

# MAPSI — cours 4 : Expectation-Maximization (EM)

Pierre-Henri Wuillemin – Christophe Gonzales

LIP6 / ISIR – Sorbonne Université, France

- 1 Principes d'apprentissage avec données manquantes
- 2 Quelques rappels de maths
- 3 L'algorithme EM
- 4 Pourquoi fonctionne-t-il ?
- 5 Mixtures de Gaussiennes et EM

X	Y
a	?
a	?
a	?
a	?
a	c
a	c
a	c
a	d
b	c
b	c
b	d
b	d

## Algorithme naïf :

- Supprimer les enregistrements avec ?

⇒

X	Y
a	c
a	c
a	c
a	d
b	c
b	c
b	d
b	d

- Données complètes
- Apprentissage : cf. cours précédents
- **Problème :**

Tableau gauche :  $P(X = a) = 2/3$

Tableau droit :  $P(X = a) = 1/2$

⇒ essayer de tenir compte de tous les enregistrements

## Algorithme k-means :

- Remplacer les ? par leur valeur la plus probable

X	Y
a	?
a	?
a	?
a	?
a	c
a	c
a	c
a	d
b	c
b	c
b	d
b	d



X	Y
a	c
a	c
a	c
a	c
a	c
a	c
a	c
a	d
b	c
b	c
b	d
b	d

- Données complètes
- Apprentissage : cf. cours précédents
- **Problème :**

Tableau gauche :  $P(Y = c|X = a) = 3/4$   
(sur les données observées)

Tableau droit :  $P(Y = c|X = a) = 7/8$

⇒ essayer de tenir compte de toutes les valeurs possibles de Y

## Algorithme naïf 2 :

- Remplacer les ? par toutes les valeurs possibles

X	Y
a	?
a	?
a	?
a	?
a	c
a	c
a	c
a	d
b	c
b	c
b	d
b	d



X	Y
a	c
a	d
a	c
a	d
a	c
a	d
a	c
a	d
a	c
a	d

X	Y
a	c
a	c
a	c
a	d
a	d
a	d
b	c
b	c
b	d
b	d

### Problèmes :

Tableau gauche :  $P(X = a) = 2/3$

Tableau droit :  $P(X = a) = 3/4$

Tableau gauche :  $P(Y = c|X = a) = 3/4$

Tableau droit :  $P(Y = c|X = a) = 7/12$

⇒ essayer de tenir compte des distributions des valeurs

## Algorithme EM :

- Remplacer les ? par toutes les valeurs possibles pondérées par leur probabilité d'apparition

X	Y
a	?
a	?
a	?
a	?
a	c
a	c
a	c
a	d
b	c
b	c
b	d
b	d

⇒

X	Y	w
a	c	3/4
a	d	1/4
a	c	3/4
a	d	1/4
a	c	3/4
a	d	1/4
a	c	3/4
a	d	1/4
a	c	3/4
a	d	1/4

X	Y	w
a	c	1
a	c	1
a	c	1
a	d	1
b	c	1
b	c	1
b	d	1
b	d	1

Apprentissage ⇒ comptages

⇒ sommer les poids

$$\Rightarrow P(X = a) = \frac{8}{12} = \frac{2}{3}$$

$$\Rightarrow P(Y = c | X = a) = \frac{6}{8} = \frac{3}{4}$$

⇒ Tableau gauche = droit

- K-means : remplacer ? par la valeur la plus probable
- EM : Remplacer ? par toutes les valeurs possibles pondérées par leur probabilité d'apparition

⇒ On connaît la probabilité des valeurs de  $Y|X$

⇒ On a un modèle probabiliste de ces valeurs

Or, c'est justement le modèle qu'on souhaite apprendre !

Idée clef de K-means et EM : algos itératifs

- 1 se donner un modèle initial (pas trop mauvais)
- 2 ce modèle ⇒ données complètes
- 3 apprendre nouveau modèle avec ces données
- 4 revenir en 2 avec nouveau modèle si  $\neq$  ancien modèle

## Problèmes :

- 1 Y a-t-il convergence ?
- 1 À convergence, est-ce que l'on a obtenu un bon modèle ?

But du reste du cours 4 : répondre à ces questions pour EM



- $\mathbf{x}^o$  : données observées,  $\mathbf{x}^h$  : données manquantes
- $\mathbf{x} = \mathbf{x}^o \cup \mathbf{x}^h$

Film	$r_A$	$r_B$	$r_C$	$r_D$
I robot	4	?	3	3
Forest Gump	?	?	2	4
Intouchables	2	2	3	?
Le parrain	1	?	2	?
Pulp fiction	2	4	3	4

- $\mathcal{M}_{ij} = P(r_i^j \in \mathbf{x}^h)$  : position des données manquantes

## Typologies

- **Missing Completely at Random (MCAR)** :  $P(\mathcal{M}|\mathbf{x}) = P(\mathcal{M})$  Aucune relation entre le fait qu'une donnée soit manquante ou observée
- **Missing at Random (MAR)** :  $P(\mathcal{M}|\mathbf{x}) = P(\mathcal{M}|\mathbf{x}^o)$  données manquantes en relation avec les données observées mais pas avec les autres données manquantes
- **Not Missing At Random (NMAR)** :  $P(\mathcal{M}|\mathbf{x})$  données manquantes en relation avec *toutes* les données



On n'étudiera que MCAR !

- Échantillon  $\mathbf{x} = \{x_1, \dots, x_n\}$  de taille  $n$

- données complètes :  $\log L(\mathbf{x}, \Theta) = \sum_{i=1}^n \log P(x_i | \Theta)$

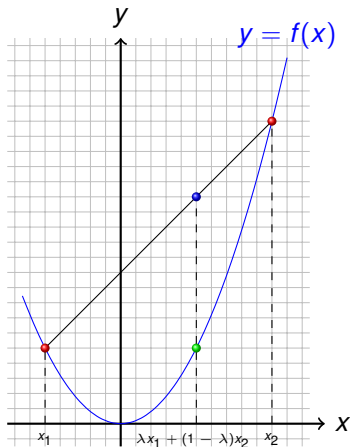
- $\mathbf{x}^o$  : données observées,  $\mathbf{x}^h$  : données manquantes

- $\log L(\mathbf{x}^o, \Theta) = \log$ -vraisemblance des données observées  
$$= \sum_{i=1}^n \log P(x_i^o | \Theta) = \sum_{i=1}^n \log \left( \sum_{\mathbf{x}_i^h \in \mathbf{x}^h} P(x_i^o, \mathbf{x}_i^h | \Theta) \right)$$

- Soit  $Q_i(x_i^h)$  une loi de proba **quelconque** alors :

$$\log L(\mathbf{x}^o, \Theta) = \sum_{i=1}^n \log \left( \sum_{\mathbf{x}_i^h \in \mathbf{x}^h} Q_i(\mathbf{x}_i^h) \frac{P(x_i^o, \mathbf{x}_i^h | \Theta)}{Q_i(\mathbf{x}_i^h)} \right)$$

# Rappel : fonctions convexes



## Définition

$f$  convexe  $\iff \forall \lambda \in [0, 1], \forall x_1, x_2 :$

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

## fonction concave

$f$  concave  $\iff -f$  convexe

## *Inégalité de Jensen*

- $f$  convexe définie sur  $D$
- $x_1, \dots, x_n \in D$
- $\lambda_1, \dots, \lambda_n \geq 0, \quad \sum_{i=1}^n \lambda_i = 1$

Alors :

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

## *Inégalité de Jensen*

- $f$  convexe
- $X$  : variable aléatoire à  $n$  dimensions  $x_1, \dots, x_n$
- $\lambda_1, \dots, \lambda_n \geq 0, \quad \sum_{i=1}^n \lambda_i = 1 \implies$  probabilité  $P_\lambda$
- $f(\mathbb{E}_{P_\lambda}(X)) \leq \mathbb{E}_{P_\lambda}(f(X))$  où  $\mathbb{E}_{P_\lambda}$  = espérance

# Démonstration de l'inégalité de Jensen

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

- par récurrence : si  $n = 1$  : trivial, si  $n = 2$  : convexité

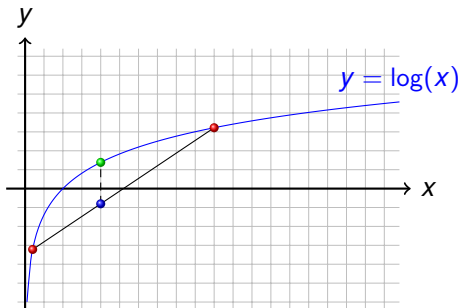
- $$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^n \lambda_i x_i\right) \\ &= f\left(\lambda_{n+1} x_{n+1} + (1 - \lambda_{n+1}) \sum_{i=1}^n \frac{\lambda_i}{1 - \lambda_{n+1}} x_i\right) \\ &\leq \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) f\left(\sum_{i=1}^n \frac{\lambda_i}{1 - \lambda_{n+1}} x_i\right) \\ &\leq \lambda_{n+1} f(x_{n+1}) + (1 - \lambda_{n+1}) \sum_{i=1}^n \frac{\lambda_i}{1 - \lambda_{n+1}} f(x_i) \\ &= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^n \lambda_i f(x_i) = \sum_{i=1}^{n+1} \lambda_i f(x_i) \end{aligned}$$

# Conséquences de l'inégalité de Jensen

## *Inégalité de Jensen pour le logarithme*


Logarithme = fonction concave :

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




$$\mathbb{E}(\log(X)) = \log(\mathbb{E}(X)) \implies X = \mathbb{E}(X) = \text{constante}$$

- $\log L(\mathbf{x}^o, \Theta) = \sum_{i=1}^n \log \left( \sum_{x_i^h \in \mathbf{x}^h} \textcolor{red}{Q_i(x_i^h)} \frac{P(x_i^o, x_i^h | \Theta)}{\textcolor{red}{Q_i(x_i^h)}} \right)$

 inégalité de Jensen  $\implies \log \left( \sum_{i=1}^n \lambda_i y_i \right) \geq \sum_{i=1}^n \lambda_i \log(y_i)$

$$\log L(\mathbf{x}^o, \Theta) \geq \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{Q_i(x_i^h)} \right)$$

 Jensen  $\implies$  égalité ssi  $\frac{P(x_i^o, x_i^h | \Theta)}{Q_i(x_i^h)} = \text{constante}$

choisir  $Q_i(x_i^h) \propto P(x_i^o, x_i^h | \Theta) \implies \textcolor{red}{Q_i(x_i^h)} = \textcolor{red}{P(x_i^h | x_i^o, \Theta)}$



## Algorithme

- 1 choisir une valeur initiale  $\Theta = \Theta^0$
- 2 **Étape E (expectation) :**
  - $Q_i^{t+1}(x_i^h) \leftarrow P(x_i^h | x_i^o, \Theta^t) \quad \forall i \in \{1, \dots, n\}$
  - $\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right)$
- 3 **Étape M (maximization) :**
  - $\Theta^{t+1} \leftarrow \text{Argmax}_{\Theta} \log L^{t+1}(\mathbf{x}^o, \Theta)$
- 4 Tant qu'on n'a pas convergé, revenir en 2

À convergence,  $\Theta^{t+1} = \text{optimum local par max de vraisemblance}$

# Algorithme EM : un exemple

- 2 variables aléatoires  $A \in \{a, b\}$  et  $C \in \{c, d\}$

$$P(A, C|\Theta) = \begin{bmatrix} \theta_{ac} & \theta_{ad} \\ \theta_{bc} & \theta_{bd} \end{bmatrix} \implies \Theta = \{\theta_{ac}, \theta_{ad}, \theta_{bc}, \theta_{bd}\}$$

but : estimer  $\Theta$  par EM

A	C
a	?
b	?
a	d
b	d
a	c

- A toujours observé :

$$\implies \theta_{ac} + \theta_{ad} = \frac{3}{5} \text{ par max de vraisemblance}$$

$$\theta_{bc} + \theta_{bd} = \frac{2}{5} \text{ par max de vraisemblance}$$

- initialisation possible :

$$\Theta^0 = \{\theta_{ac}^0 = 0.3, \theta_{ad}^0 = 0.3, \theta_{bc}^0 = 0.2, \theta_{bd}^0 = 0.2\}$$

- Étape E (expectation) :  $Q_i^1(x_i^h) \leftarrow P(x_i^h|x_i^o, \Theta^0) \quad \forall i \in \{1, 2\}$

$$Q_1^1(C) = P(C|A = a, \Theta^0) = \frac{P(A=a, C|\Theta^0)}{\sum_C P(A=a, C|\Theta^0)} = \left[\frac{0.3}{0.6}, \frac{0.3}{0.6}\right] = [0.5, 0.5]$$

$$Q_2^1(C) = P(C|A = b, \Theta^0) = \frac{P(A=b, C|\Theta^0)}{\sum_C P(A=b, C|\Theta^0)} = \left[\frac{0.2}{0.4}, \frac{0.2}{0.4}\right] = [0.5, 0.5]$$

# Algorithme EM : un exemple

$$\Theta^0 = \{\theta_{ac}^0 = 0.3, \theta_{ad}^0 = 0.3, \theta_{bc}^0 = 0.2, \theta_{bd}^0 = 0.2\}$$

$$Q_1^1(C) = [0.5, 0.5] \quad Q_2^1(C) = [0.5, 0.5] \quad P(x_i^h, x_i^o | \Theta^0) = \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix}$$

$$\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right)$$

A	B			
a	a	0.5	$\theta_{ac}/0.5$	$\log \theta_{ac} - \log 0.5$
a	d	0.5	$\theta_{ad}/0.5$	$\log \theta_{ad} - \log 0.5$
b	b	0.5	$\theta_{bc}/0.5$	$\log \theta_{bc} - \log 0.5$
a	d	0.5	$\theta_{bd}/0.5$	$\log \theta_{bd} - \log 0.5$
b	d	1	$\theta_{ad}$	$\log \theta_{ad}$
b	d	1	$\theta_{bd}$	$\log \theta_{bd}$
a	c	1	$\theta_{ac}$	$\log \theta_{ac}$



⇒ revient à observer l'échantillon avec poids  $Q_i^{t+1}$

$$\begin{aligned}\log L^{t+1}(\mathbf{x}^o, \Theta) &= \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right) \\ &= \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \left[ \log(P(x_i^o, x_i^h | \Theta)) - \log(Q_i^{t+1}(x_i^h)) \right]\end{aligned}$$

$$\begin{aligned}\Rightarrow \Theta^{t+1} &= \text{Argmax}_{\Theta} \log L^{t+1}(\mathbf{x}^o, \Theta) \\ &= \text{Argmax}_{\Theta} \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log(P(x_i^o, x_i^h | \Theta))\end{aligned}$$

## *Principe de EM*

Étape M  $\Rightarrow$  maximum de vraisemblance avec un échantillon dont chaque enregistrement  $x_i$  a un poids  $Q_i^{t+1}$

# Algorithme EM : un exemple

$$\Theta^1 = \operatorname{Argmax}_{\Theta} \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right)$$

A	B	$Q_i^{t+1}$	$\log \theta$
a	c	0.5	$\log \theta_{ac}$
a	d	0.5	$\log \theta_{ad}$
b	c	0.5	$\log \theta_{bc}$
b	d	0.5	$\log \theta_{bd}$
a	d	1	$\log \theta_{ad}$
b	d	1	$\log \theta_{bd}$
a	c	1	$\log \theta_{ac}$

$$\Theta^1 = \operatorname{Argmax}_{\Theta} [0.5 + 1] \log \theta_{ac} + [0.5 + 1] \log \theta_{ad} + 0.5 \log \theta_{bc} + [0.5 + 1] \log \theta_{bd}$$

Sous contrainte :  $\theta_{ac} + \theta_{ad} + \theta_{bc} + \theta_{bd} = 1$

$$\Theta^1 = \left\{ \theta_{ac}^1 = \frac{3}{10}, \theta_{ad}^1 = \frac{3}{10}, \theta_{bc}^1 = \frac{1}{10}, \theta_{bd}^1 = \frac{3}{10} \right\}$$

# Algorithme EM : un exemple

$$\Theta^1 = \{\theta_{ac}^1 = \frac{3}{10}, \theta_{ad}^1 = \frac{3}{10}, \theta_{bc}^1 = \frac{1}{10}, \theta_{bd}^1 = \frac{3}{10}\}$$

A	C
a	?
b	?
a	d
b	d
a	c

● **Étape E (expectation)** :  $Q_i^2(x_i^h) \leftarrow P(x_i^h | x_i^o, \Theta^1) \quad \forall i \in \{1, 2\}$

$$Q_1^2(C) = P(C|A = a, \Theta^1) = \frac{P(A=a, C|\Theta^1)}{\sum_C P(A=a, C|\Theta^1)} = \left[\frac{0.3}{0.6}, \frac{0.3}{0.6}\right] = [0.5, 0.5]$$

$$Q_2^2(C) = P(C|A = b, \Theta^1) = \frac{P(A=b, C|\Theta^1)}{\sum_C P(A=b, C|\Theta^1)} = \left[\frac{0.1}{0.4}, \frac{0.3}{0.4}\right] = [0.25, 0.75]$$

# Algorithme EM : un exemple

$$\Theta^1 = \{\theta_{ac}^1 = \frac{3}{10}, \theta_{ad}^1 = \frac{3}{10}, \theta_{bc}^1 = \frac{1}{10}, \theta_{bd}^1 = \frac{3}{10}\}$$

$$Q_1^2(C) = [0.5, 0.5] \quad Q_2^2(C) = [0.25, 0.75] \quad P(x_i^h, x_i^o | \Theta^0) = \begin{bmatrix} 0.3 & 0.3 \\ 0.1 & 0.3 \end{bmatrix}$$

$$\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right)$$

A	B			
a	c	0.5	$\theta_{ac}/0.5$	$\log \theta_{ac} - \log 0.5$
a	d	0.5	$\theta_{ad}/0.5$	$\log \theta_{ad} - \log 0.5$
b	c	0.25	$\theta_{bc}/0.25$	$\log \theta_{bc} - \log 0.25$
a	d	0.75	$\theta_{bd}/0.75$	$\log \theta_{bd} - \log 0.75$
b	d	1	$\theta_{ad}$	$\log \theta_{ad}$
b	d	1	$\theta_{bd}$	$\log \theta_{bd}$
a	c	1	$\theta_{ac}$	$\log \theta_{ac}$

$$\Theta^2 = \text{Argmax}_{\Theta} [0.5 + 1] \log \theta_{ac} + [0.5 + 1] \log \theta_{ad} + 0.25 \log \theta_{bc} + [0.75 + 1] \log \theta_{bd}$$

Sous contrainte :  $\theta_{ac} + \theta_{ad} + \theta_{bc} + \theta_{bd} = 1$

# Algorithme EM : un exemple

- $\Theta^2 = \{\theta_{ac}^2 = \frac{3}{10}, \theta_{ad}^2 = \frac{3}{10}, \theta_{bc}^2 = \frac{1}{20}, \theta_{bd}^2 = \frac{7}{20}\}$

- $\Theta^3 = \{\theta_{ac}^3 = \frac{3}{10}, \theta_{ad}^3 = \frac{3}{10}, \theta_{bc}^3 = \frac{1}{40}, \theta_{bd}^3 = \frac{15}{40}\}$

...

---

- $\theta_{ac} = \theta_{bc} = 0,3$

- $\theta_{bc} + \theta_{bd} = 0,4$  et  $\theta_{bc}$  divisé par 2 à chaque étape.

$\Rightarrow$  à convergence :  $\Theta = \{\theta_{ac} = \frac{3}{10}, \theta_{ad} = \frac{3}{10}, \theta_{bc} = 0, \theta_{bd} = \frac{4}{10}\}$



# Système de recommandation : le retour

Film	$r_A$	$r_B$	$r_C$	$r_D$		
I robot	4	?	3	3	$\leftarrow \alpha_1^{t+1}(r_B)$	$\implies 4$ enregistrements
Forest Gump	?	?	2	4	$\leftarrow \alpha_2^{t+1}(r_A, r_B)$	$\implies 16$ enregistrements
Intouchables	2	2	3	?	$\leftarrow \alpha_3^{t+1}(r_D)$	$\implies 4$ enregistrements
Le parrain	1	?	2	?	$\leftarrow \alpha_4^{t+1}(r_B, r_D)$	$\implies 16$ enregistrements
Pulp fiction	2	4	3	4	$\leftarrow \alpha_5^{t+1}() = 1$	$\implies 1$ enregistrement

$\implies 4 + 16 + 4 + 16 + 1 = 41$  enregistrements pour calculer  $\Theta^{t+1}$

# Convergence de EM : monotonie

Étape E :  $\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} \boxed{Q_i^{t+1}}(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta)}{\boxed{Q_i^{t+1}}(x_i^h)} \right)$

Étape M :  $\Theta^{t+1} \leftarrow \text{Argmax}_{\Theta} \log L^{t+1}(\mathbf{x}^o, \Theta)$

$$\Rightarrow \log L^{t+1}(\mathbf{x}^o, \Theta^{t+1}) \geq \log L^{t+1}(\mathbf{x}^o, \Theta^t)$$

*Rappel : inégalité de Jensen*

$\forall$  loi de proba  $Q_i(x_i^h)$  :

$$\log L(\mathbf{x}^o, \Theta^t) \geq \sum_{i=1}^n \sum_{x_i^h \in \mathbf{x}^h} Q_i(x_i^h) \log \left( \frac{P(x_i^o, x_i^h | \Theta^t)}{Q_i(x_i^h)} \right)$$

$$\text{égalité} \iff Q_i(x_i^h) = P(x_i^h | x_i^o, \Theta^t) = Q_i^{t+1}(x_i^h)$$

$$\Rightarrow \begin{cases} \log L^{t+1}(\mathbf{x}^o, \Theta^t) = \log L(\mathbf{x}^o, \Theta^t) \geq \log L^t(\mathbf{x}^o, \Theta^t) \\ \log L(\mathbf{x}^o, \Theta^{t+1}) \geq \log L^{t+1}(\mathbf{x}^o, \Theta^{t+1}) \end{cases}$$

$$\Rightarrow L(\mathbf{x}^o, \Theta^{t+1}) \geq L^{t+1}(\mathbf{x}^o, \Theta^{t+1}) \geq L(\mathbf{x}^o, \Theta^t) \geq L^t(\mathbf{x}^o, \Theta^t)$$

$$L(\mathbf{x}^o, \Theta^{t+1}) \geq L^{t+1}(\mathbf{x}^o, \Theta^{t+1}) \geq L(\mathbf{x}^o, \Theta^t) \geq L^t(\mathbf{x}^o, \Theta^t)$$

## *Propriété de EM*

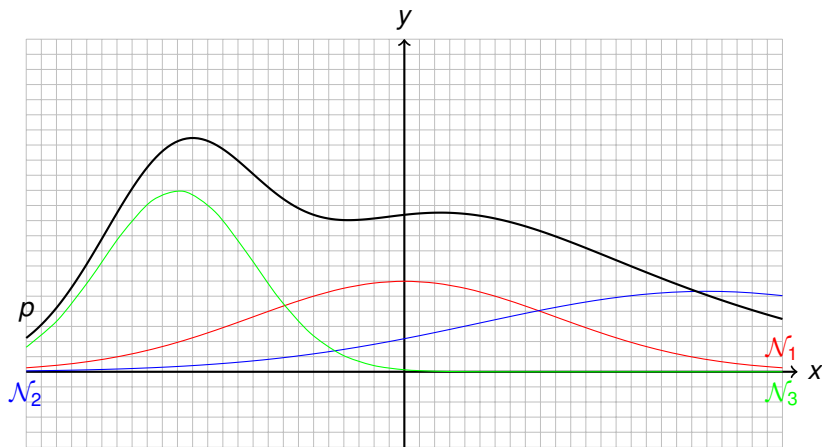
EM converge vers un maximum local de la vraisemblance



si  $\text{Argmax}_{\Theta} L^{t+1}(\mathbf{x}^o, \Theta)$  estimé par descente de gradient,  
on peut perdre cette propriété !

# Mixture de gaussiennes

$$p(\cdot) = 0,3 \times \mathcal{N}(0, 2^2) + 0,4 \times \mathcal{N}(4, 3^2) + 0,3 \times \mathcal{N}(-3, 1^2)$$



# Application : apprentissage de prix fonciers

postulat : prix de biens similaires dans un quartier  $\sim$  identiques

$\Rightarrow$  prix dépendent  $\left\{ \begin{array}{l} \text{des caractéristiques du bien (e.g. nombre de pièces)} \\ \text{du quartier} \end{array} \right.$

$\Rightarrow$  modélisation par une mixture de gaussiennes (ici 2 gaussiennes)

## *Modélisation du problème*

- $\Theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2\}$
- $p(x|\Theta) = \pi_1 \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(\mu_2, \sigma_2^2)$

## *Apprentissage non supervisé*

- échantillon  $\mathbf{x} = \langle x_1, \dots, x_n \rangle$
- $x_i = \text{prix} \Rightarrow$  on ne connaît pas la Gaussienne à laquelle le bien appartient



échantillon supposé complet (pas de données manquantes)

échantillon complet  $\implies$  estimation par max de vraisemblance

$$L(\mathbf{x}, \Theta) = \prod_{i=1}^n p(x_i | \Theta) = \prod_{i=1}^n \sum_{k=1}^2 \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp \left\{ -\frac{1}{2} \left( \frac{x_i - \mu_k}{\sigma_k} \right)^2 \right\}$$

$$\log L(\mathbf{x}, \Theta) = \sum_{i=1}^n \log \left[ \sum_{k=1}^2 \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp \left\{ -\frac{1}{2} \left( \frac{x_i - \mu_k}{\sigma_k} \right)^2 \right\} \right]$$



trop compliqué à maximiser analytiquement !

*Solution : EM*

- ❶  $x_i$  appartient à une classe  $y_{k(i)}$  **non observée**  $\sim \mathcal{N}(\mu_{k(i)}, \sigma_{k(i)})$
- ❷ échantillon  $\mathbf{x} = \langle (x_i, y_{k(i)}) \rangle$
- ❸ échantillon maintenant avec données manquantes  $\implies$  EM

## Nouvelle modélisation du problème

$$p(X_i, Y_{k(i)}|\Theta) = p(X_i|Y_{k(i)}, \Theta)P(Y_{k(i)}|\Theta) = \begin{bmatrix} \mathcal{N}(\mu_1, \sigma_1^2) \pi_1 \\ \mathcal{N}(\mu_2, \sigma_2^2) \pi_2 \end{bmatrix}$$

$\Rightarrow$  pour  $x_i$  connu :

$$P(Y_{k(i)}|x_i, \Theta) = \frac{p(x_i, Y_{k(i)}|\Theta)}{p(x_i|\Theta)} \propto \begin{bmatrix} \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_1}{\sigma_1}\right)^2\right\} \times \pi_1 \\ \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_2}{\sigma_2}\right)^2\right\} \times \pi_2 \end{bmatrix}$$

● **Initialisation d'EM** : choisir une valeur  $\Theta^0 = \{\mu_1^0, \mu_2^0, \sigma_1^0, \sigma_2^0, \pi_1^0, \pi_2^0\}$

● **Étape E** :  $Q_i^1(y_k) \leftarrow P(y_k|x_i, \Theta^0)$  pour  $k = 1, 2$

$\Rightarrow Q_i^1(\cdot)$  très facile à calculer

● **Étape M** :

$$\underset{\Theta}{\text{Argmax}} \log L^{t+1}(\mathbf{x}^o, \Theta) = \underset{\Theta}{\text{Argmax}} \sum_{i=1}^n \sum_{k=1}^2 Q_i^{t+1}(y_k) \log \left( \frac{p(x_i, y_k|\Theta)}{Q_i^{t+1}(y_k)} \right)$$

## Étape M :

$$\begin{aligned} & \text{Argmax}_{\Theta} \log L^{t+1}(\mathbf{x}^o, \Theta) \\ &= \text{Argmax}_{\Theta} \sum_{i=1}^n \sum_{k=1}^2 Q_i^{t+1}(y_k) \log \left( \frac{p(x_i, y_k | \Theta)}{Q_i^{t+1}(y_k)} \right) \\ &= \text{Argmax}_{\Theta} \sum_{i=1}^n Q_i^{t+1}(y_1) \log \left( \pi_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp \left\{ -\frac{1}{2} \left( \frac{x_i - \mu_1}{\sigma_1} \right)^2 \right\} \right) + \\ & \quad Q_i^{t+1}(y_2) \log \left( \pi_2 \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp \left\{ -\frac{1}{2} \left( \frac{x_i - \mu_2}{\sigma_2} \right)^2 \right\} \right) \\ &= \text{Argmax}_{\Theta} \sum_{i=1}^n Q_i^{t+1}(y_1) \left[ \log(\pi_1) - \frac{1}{2} \log(\sigma_1^2) - \frac{1}{2} \left( \frac{x_i - \mu_1}{\sigma_1} \right)^2 \right] + \\ & \quad Q_i^{t+1}(y_2) \left[ \log(\pi_2) - \frac{1}{2} \log(\sigma_2^2) - \frac{1}{2} \left( \frac{x_i - \mu_2}{\sigma_2} \right)^2 \right] \end{aligned}$$



Argmax facile à calculer !



## Signatures spectrales en teintes de gris

- neige  $\sim \mathcal{N}(\mu_1, \sigma_1^2)$
- forêt  $\sim \mathcal{N}(\mu_2, \sigma_2^2)$
- désert  $\sim \mathcal{N}(\mu_3, \sigma_3^2)$
- mer  $\sim \mathcal{N}(\mu_4, \sigma_4^2)$
- $Y$  = observation en teinte de gris = pixels d'une image
- $Z$  = classe paysage  $\in \{1, 2, 3, 4\} \sim \text{distribution } (\pi_1, \pi_2, \pi_3, \pi_4)$

### *Paramètres du problème*

- $\Theta = \{(\mu_j, \sigma_j)\}_{j=1}^4 \cup \{\pi_j\}_{j=1}^4$
- $p(Y, Z|\Theta) = p(Y|Z, \Theta)P(Z|\Theta) = \sum_{j=1}^4 \pi_j \mathcal{N}(\mu_j, \sigma_j^2)$

- $Y$  : observations,  $Z$  : classes de paysage

- $p(Y, Z|\Theta) = \sum_{j=1}^4 \pi_j \mathcal{N}(\mu_j, \sigma_j^2)$

- base de données incomplète ou  $Z$  non observé

$\implies$  estimation de  $\Theta$  par EM (similaire aux prix fonciers)