) - \frac{1}{2} \log(\det(\Sigma_k)) - \frac{1}{2}(x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \right)$$

car les $\tilde{p}_{k,i}$ ont été calculés avec θ_m

3) Algorithme E.M.

- et on maximise enfin cette vraisemblance estimée pour déterminer la nouvelle valeur du paramètre (« étape M » de l'algorithme).

Ainsi, le passage de l'itération m à l'itération $m + 1$ de l'algorithme consiste à déterminer :

$$\theta_{m+1} = \arg \max_{\theta} \left\{ \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m} [\log P(\mathbf{X}, \mathbf{z}|\theta)] \right\}$$

Optimisation de θ dans la fonctionnelle :

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta_m} [\log(L_n(\mathbf{X}, \mathbf{Z}|\theta))] = \sum_{i=1}^n \sum_{k=1}^K \tilde{p}_{k,i} \left(\log(p_k) - \log(2\pi) - \frac{1}{2} \log(\det(\Sigma_k)) - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right)$$

Peut être résolu avec les zéros du gradient ou bien par descente de gradient !

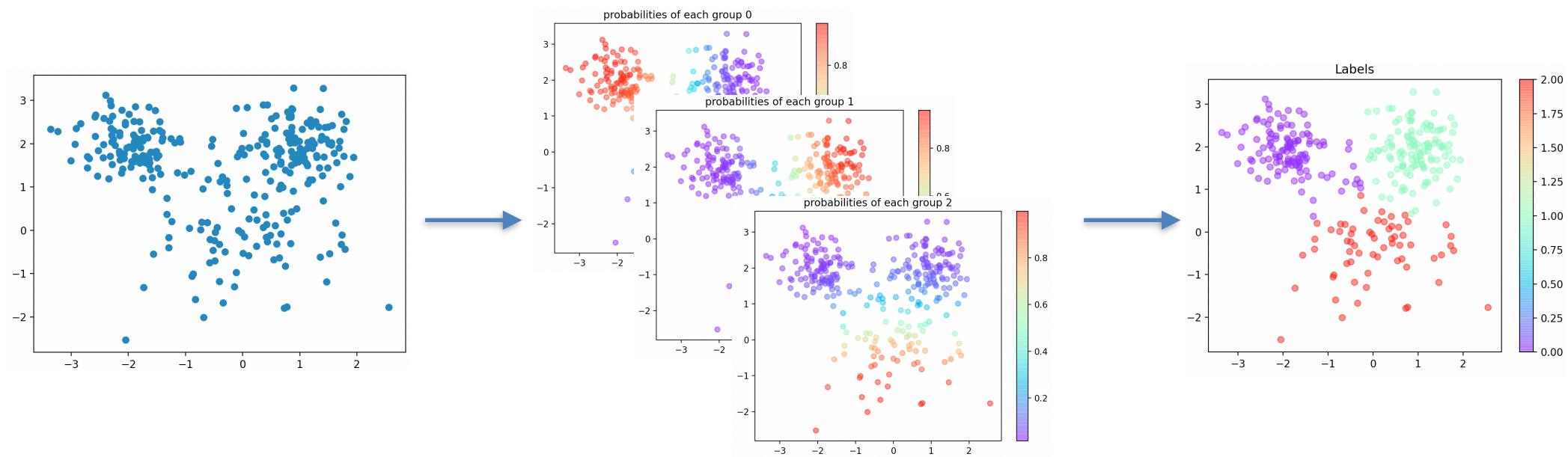
3) Algorithme Expectation-Maximisation (E.M.)

Tant que non-convergence :

- Une étape d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées,
- Une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

A la convergence :

On attribue à chaque observation son label le plus probable



4) Preuve de la croissance de la vraisemblance avec E.M.

Rappels d'analyse convexe

Définition 2 (Fonction convexe). — Une application $f : [a, b] \rightarrow \mathbb{R}$ est dite *convexe* sur $[a, b]$ si pour tous x_1, x_2 de cet intervalle et tout $\lambda \in [0, 1]$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \quad (5)$$

f est dite *strictement convexe* si l'inégalité (5) est stricte, et f est dite *concave* si $-f$ est convexe.

THÉORÈME 1. — Si f est deux fois dérivable sur $[a, b]$ et si $f''(x) \geq 0$ pour tout $x \in [a, b]$, alors f est convexe sur cet intervalle. \diamond

THÉORÈME 2 (Inégalité de Jensen). — Soit f une fonction convexe définie sur un intervalle I . Si $x_1, \dots, x_n \in I$ et $\lambda_1, \dots, \lambda_n \geq 0$ tels que $\sum \lambda_i = 1$, alors :

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i)$$

4) Preuve de la croissance de la vraisemblance avec E.M.

À l'itération m , nous disposons d'une valeur $\theta_m \in \mathbb{R}^d$ du vecteur de paramètres. Le but est de la mettre à jour avec une « meilleure » valeur θ , augmentant la vraisemblance, donc telle que $\Delta(\theta, \theta_m) := \log P(\mathbf{X}|\theta) - \log P(\mathbf{X}|\theta_m) \geq 0$. On souhaite bien sûr que cette différence soit la plus grande possible.

Cependant, comme précédemment exposé en section 2, on ne sait pas maximiser $P(\mathbf{X}|\theta)$, donc on ne sait pas non plus maximiser $\Delta(\theta, \theta_m)$... Un moyen d'optimiser malgré tout, dans une certaine mesure, cette différence, peut consister à chercher une fonction $\theta \mapsto \delta(\theta|\theta_m)$ que l'on sait maximiser, et qui est telle que :

$$\begin{cases} \Delta(\theta, \theta_m) & \geq \delta(\theta|\theta_m) \quad \forall \theta \in \mathbb{R}^d \\ \delta(\theta_m|\theta_m) & = 0 \end{cases} \quad (8)$$

Ainsi, $\delta(\theta|\theta_m)$ borne inférieurement $\Delta(\theta, \theta_m)$, et son maximum est au moins égal à 0. Trouver un θ' qui maximise $\theta \mapsto \delta(\theta|\theta_m)$ conduit donc mécaniquement à obtenir $\Delta(\theta', \theta_m) \geq 0$, c'est à dire une nouvelle valeur θ' plus vraisemblable des paramètres.

4) Preuve de la croissance de la vraisemblance avec E.M.

Afin de trouver une telle fonction δ , nous utilisons une représentation marginale de la vraisemblance selon les « données cachées » $\mathbf{Z} = (Z_1, \dots, Z_n)$:

$$P(\mathbf{X}|\theta) = \sum_{\mathbf{z}} P(\mathbf{X}, \mathbf{z}|\theta) = \sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta) \quad (9)$$

Il vient alors :

$$\begin{aligned} \Delta(\theta, \theta_m) &= \log P(\mathbf{X}|\theta) - \log P(\mathbf{X}|\theta_m) \\ &= \log \left(\sum_{\mathbf{z}} P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta) \right) - \underbrace{\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m)}_{=1} \log P(\mathbf{X}|\theta_m) \end{aligned} \quad (10)$$

Cette expression utilise le logarithme d'une somme : en se souvenant de l'inégalité de Jensen (6), on commence à voir apparaître clairement une façon de minorer $\Delta(\theta, \theta_m)...$

4) Preuve de la croissance de la vraisemblance avec E.M.

Nous réécrivons (10) en introduisant dans la somme de gauche les $P(\mathbf{z}|\mathbf{X}, \theta_m)$ présents dans la somme de droite, afin de se rapprocher de l'expression (7) :

$$\Delta(\theta, \theta_m) = \log \left(\sum_{\mathbf{z}} \frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta_m)} \cdot P(\mathbf{z}|\mathbf{X}, \theta_m) \right) - \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log P(\mathbf{X}|\theta_m) \quad (11)$$

Et enfin, en remarquant que $\sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) = 1$, nous appliquons l'inégalité de Jensen :

$$\begin{aligned} \Delta(\theta, \theta_m) &\geq \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log \left(\frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta_m)} \right) - \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log P(\mathbf{X}|\theta_m) \\ &= \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log \left(\frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta_m)} \right) - \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log P(\mathbf{X}|\theta_m) \\ &= \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log \left(\frac{P(\mathbf{X}|\mathbf{z}, \theta)P(\mathbf{z}|\theta)}{P(\mathbf{z}|\mathbf{X}, \theta_m)P(\mathbf{X}|\theta_m)} \right) \end{aligned} \quad (12)$$

$$= \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log \left(\frac{P(\mathbf{X}, \mathbf{z}|\theta)}{P(\mathbf{X}, \mathbf{z}|\theta_m)} \right) \quad (13)$$

$$=: \delta(\theta|\theta_m)$$

4) Preuve de la croissance de la vraisemblance avec E.M.

Nous avons donc obtenu une fonction $\theta \mapsto \delta(\theta|\theta_m)$ vérifiant les conditions (8) — il est évident avec (13) que $\delta(\theta_m|\theta_m) = 0$.

Finalement, nous posons :

$$\begin{aligned}
 \theta_{m+1} &= \arg \max_{\theta} \delta(\theta|\theta_m) \\
 &= \arg \max_{\theta} \left\{ \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log \left(\frac{P(\mathbf{X}, \mathbf{z}|\theta)}{P(\mathbf{X}, \mathbf{z}|\theta_m)} \right) \right\} \\
 &= \arg \max_{\theta} \left\{ \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{X}, \theta_m) \log P(\mathbf{X}, \mathbf{z}|\theta) \right\} \\
 &= \arg \max_{\theta} \left\{ \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta_m} [\log P(\mathbf{X}, \mathbf{z}|\theta)] \right\}
 \end{aligned} \tag{14}$$

On détermine bien ainsi une valeur θ_{m+1} plus vraisemblable que θ_m , puisque :

$$\log P(\mathbf{X}|\theta_{m+1}) - \log P(\mathbf{X}|\theta_m) = \Delta(\theta_{m+1}, \theta_m) \geq \delta(\theta_{m+1}|\theta_m) \geq \delta(\theta_m|\theta_m) \geq 0$$

MERCI !!!