Analyse de données et Traitement de l'Information: Analyse Statique

Stéphane Marchand-Maillet

Département d'Informatique
Université de Genève
stephane.marchand-maillet@unige.ch

Master en Sciences Informatiques - Semestre d'Automne

Analyse Sémantique Latente (ASL)

Analyse de documents textuels : construction de la matrice des occurences Termes-Documents

Analyse sémantique latente

L'**analyse sémantique latente** (**LSA**, de l'anglais : *Latent semantic analysis*) ou **indexation sémantique latente** (ou **LSI**, de l'anglais : *Latent semantic indexation*) est un procédé de traitement des langues naturelles, dans le cadre de la sémantique vectorielle. La LSA fut brevetée en 1988[1] et publiée en 1990[2].

Elle permet d'établir des relations entre un ensemble de documents et les termes qu'ils contiennent, en construisant des « concepts » liés aux documents et aux termes.

Sommaire [masquer]

1 Matrice des occurrences
2 Applications
3 Réduction du rang
4 Description
4.1 Construction de la matrice des occurrences
4.2 Corrélations
4.3 Décomposition en valeurs singulières
4.4 Espace des concepts
5 Implémentations
6 Limitations
7 Analyse sémantique latente probabiliste (PLSA)
8 Annexes
8.1 Bibliographie
8.2 Notes et références
9 Source
10 Voir aussi
10.1 Articles connexes
10.2 Liens externes

Matrice des occurrences [modifier]

La LSA utilise une matrice qui décrit l'occurrence de certains termes dans les documents. C'est une matrice creuse dont les lignes correspondent aux « termes » et dont les colonnes correspondent aux « documents ».

Les « termes » sont généralement des mots tronqués ou ramenés à leur radical, issus de l'ensemble du corpus. On a donc le nombre d'apparition d'un mot dans chaque document, et pour tous les mots. Ce nombre est normalisé en utilisant la pondération *tf-idf* (de l'anglais : *term frequency – inverse document frequency*), combinaison de deux techniques : un coefficient de la matrice est d'autant plus grand qu'il apparaît beaucoup dans un document, et qu'il est rare — pour les mettre en avant.

Cette matrice est courante dans les modèles sémantiques standards, comme le modèle vectoriel, quoique sa forme matricielle ne soit pas systématique, étant donné qu'on ne se sert que rarement des propriétés mathématiques des matrices.

La LSA transforme la matrice des occurrences en une « relation » entre les termes et des « concepts », et une relation entre ces concepts et les documents. On peut donc relier des documents entre eux.

Applications [modifier]

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 1 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 1 \\ & & \vdots & & & \\ 0 & 1 & 1 & 0 & \dots & 0 \end{pmatrix}$$

- Matrice creuse (composée majoritairement de 0), dont le rang (dimension de l'espace dans lequel sont représentées les données) n'est pas égal au nombres de termes.

Espace des concepts

Reduction de rang/reduction de dimension par l'algorithme de decomposition en valeurs singulières

$$X = U\Lambda V^{\mathsf{T}} \Longleftrightarrow X^{\mathsf{T}}X = V\Lambda^2 V^{\mathsf{T}},\ XX^{\mathsf{T}} = U\Lambda^2 U^{\mathsf{T}}$$

Il est facile de voir que l'ASL est une ACP appliquée à la matrice de contingence termes-documents

- Les colonnes de V sont les composantes principales de l'espace des documents (les termes sont les variables).
- Les colonnes de U sont les composantes principales de l'espace des termes (les documents sont les variables).

Approximation de la matrice d'occurences et recherche d'information

En selectionnnant les k plus grandes valeurs singulières de Λ, on obtient une approximation de rang k de la matrice des occurrences

$$X_k = U_k \Lambda_k V_k^{\mathsf{T}}$$

- distance inter-documents $= \cos(\widehat{d_i, d_j})$
- une requête q, mini document composée de termes est projetée dans l'espace des concepts

$$\hat{q} = \Lambda_k^{-1} U_k^{\mathsf{T}} q$$

Équivalence des méthodes spectrales

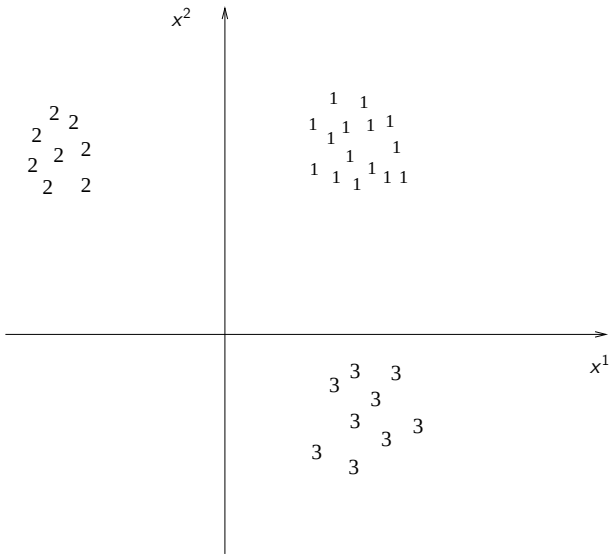
ACP, AFC, ASL sont en fait les même techniques appliquées à des tableaux de données différents

- ACP : tableau éléments-variables
- AFC : tableau de contingence variables-variables
- ASL : tableau de contingence termes-documents

$\Rightarrow$La recherche de la base diagonalisant la matrice de covariance fourni une nouvelle représentation de ces tableaux, avec comme critère la maximisation de la variance des données analysées

Analyse Discriminante Linéaire (ADL)

Nous nous intéressons à l'analyse de données décrites par p variables quantitatives et 1 une variable qualitative ayant q valeurs possibles. Chaque élément est alors caractérisé par le couple $(\mathbf{x}_i, y_i)$, où $\mathbf{x}_i \in \mathbb{R}^p$ et $y_i = 1, \ldots, q$.



Présentation de la méthode

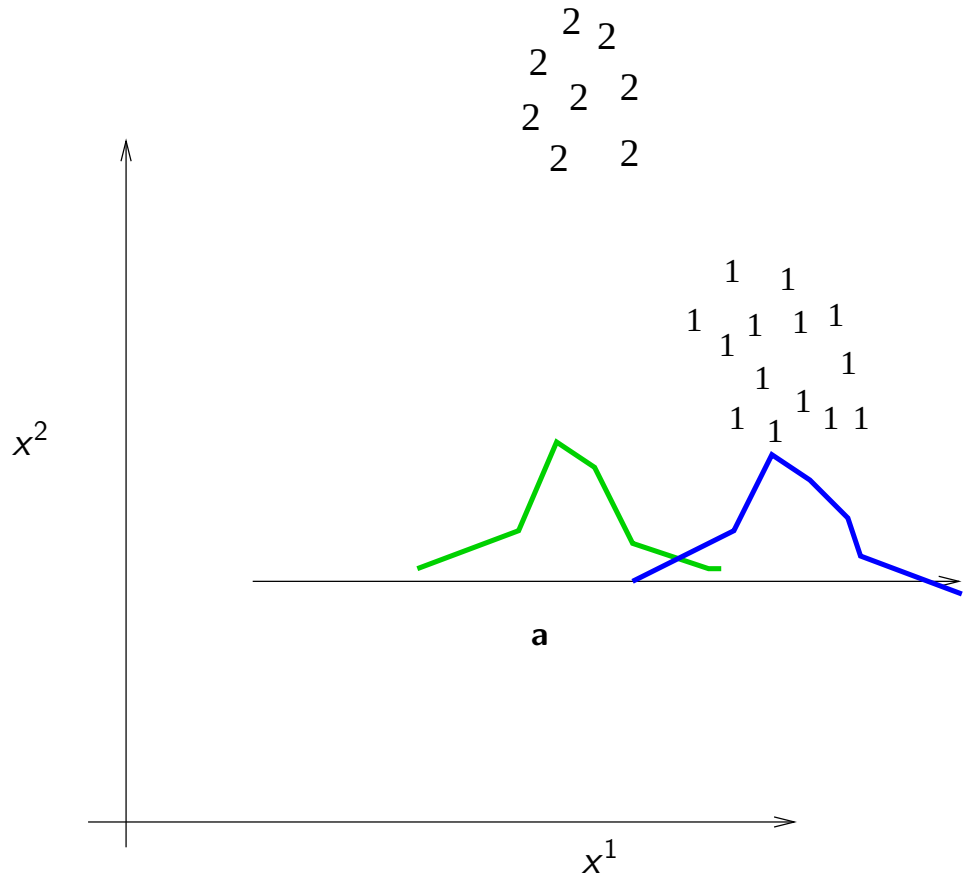
Définition des classes

Autre interprétation : la variable qualitative y_i définit la classe à laquelle appartient le point i, caractérisé par les variables $\mathbf{x}_i$. On dit que i appartient à C_k si $y_i = k$.

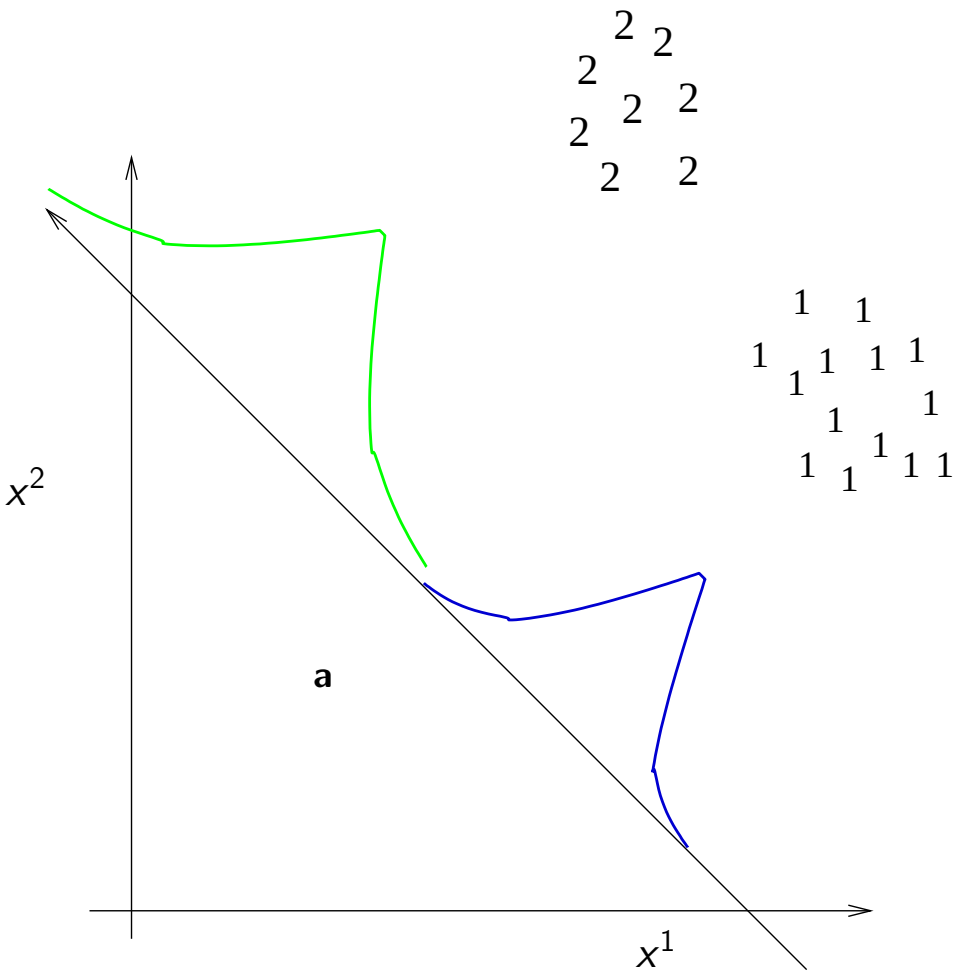
Analyse discriminante

Les q classes diffèrent-elles sur l'ensemble des variables quantitatives $\mathbf{x}$? Existe t'il une transformation lineaire de $\mathbf{x}$ telle que les q classes soient mieux séparées (discriminées) ?

2 Classes - Axe discriminant



2 Classes - Axe discriminant



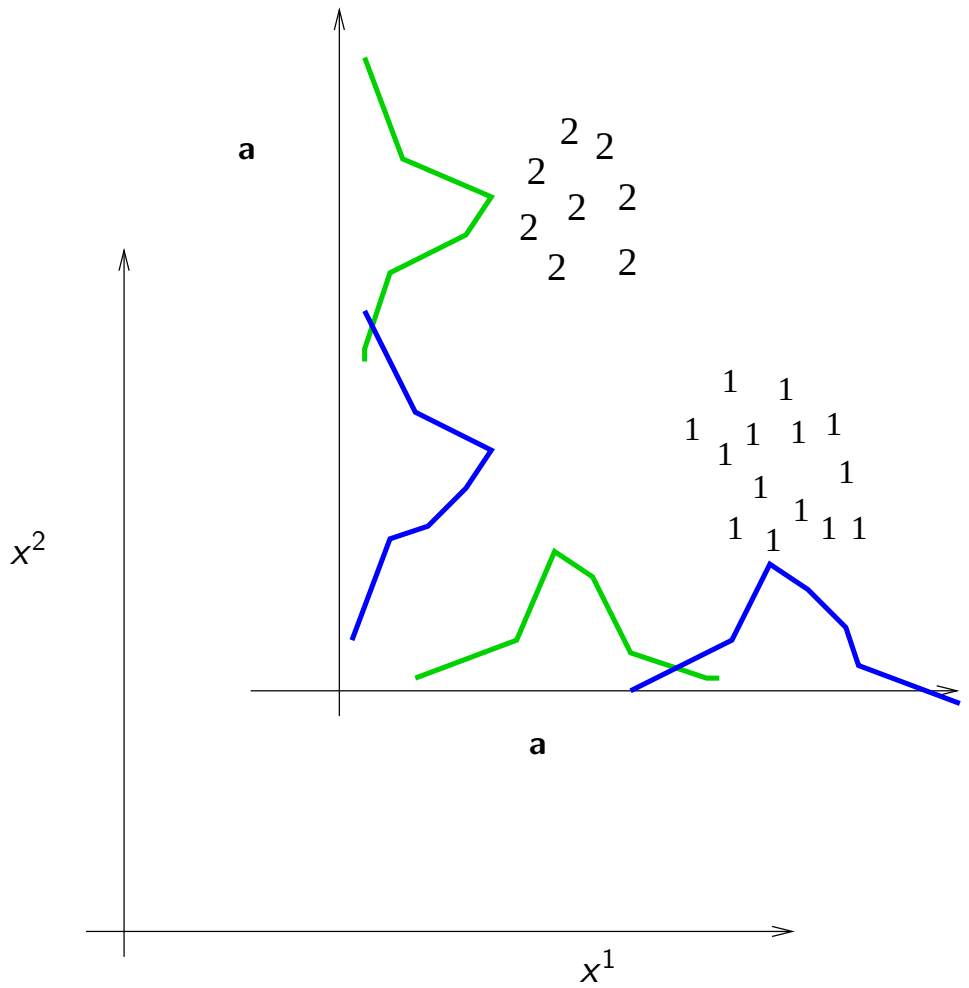
Critères de discrimination inter-classe

- Axe $\mathbf{a}$ où la distance entre les classes (*inter-classe*) est maximum.
- On voit clairement que l'axe doit être aligné avec la droite $(\mathbf{g}_1, \mathbf{g}_2)$ formée par les centres de gravité des classes, car

$$(\hat{\mathbf{g_1}} - \hat{\mathbf{g_2}})^2 = \left(\frac{\mathbf{a}^T \mathbf{g_1}}{||\mathbf{a}||} - \frac{\mathbf{a}^T \mathbf{g_2}}{||\mathbf{a}||} \right)^2 = \left(\frac{\mathbf{a}^T}{||\mathbf{a}||} (\mathbf{g_1} - \mathbf{g_2}) \right)^2$$

qui est maximum quand $\mathbf{a} \simeq \mathbf{g_1} - \mathbf{g_2}$

Critères de discrimination intra-classe



Critères de discrimination intra-classe

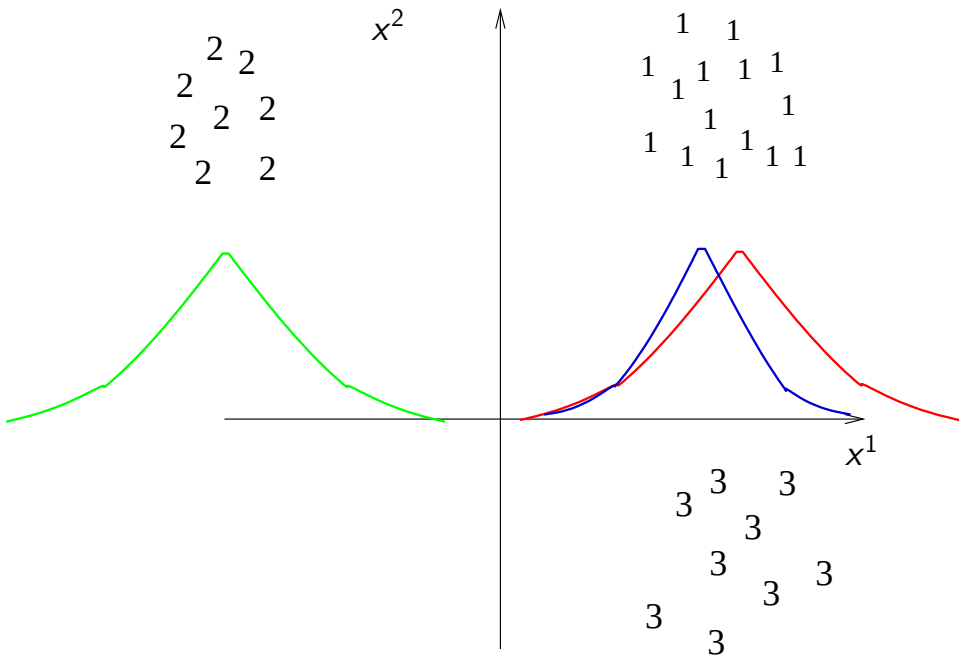
- On doit prendre en compte la variance *intra*-classe des classes projetées
- Le critère de Fisher propose de maximiser

$$\max_{\mathbf{a}} \frac{(\hat{\mathbf{g_1}} - \hat{\mathbf{g_2}})^2}{\sigma_1^2 + \sigma_2^2}$$

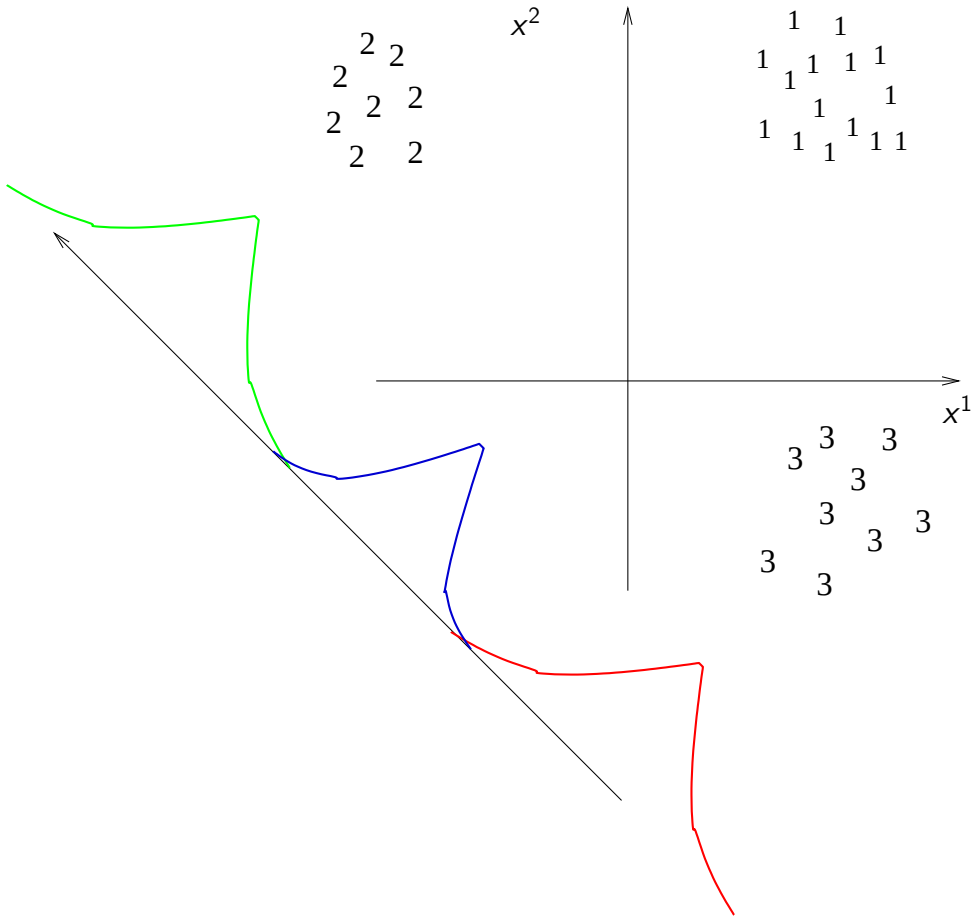
où σ_k est la variance normalisée des points projetés de la classe k

$$\sigma_k^2 = \sum_{\mathbf{x} \in \mathbf{C_k}} (\hat{\mathbf{x}} - \hat{\mathbf{g_k}})^T (\hat{\mathbf{x}} - \hat{\mathbf{g_k}})$$

Axes discriminants



Axes discriminants



Généralisation : critères intra-classe

- Soit $A_k = [\mathbf{x}_1 - \mathbf{g}_k, \ldots, \mathbf{x}_{n_k} - \mathbf{g}_k], x_i \in C_k$ la matrice des données centrées
- $\frac{1}{n_k} A_k A_k^T$ est la matrice de covariance *intra*-classe
- $S_w = \sum_k \frac{1}{n_k} A_k A_k^T$ est la somme des matrices de covariance *intra*-classes
- On veut minimiser

$$\sum_k \sum_{x_i \in c_k} (\hat{\mathbf{x_i}} - \hat{\mathbf{g}_k})^T (\hat{\mathbf{x_i}} - \hat{\mathbf{g}_k}) = \sum_k \sum_{x_i \in c_k} \frac{(\mathbf{a}^T(\mathbf{x_i} - \mathbf{g}_k))^T \mathbf{a}^T(\mathbf{x_i} - \mathbf{g}_k)}{||\mathbf{a}||^2}$$

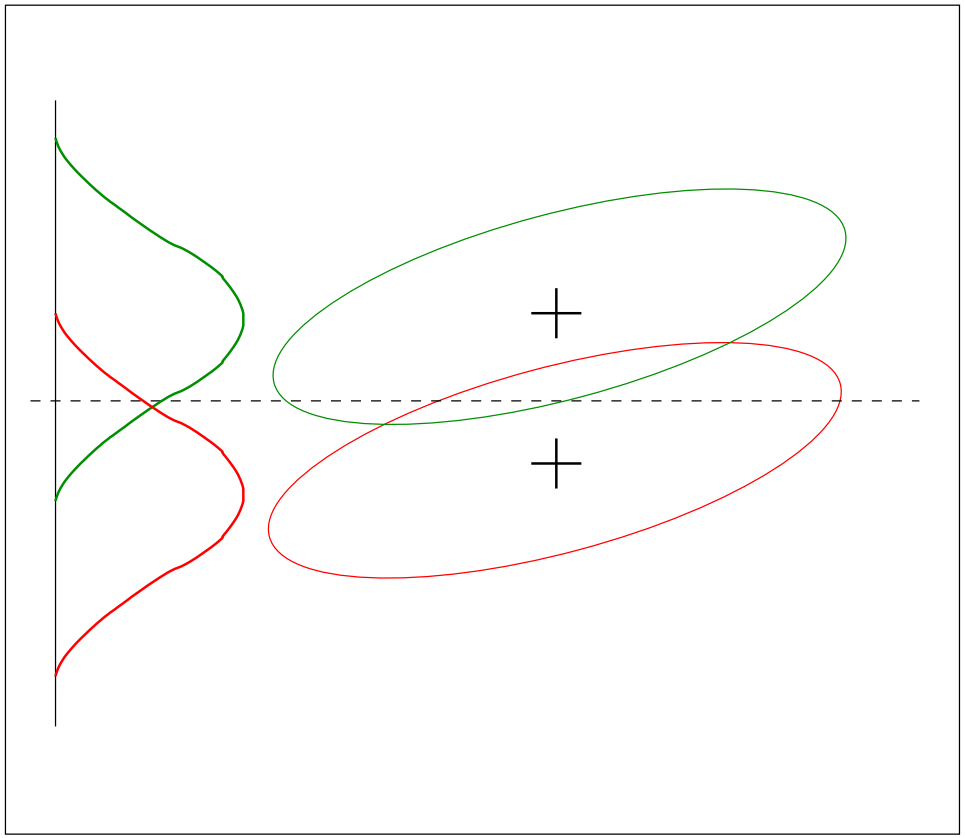
$$= \sum_k \frac{1}{||\mathbf{a}||^2} \mathbf{a}^T A_k A_k^T \mathbf{a} = \frac{1}{||\mathbf{a}||^2} \mathbf{a}^T S_w \mathbf{a}$$

Généralisation : critères inter-classes

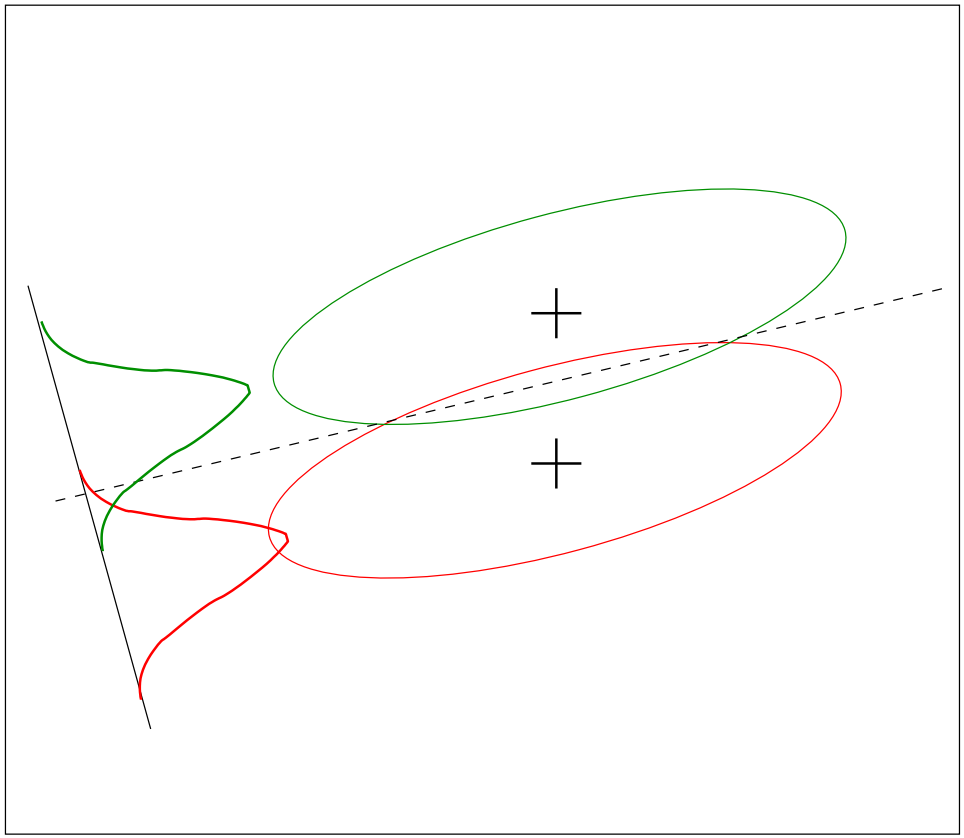
- On note $B = [\mathbf{g}_1 - \mathbf{g}, \ldots, \mathbf{g}_q - \mathbf{g}]$ la matrice des centres des données centrées ($\mathbf{g} = \frac{1}{N}\sum_N \mathbf{x_i}$ et $N = \sum_k n_k$)
- $S_b = \frac{1}{q} BB^\mathsf{T}$ est la matrice de covariance entre les centres de classes
- On veut maximiser

$$\sum_k (\hat{\mathbf{g}}_k - \hat{\mathbf{g}})^\mathsf{T} (\hat{\mathbf{g}}_k - \hat{\mathbf{g}}) = \frac{1}{||\mathbf{a}||^2} \mathbf{a}^\mathsf{T} S_b \mathbf{a}$$

Combinaisons des deux critères



Combinaisons des deux critères



Critères de discrimination de Fisher : coefficient de Raleigh

- On rassemble en un seul critère

$$max_{\mathbf{a}} J_{\mathbf{a}} = max_{\mathbf{a}} \frac{\mathbf{a}^T S_b \mathbf{a}}{\mathbf{a}^T S_w \mathbf{a}}$$

- que l'on obtient si

$$\frac{\partial J_{\mathbf{a}}}{\partial \mathbf{a}} = \frac{S_b\mathbf{a}(\mathbf{a}^T S_w \mathbf{a}) - S_w\mathbf{a}(\mathbf{a}^T S_b \mathbf{a})}{(\mathbf{a}^T S_w \mathbf{a})^2} = 0$$

- on obtient le système propre généralisé : $S_b\mathbf{a} = J_{\mathbf{a}} S_w \mathbf{a}$
- Donc **a** est le premier vecteur propre de $S_w^{-1} S_b$

Sous-espaces discriminants

- Les vecteurs correspondant aux valeur propres λ_i les plus élevées sont les plus discriminants.

$$\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_p \text{ avec } \lambda_1 > \lambda_2 > \ldots \lambda_p$$

- q classes se discriminent dans un sous-espace de dimension $q-1$ au maximum

$\Rightarrow$ seules $q-1$ v.p. sont différentes de zéros

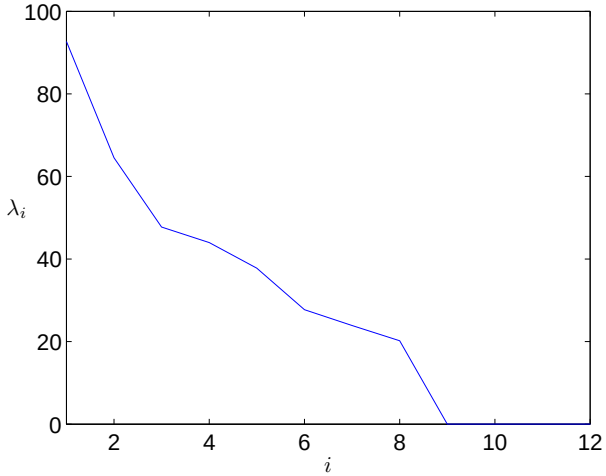
Cas particulier : 2 classes

- $BB^{\mathsf{T}} = (\mathbf{g}_1 - \mathbf{g})(\mathbf{g}_1 - \mathbf{g})^{\mathsf{T}} + (\mathbf{g}_2 - \mathbf{g})(\mathbf{g}_2 - \mathbf{g})^{\mathsf{T}} = (\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)^{\mathsf{T}}$
- donc $BB^{\mathsf{T}}\mathbf{a}$ est un vecteur dans la direction de $(\mathbf{g}_1 - \mathbf{g}_2)$
- donc on aurait $\mathbf{a} \simeq S_w^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$

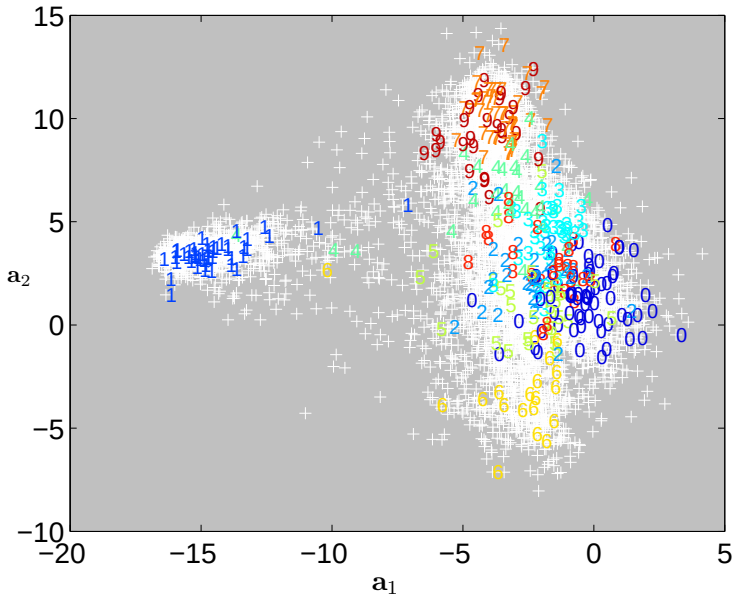
Illustrations : reconnaissance de caractères

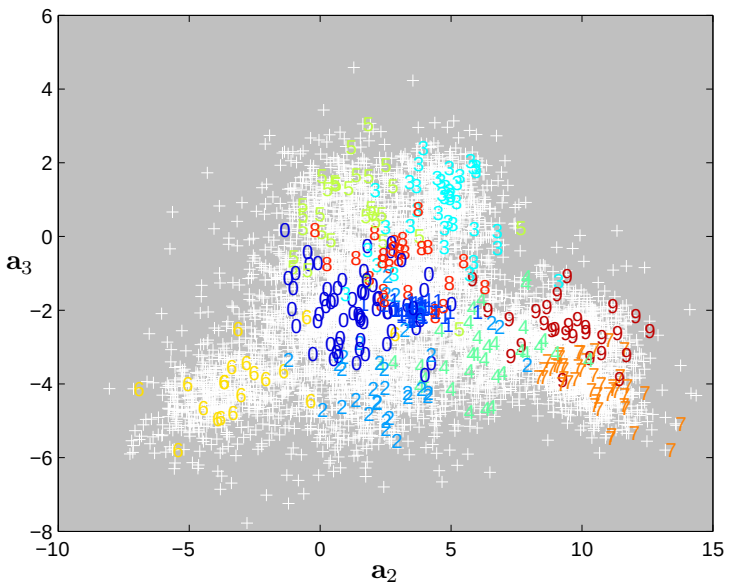
Rappel : 7291 images 16×16 (8 bits) de chiffres de 0 à 9

$\Rightarrow \{\mathbf{x}_i, y_i\}$ avec $\mathbf{x}_i \in \mathbb{R}^{256}$ et $y_i = 1, \ldots, 10$, $i = 1 \ldots 7291$



Projection des données





$\Rightarrow$ L'ADL détermine automatiquement le sous-espace optimal pour séparer linéairement les données selon les labels y_i.

Analyse Discriminante Décisionnelle

- Nouvelle donnée $j \rightarrow \mathbf{x}_j$ connus, mais y_j inconnus.
- A quelle classe C_k doit être affectée j ?

$\Rightarrow$ Calcul de $P(C_k|\mathbf{x}_j)$ (loi de Bayes) :

$$P(C_k|\mathbf{x}_j) = \frac{p(\mathbf{x}_j|C_k)P(C_k)}{p(\mathbf{x}_j)}$$

Approximation gaussienne

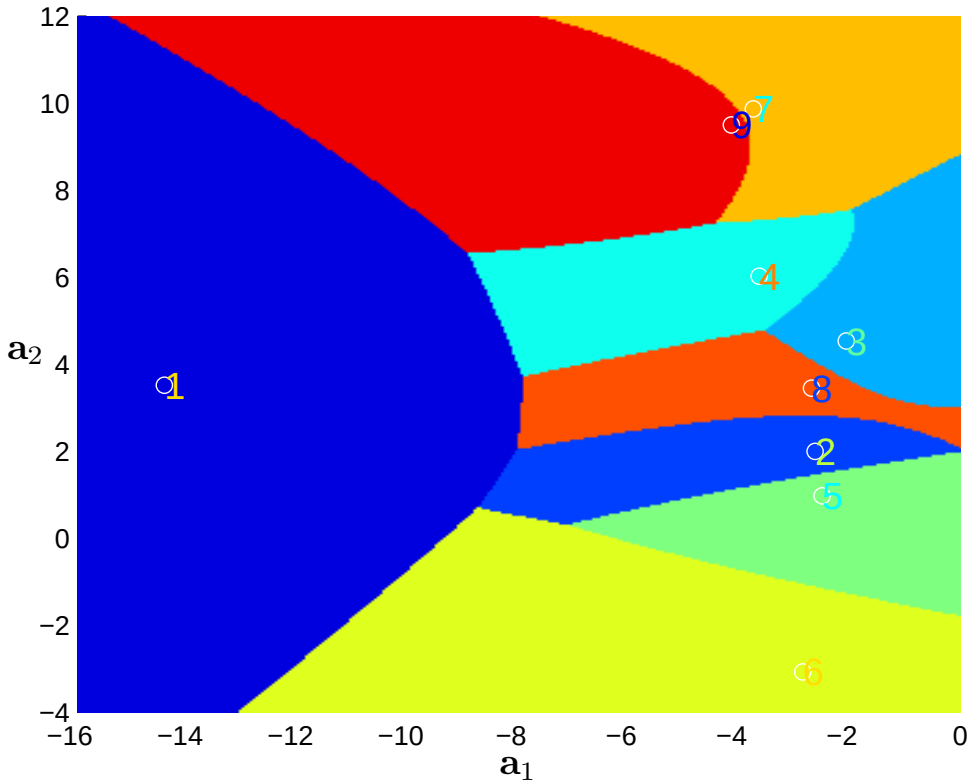
- Chaque classe est modélisée par $\mathcal{N}(\mu_k, W_k)$
- Probabilité a priori $P(C_k) = 1/q$
- Évidence $p(\mathbf{x})$ pas pris en compte

$\Rightarrow$ Maximisation de la vraissemblance

$$p(\mathbf{x}|C_k) \approx \exp\left(-(\mathbf{x} - \mu_k)^{\mathsf{T}} W_k (\mathbf{x} - \mu_k)\right)$$

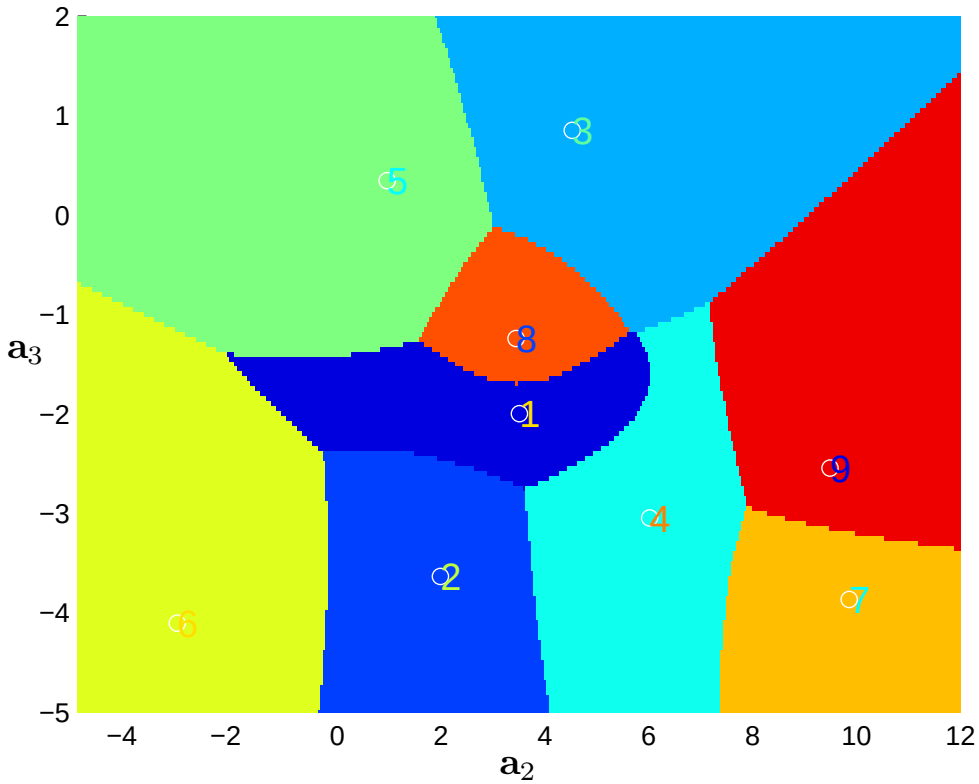
Décision (classification)

$$\delta(\mathbf{x}) = \arg\max_k p(x|C_k)$$



Décision (classification)

$$\delta(\mathbf{x}) = \arg\max_k p(x|C_k)$$



Optimalité de la ADL

- Comme pour l'ACP, l'ADL est optimale dans le cas ou les q classes suivent une distribution gaussienne

$\Rightarrow$ Critère de séparation basé sur les matrices de covariance S_w et S_b

- La discrimination linéaire $\rightarrow$ ne prend pas en compte les relations **non-linéaires** existant entre les variables (idem ACP)

Modélisation des données

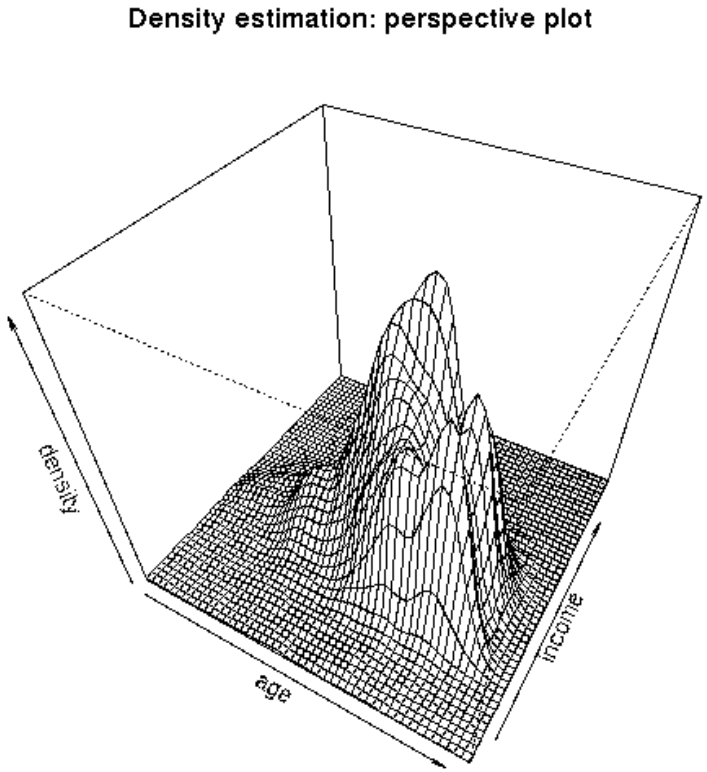
- Jusque ici, nous avons vu des analyses des données basées sur une analyse implicite de la distribution des données
  - ACP, AFC $\rightarrow$ Espace maximisant la variance de la distribution des données
  - ADL $\rightarrow$ Espace maximisant la variance inter-classe, minimisant l'intra-classe

$\Rightarrow$ On ne cherche pas la forme de la distribution, mais des axes où elle s'explicite le mieux

- Une alternative est de donner une formulation explicite à la densité des données dans leur espace.
- On note $f(\mathbf{x}) : \mathcal{F} \rightarrow \mathbb{R}$ cette densité.

Estimation de densité

- Méthodes des plus proches voisins (knn)
- Noyaux de Parzen, réseaux RBF
- Histogrammes
- **Modèle de mélange**



Mélange de densités

Définition

- La densité $f(x)$ est générée par un nombre c de fonctions de "base"

$$f(x) = \sum_{j=1}^{c} \pi_j \phi(x, \boldsymbol{\theta}_j)$$

- π_j sont les paramètres de proportion du mélange
- $\phi(x, \boldsymbol{\theta}_j)$ sont des fonctions paramétrées par $\boldsymbol{\theta}_j$

Hypothèses

1. Le nombre de composantes c est a priori connu
2. La forme de ϕ est connue

Interprétation probabiliste

- La densité $f(x)$ représente un processus aléatoire où x est tiré sur un ensemble d'états ω_j avec la probabilité à priori $P(\omega_j)$

$$f(x) = p(x|\boldsymbol{\theta}) = \sum_{j=1}^{c} p(x|\omega_j, \boldsymbol{\theta}_j) P(\omega_j)$$

- Par identification $\pi_j = P(\omega_j)$ est donc $\sum_j \pi_j = 1$
- $\phi(x, \boldsymbol{\theta}_j)$ est la proba. conditionnelle que x soit généré par ω_j

Mélange de Gaussiennes

La fonction ϕ représentant une densité probabilité, le choix de la loi normale $\mathcal{N}(\mu, \Sigma)$ parait judicieux :

$$f(x) = \sum_{j=1}^{c} \pi_j \mathcal{N}(\mu_j, \Sigma_j)$$

- Peut approximer toutes autres densités
- Permet une résolution linéaire par Maximum de Vraisemblance (MV) de ses paramètres (log-vraissemblance)

Maximum de vraisemblance

- Soit $\Omega = \{x_1, \ldots, x_N\}$ échantillons non-labelés générés par le le mélange de densités $f(x) = p(x|\boldsymbol{\theta})$.
- $\boldsymbol{\theta} = \{\pi_j, \mu_j, \Sigma_j\}$ à déterminer.
- Vraisemblance :

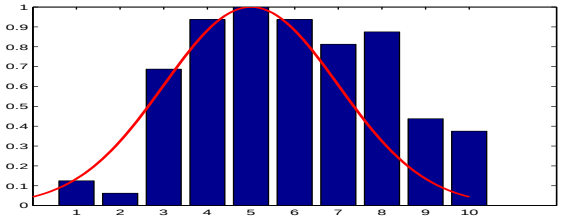
$$p(\Omega|\boldsymbol{\theta}) = \prod_i^N p(x_i|\boldsymbol{\theta})$$

- Estimation : $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\Omega|\boldsymbol{\theta})$
- ou maximisation de la log-vraisemblance

$$l(\boldsymbol{\theta}, \Omega) = \sum_{i=1}^N \log p(x_i|\boldsymbol{\theta}) = \sum_{i=1}^N \log \left[\sum_{j=1}^c \pi_j \phi(x_i, \boldsymbol{\theta}_j) \right]$$

Cas trivial : une composante, $c = 1$

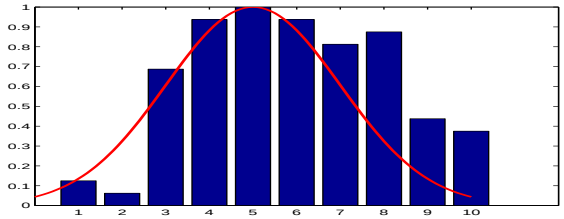
- $\boldsymbol{\theta} = \{\mu, \Sigma\}$



$$\max_{\boldsymbol{\theta}} \sum_i \log e^{-(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$$

Cas trivial : une composante, $c = 1$

- $\boldsymbol{\theta} = \{\mu, \Sigma\}$



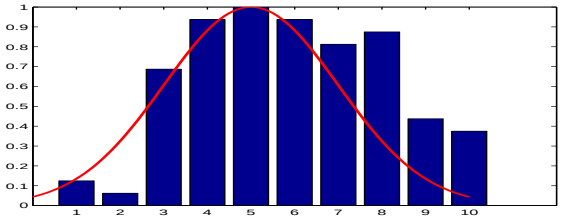
$$\max_{\boldsymbol{\theta}} \sum_i \log e^{-(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$$

$$\Longleftrightarrow$$

$$\min_{\boldsymbol{\theta}} \sum_i (x_i - \mu)^T \Sigma^{-1}(x_i - \mu)$$

Cas trivial : une composante, $c = 1$

- $\boldsymbol{\theta} = \{\mu, \Sigma\}$



$$\max_{\boldsymbol{\theta}} \sum_i \log e^{-(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}$$

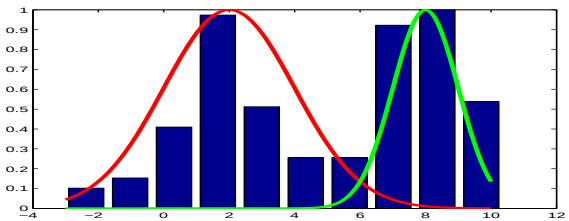
$$\Longleftrightarrow$$

$$\min_{\boldsymbol{\theta}} \sum_i (x_i - \mu)^T \Sigma^{-1}(x_i - \mu)$$

- $\hat{\mu} = \frac{1}{N}\sum_i x_i$
- $\hat{\Sigma} = \frac{1}{N}\sum_i (x_i - \mu)(x_i - \mu)^T$

Plus difficile : deux composantes !

$$\begin{aligned}\boldsymbol{\theta} &= \{\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\} \\ &= \{\pi_1, \pi_2, \mu_1, \Sigma_1, \mu_2, \Sigma_2\}\end{aligned}$$



$$l(\boldsymbol{\theta}, \Omega) = \sum_{i}^{N} \log\left[(1-\pi)\phi(x_i, \boldsymbol{\theta}_1) + \pi\phi(x_i, \boldsymbol{\theta}_2)\right]$$

$\Rightarrow$ plus complexe à maximiser à cause de la somme à l'intérieur du log !

$\Rightarrow$ Solution : algorithme itératif en 2 temps pour maximiser l

$\rightarrow$ algorithme **Expectation-Maximisation**

Algorithme EM (2 composantes)

- Ce qu'il manque au problème, c'est l'appartenance d'un point x_i à une des deux composantes.

$\Rightarrow$ Si cette appartenance était connue, on en revient au problème à une composante !

- EM consiste à introduire des variables inconnues, ou manquantes : l'appartenance $\Delta_i \in \{0, 1\}$ pour chaque x_i à une composante :

$$x_i \backsim \phi_1 \text{ si } \Delta_i = 0, x_i \backsim \phi_2 \text{ si } \Delta_i = 1$$

Étape Expectation

- Supposons une valeur initiale $\boldsymbol{\theta}^0$
- On peut calculer la contribution de chaque point x_i d'appartenir à l'une ou l'autre densité :

$$\begin{aligned}\gamma_i(\boldsymbol{\theta}^0) &= E[\Delta_i|\boldsymbol{\theta}^0, \Omega] \\ &= \frac{\pi\phi(x_i, \boldsymbol{\theta}_2^0)}{(1-\pi)\phi(x_i\boldsymbol{\theta}_1^0) + \pi\phi(x_i, \boldsymbol{\theta}_2^0)}\end{aligned}$$

- γ_i est appelée la responsabilité.
- Elle est égale à l'espérance (= expectation) de Δ_i sur toutes les composantes.

Responsabilité et assignement doux (soft-assignment)

- La valeur de γ_i pourrait permettre de déterminer $\Delta_i \Rightarrow x_i$ automatiquement assigné à ϕ_1 ou ϕ_2

$\Rightarrow$ Hard-decision, du type k-means. Un point appartient à un et un seul cluster !

L'algorithme EM est moins radical. Un point peut contribuer à plusieurs clusters (densités). On parle alors de soft-assignment

Étape Maximisation

- La responsabilité de chaque point étant calculée, il est maintenant possible d'estimer les paramètres par maximum de vraisemblance *pondéré* :

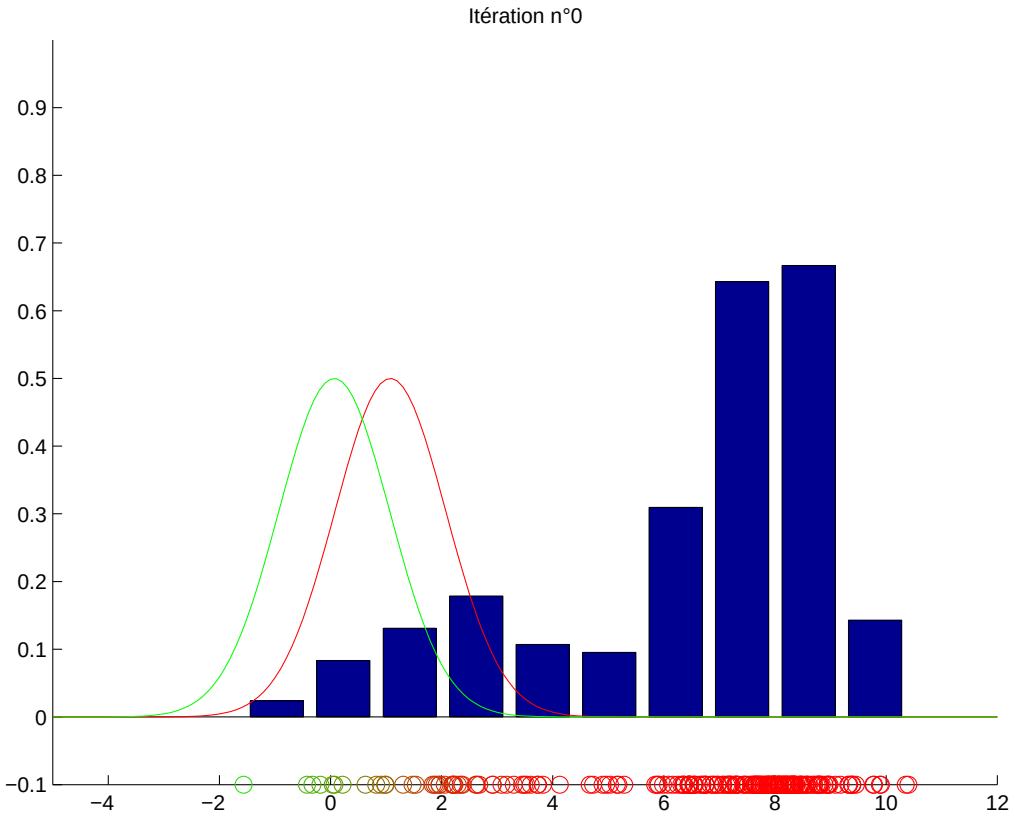
$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1-\gamma_i)x_i}{\sum_{i=1}^{N}1-\gamma_i} \qquad \hat{\Sigma}_1 = \frac{\sum_{i=1}^{N}(1-\gamma_i)(x_i-\mu_1)(x_i-\mu_1)^{\mathsf{T}}}{\sum_{i=1}^{N}1-\gamma_i}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N}\gamma_i x_i}{\sum_{i=1}^{N}\gamma_i} \qquad \hat{\Sigma}_2 = \frac{\sum_{i=1}^{N}\gamma_i(x_i-\mu_1)(x_i-\mu_1)^{\mathsf{T}}}{\sum_{i=1}^{N}\gamma_i}$$

- Proportion du mélange : $\pi = \sum_{i=1}^{N}\gamma_i/N$

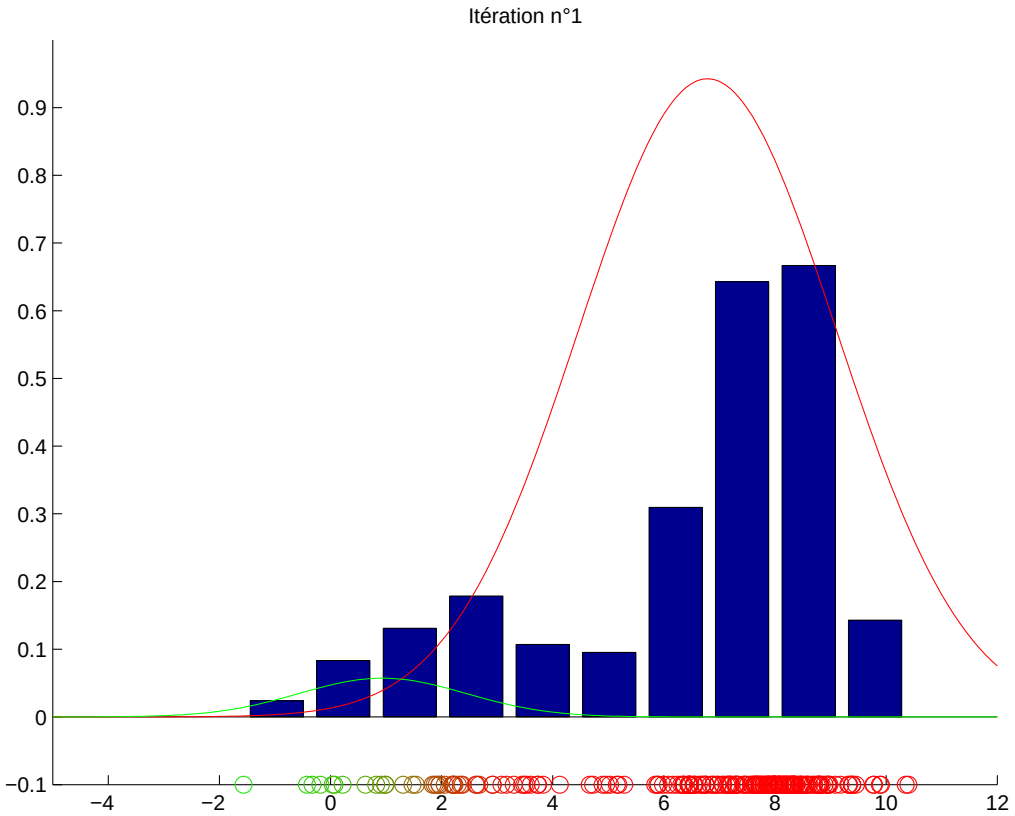
Illustration

Itérations successives de l'étape E et de l'étape M



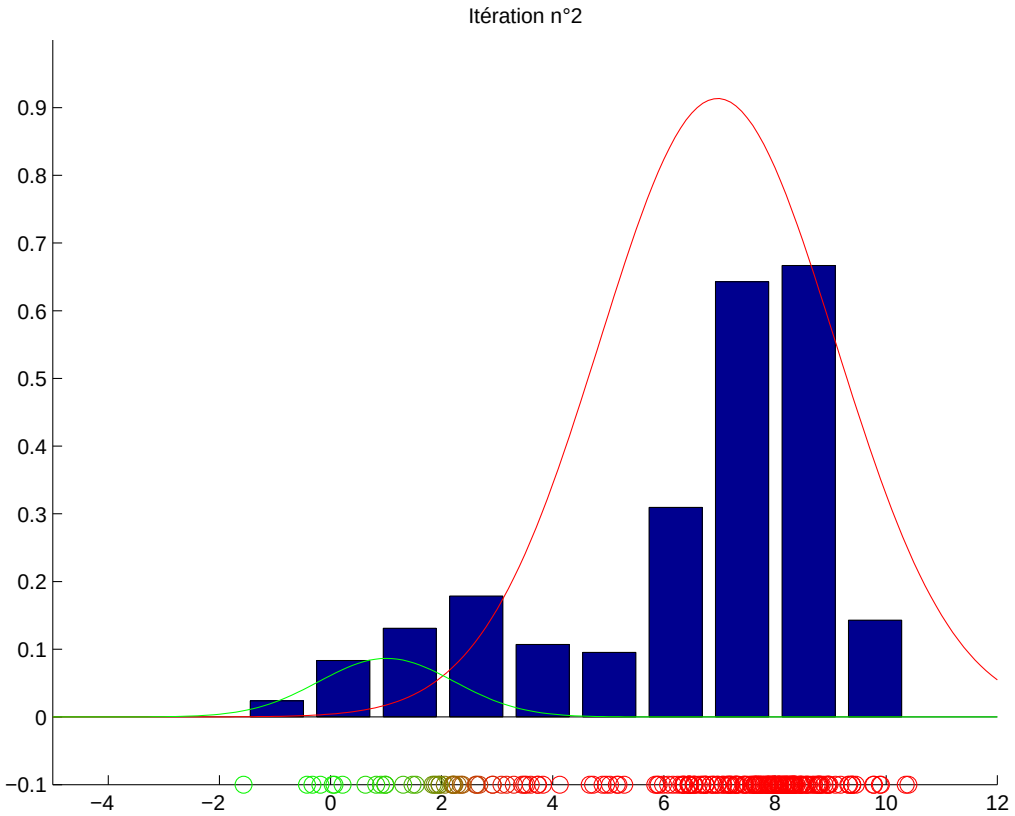
Illustration

Itérations successives de l'étape E et de l'étape M



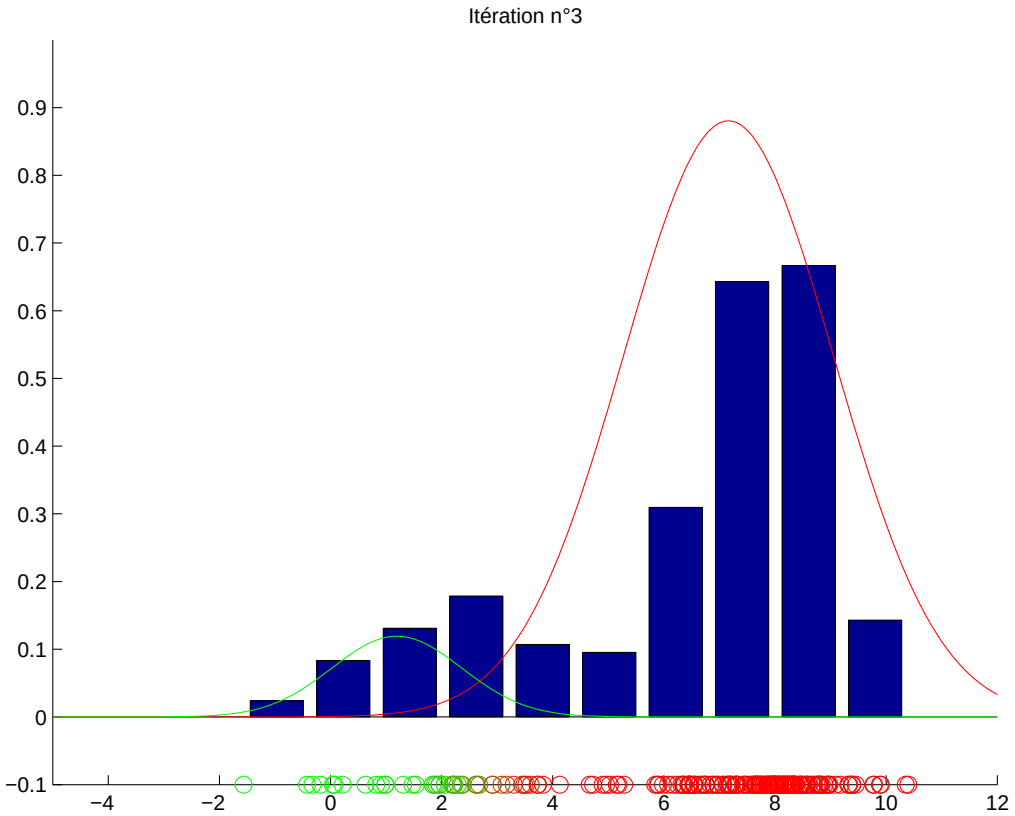
Illustration

Itérations successives de l'étape E et de l'étape M



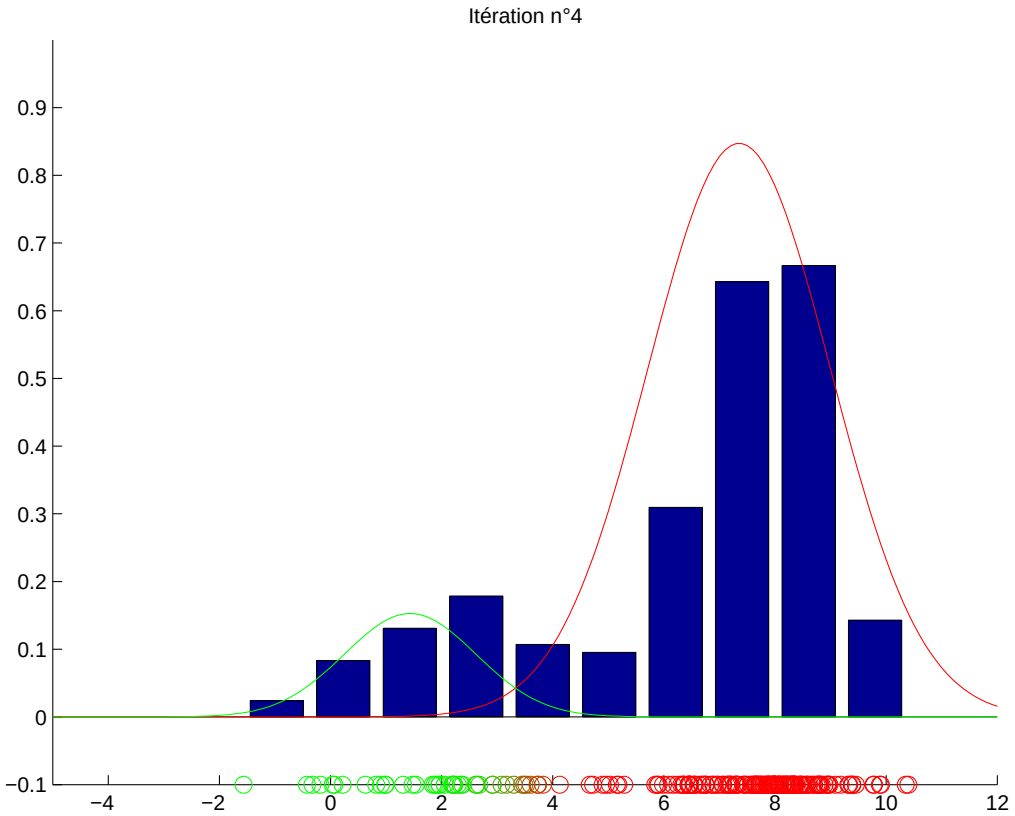
Illustration

Itérations successives de l'étape E et de l'étape M



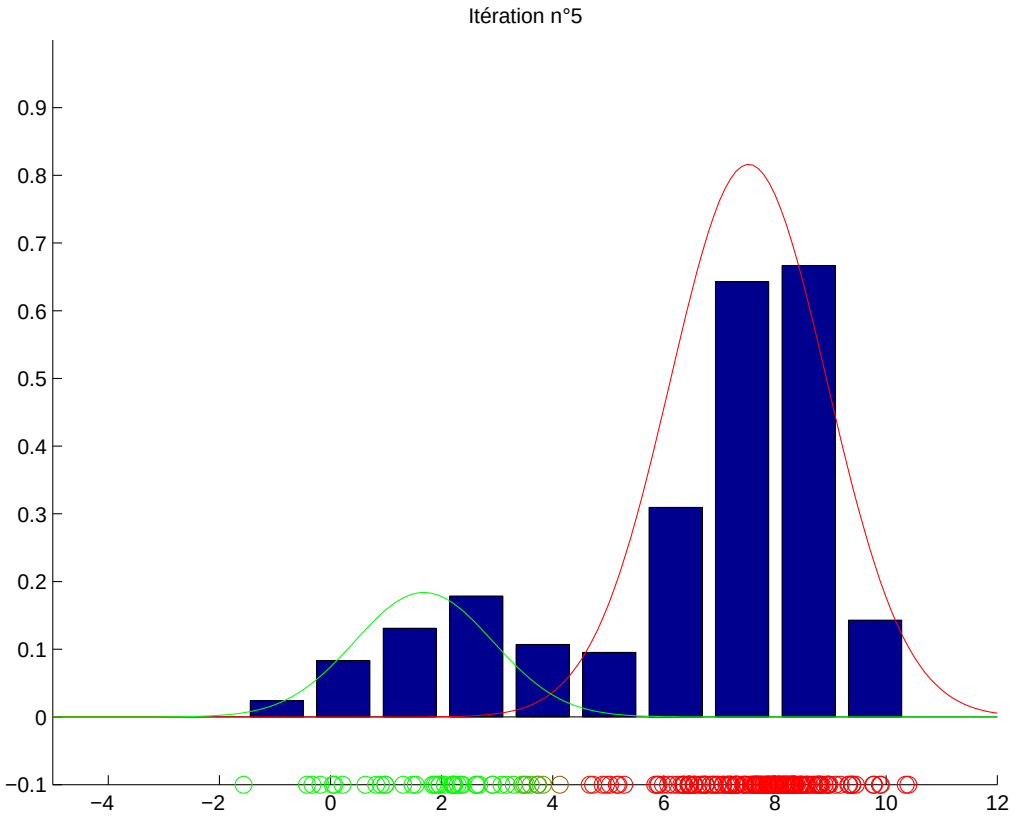
Illustration

Itérations successives de l'étape E et de l'étape M



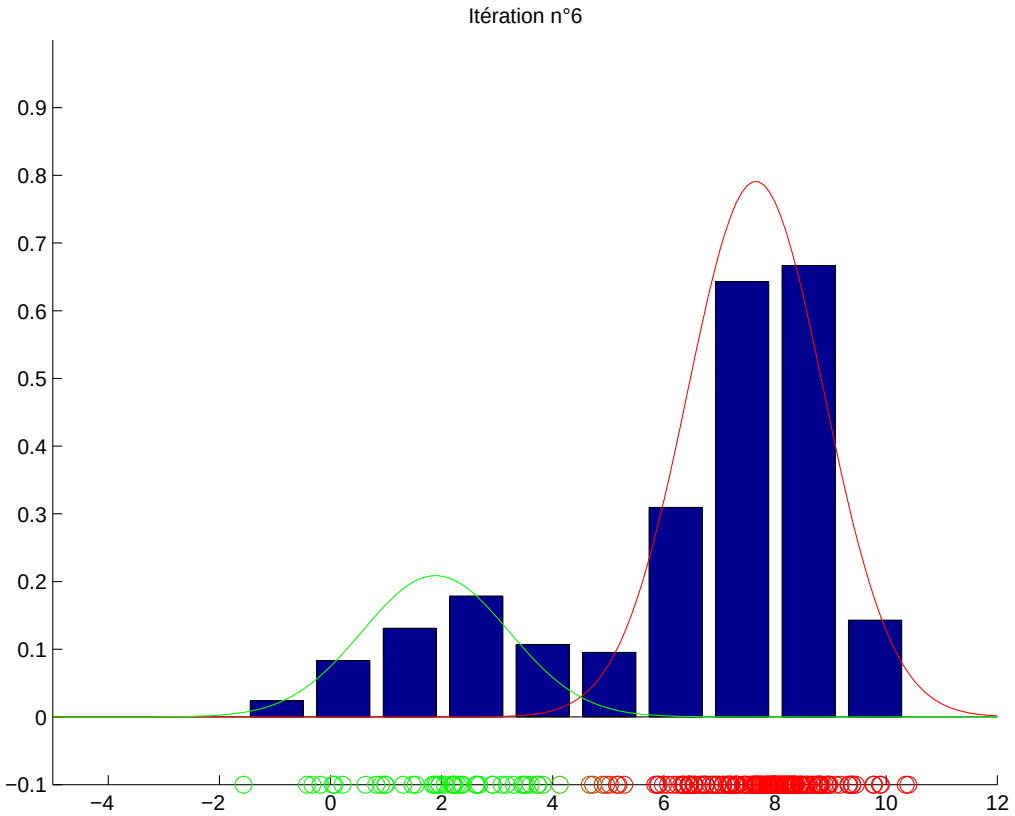
Illustration

Itérations successives de l'étape E et de l'étape M



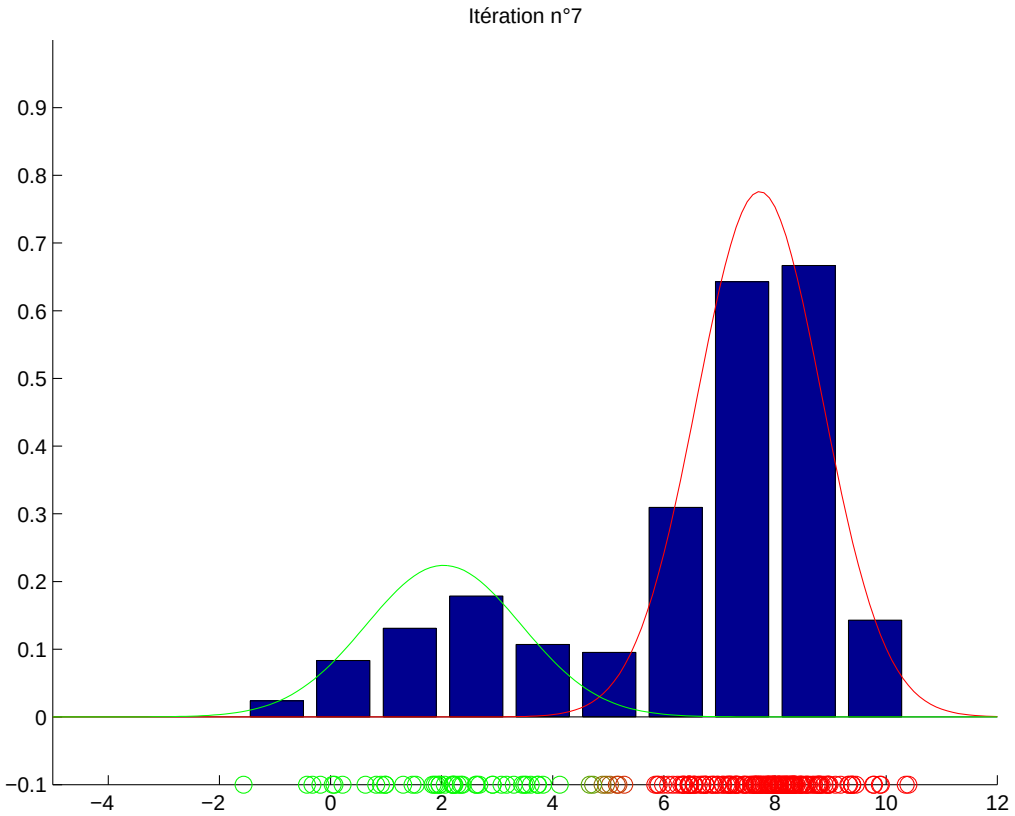
Illustration

Itérations successives de l'étape E et de l'étape M



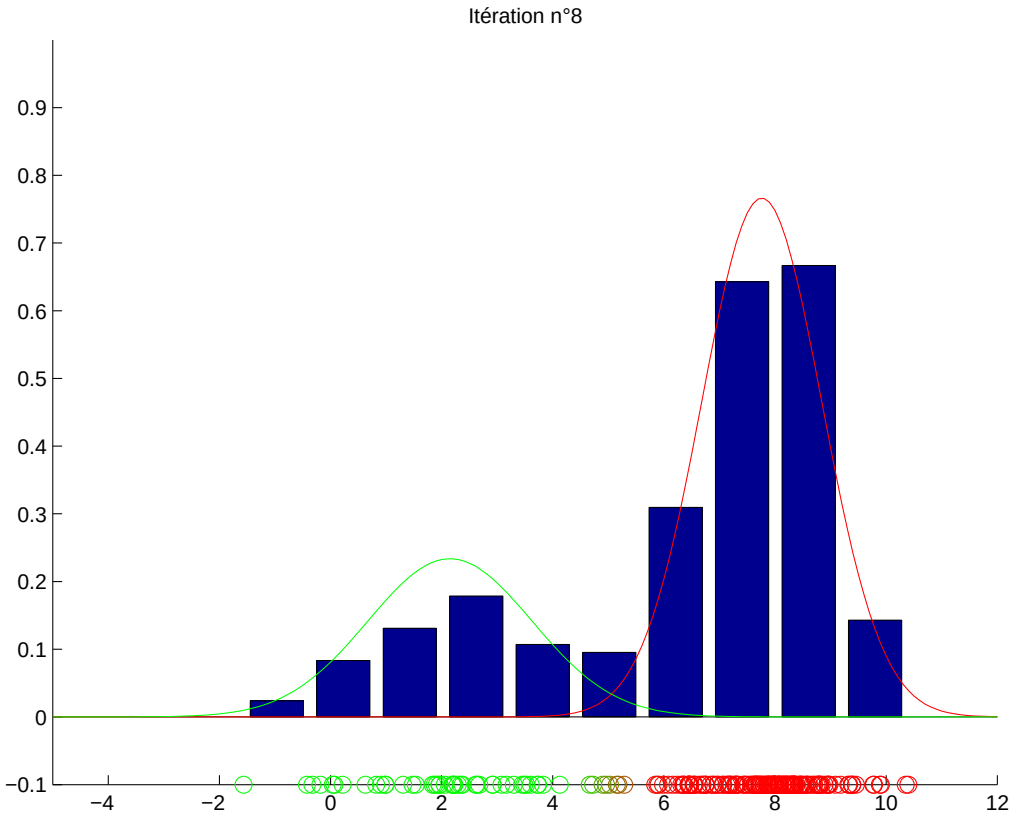
Illustration

Itérations successives de l'étape E et de l'étape M



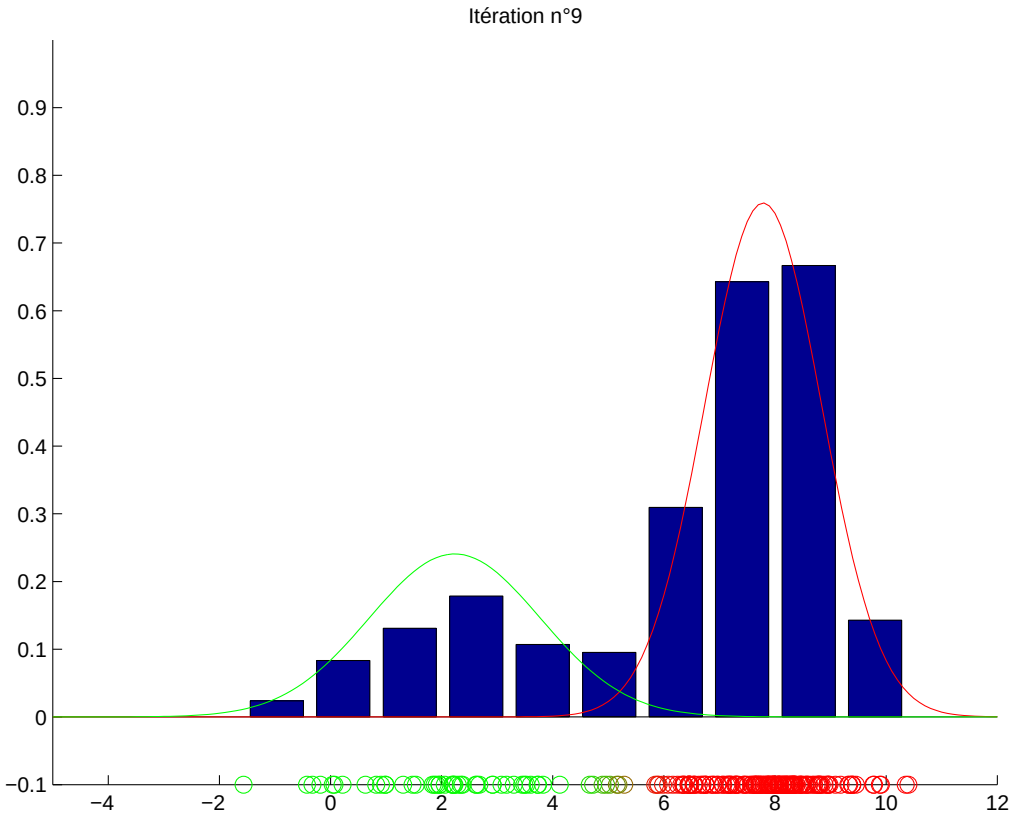
Illustration

Itérations successives de l'étape E et de l'étape M



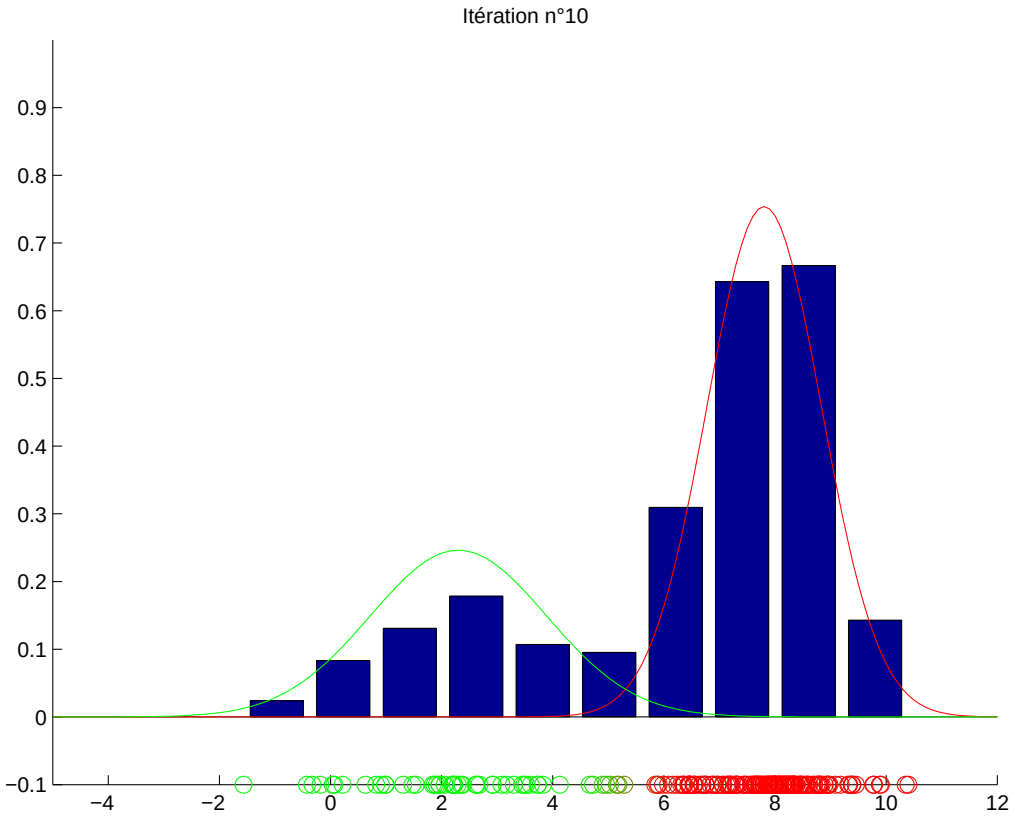
Illustration

Itérations successives de l'étape E et de l'étape M



Illustration

Itérations successives de l'étape E et de l'étape M



Résultats

- Paramètres réels

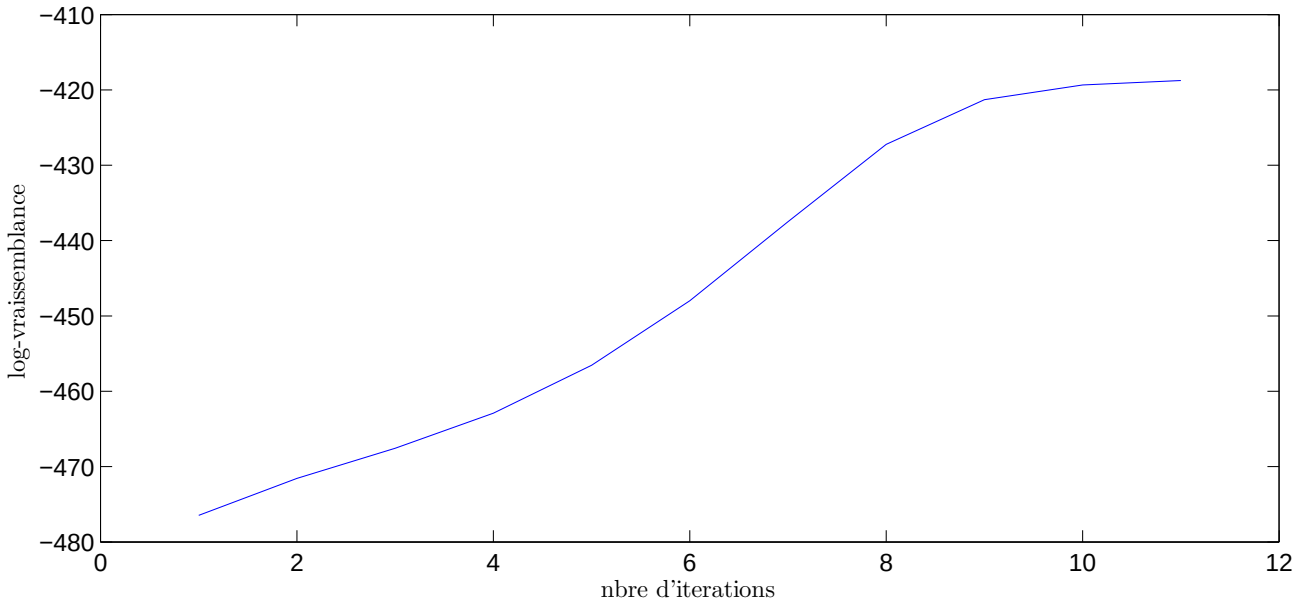
π	μ_1	σ_1	μ_2	σ_2
0.75	2	2	8	1

- Paramètres estimés

$\hat{\pi}$	$\hat{\mu_1}$	$\hat{\sigma_1}$	$\hat{\mu_2}$	$\hat{\sigma_2}$
10 itérations				
0.76	2.17	1.56	7.91	1.06
20 itérations				
0.76	2.15	1.98	8.01	0.98

Itérations

Alternance du cycle Expectation-Maximisation $\rightarrow$ augmentation de la vraisemblance entre le modèle de mélange et les données



Le processus est itéré jusqu'à convergence, c.a.d. lorsque la vraisemblance ne croît plus (ou très peu)

Limitations et détails pratiques

- Dépendant des paramètres initiaux
- Convergence plus ou moins lente selon les distributions observées
- Pb. des **maxima locaux**

Les mélanges à c composantes

- Généralisation du problème à deux composantes avec :

$$\Delta_{i1}, \Delta_{i2}, \ldots, \Delta_{ik}, \ldots, \Delta_{ic}$$

indicatrices pour la donnée i d'être générée par la composante k

$\Rightarrow$ Responsabilité γ_{ik} : espérance de Δ_{ik} sur **l'ensemble** des composantes

- $\{\pi_k, \mu_k, \Sigma_k\}_{k=1,\ldots,c}$ inconnues à déterminer

Algorithme EM, cas général à c composantes

1. Définir les paramètres initiaux $\boldsymbol{\theta}^0 = \{\pi_k^0, \mu_k^0, \Sigma_k^0\}_{k=1,\ldots,c}$
   - En général $\pi_k = 1/c$, μ_k aléatoire et $\Sigma_k = \mathbf{Id}$
   - Alternative : utiliser k-means pour l'initialisation
2. **Étape E** : calcul des responsabilités pour chaque point $i = 1, \ldots, N$ et chaque composante $k = 1, \ldots, c$

$$\gamma_{ik} = \frac{\pi_k \phi(x_i, \boldsymbol{\theta}_k)}{\sum_{j=1}^{c} \pi_j \phi(x_i, \boldsymbol{\theta}_j)}$$

3. **Étape M** Estimation des paramètres du mélange

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} x_i}{\sum_i \gamma_{ik}}; \qquad \Sigma_k = \frac{X_k \Gamma_k X_k^{\mathsf{T}}}{\mathrm{Tr}(\Gamma_k)}; \qquad \pi_k = \frac{\sum_i \gamma_{ik}}{N}$$

avec $\Gamma_k = \mathrm{diag}[\gamma_{1k}, \ldots, \gamma_{Nk}]$, X_k données centrées sur μ_k

4. Itérer étapes 2 et 3 jusqu'à convergence

Choix des modèles

- La paramétrisation *a priori* du mélange à des conséquences importantes sur la convergence de EM
- Flexibilité des paramètres
  1. Le nombre de composantes c
  2. La forme des matrices de covariance Σ_k
- Un mauvais ajustement de ces paramètres entraîne une mauvaise estimation du mélange, ou pas de convergence du tout !
- Nbre de variables : $p \times p \times c + 2 \times c$: si N faible, p grand et c grand $\rightarrow$ problème de *sous-détermination*

Choix de la covariance

Problème sous-déterminé

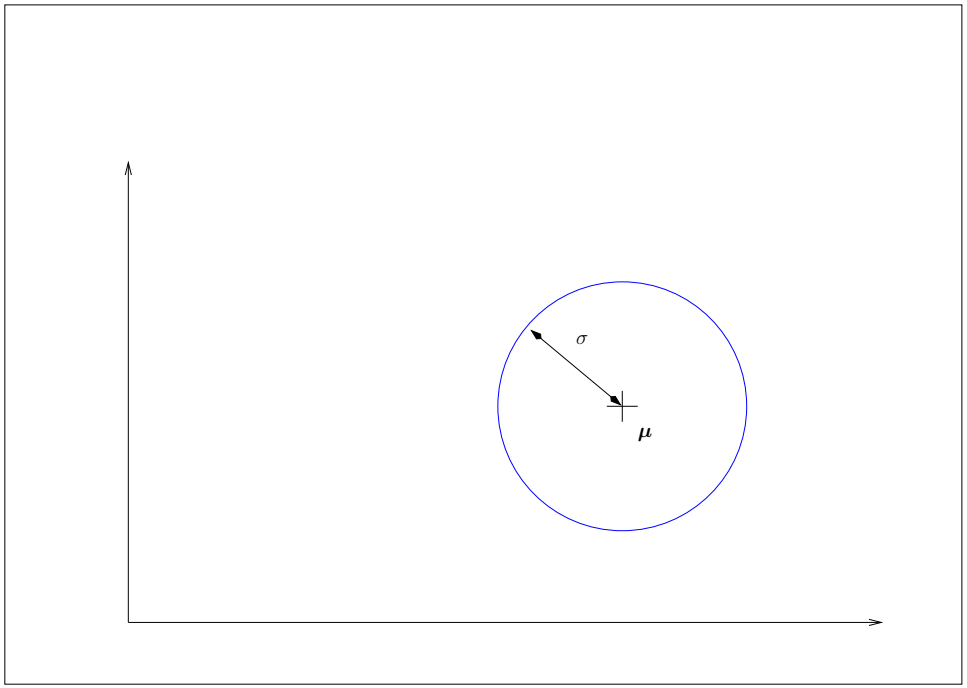
- $\Sigma \in \mathbb{R}^{p \times p}$
- Exemple de la reco. de caractères $\mathbf{x}_i \in \mathbb{R}^{256}$

$\Rightarrow$ Estimation de $256^2 \times c$ paramètres de covariance (pour à peu près 7000 éléments) !

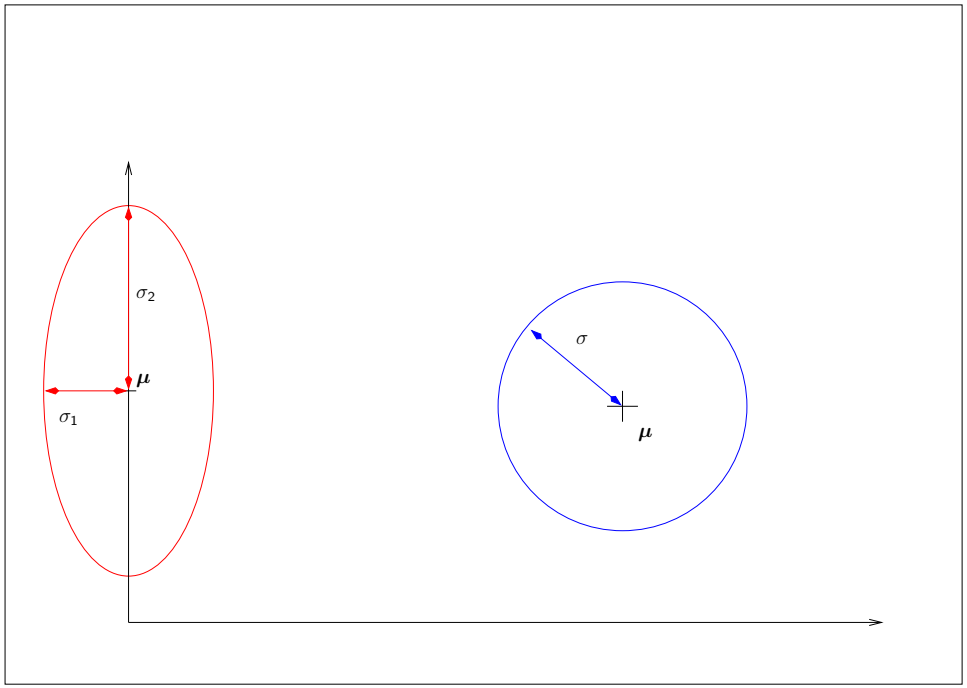
Simplification de la matrice

- Modèles sphériques $\Sigma = \sigma * \mathbf{Id}$, 1 paramètres
- Modèles diagonaux $\Sigma = \text{diag}[\sigma_1, \ldots, \sigma_p]$, p paramètres
- Modèles complets $\Sigma \in \mathbb{R}^{p \times p}$, p^2 paramètres

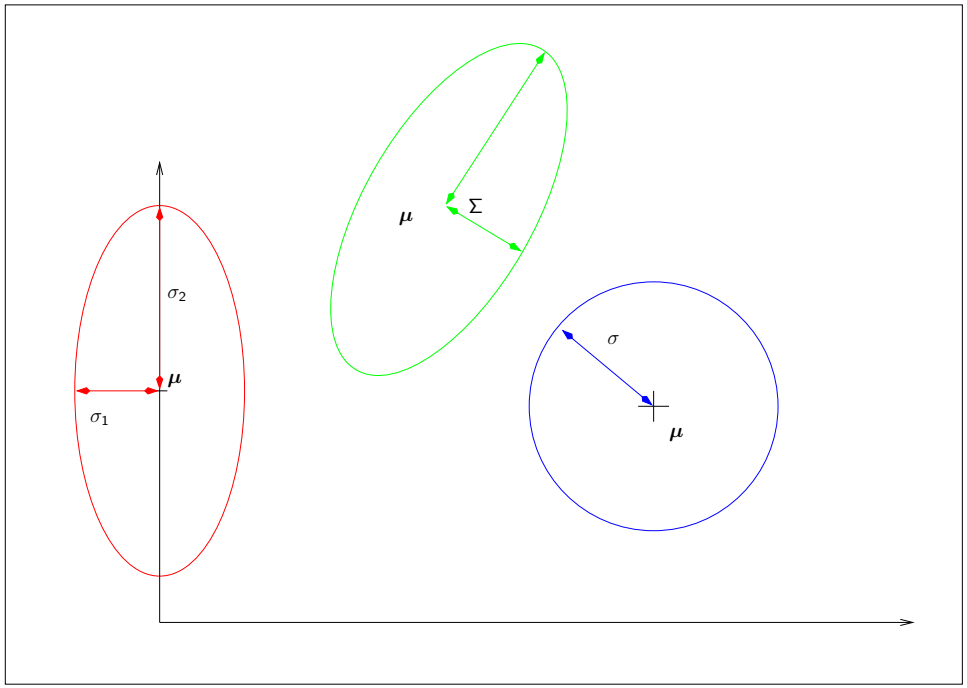
Exemple en 2D



Exemple en 2D

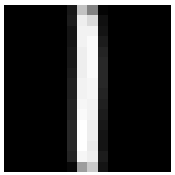
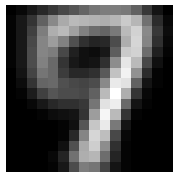
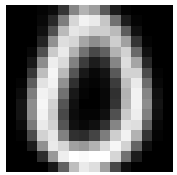
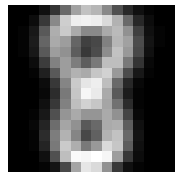
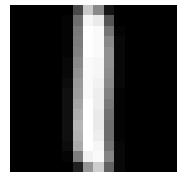


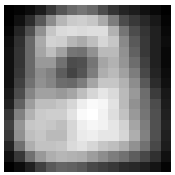
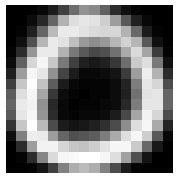
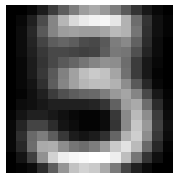
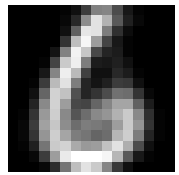
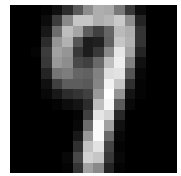
Exemple en 2D



Exemple sur la reconnaissance de caractères

10 composantes, modèle sphérique

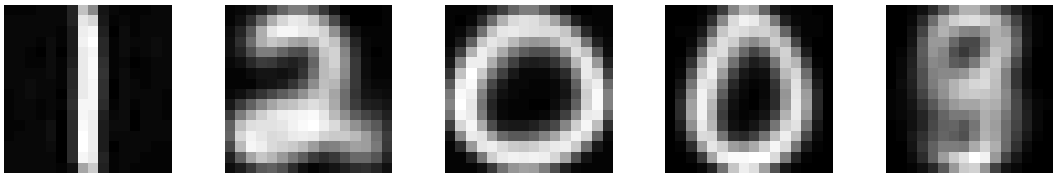
$\pi_1 = 0.0741$ 
$\pi_2 = 0.119$ 
$\pi_3 = 0.0647$ 
$\pi_4 = 0.0879$ 
$\pi_5 = 0.0577$ 

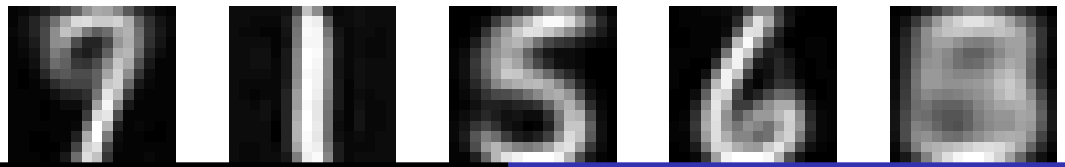
$\pi_6 = 0.187$ 
$\pi_7 = 0.0795$ 
$\pi_8 = 0.139$ 
$\pi_9 = 0.0882$ 
$\pi_{10} = 0.103$ 

Pré-processing : réduction de dimension par ACP

Rappel : on a vu que 50 composantes principales permettent de reconstruire 90% du signal

- EM dans l'espace des 50 premières CP

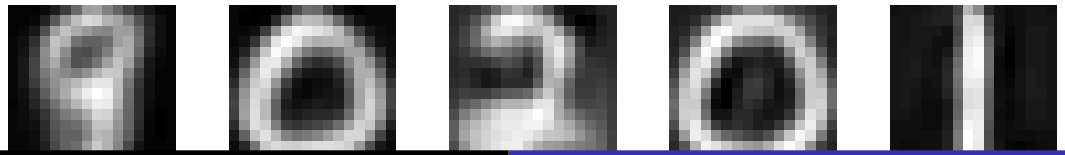




Pré-processing : réduction de dimension par ACP

Rappel : on a vu que 50 composantes principales permettent de reconstruire 90% du signal

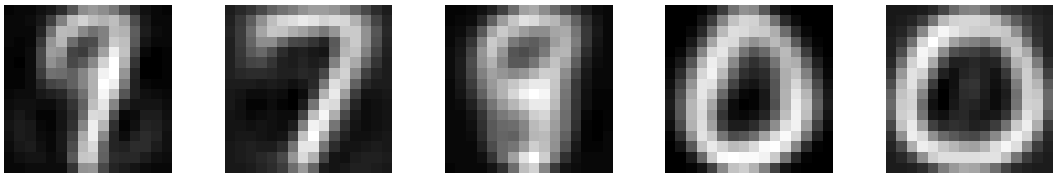
- EM dans l'espace des 10 premières CP

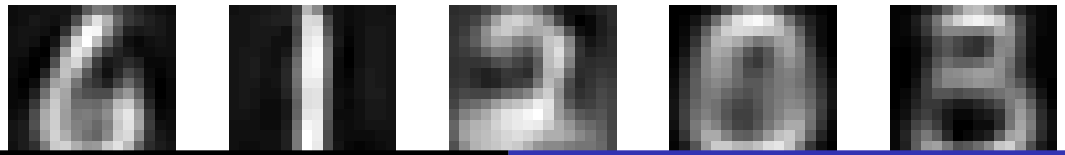


Pré-processing : réduction de dimension par ACP

Rappel : on a vu que 50 composantes principales permettent de reconstruire 90% du signal

- EM dans l'espace des 2 premières CP





Choix du nombre de composantes

Parcimonie

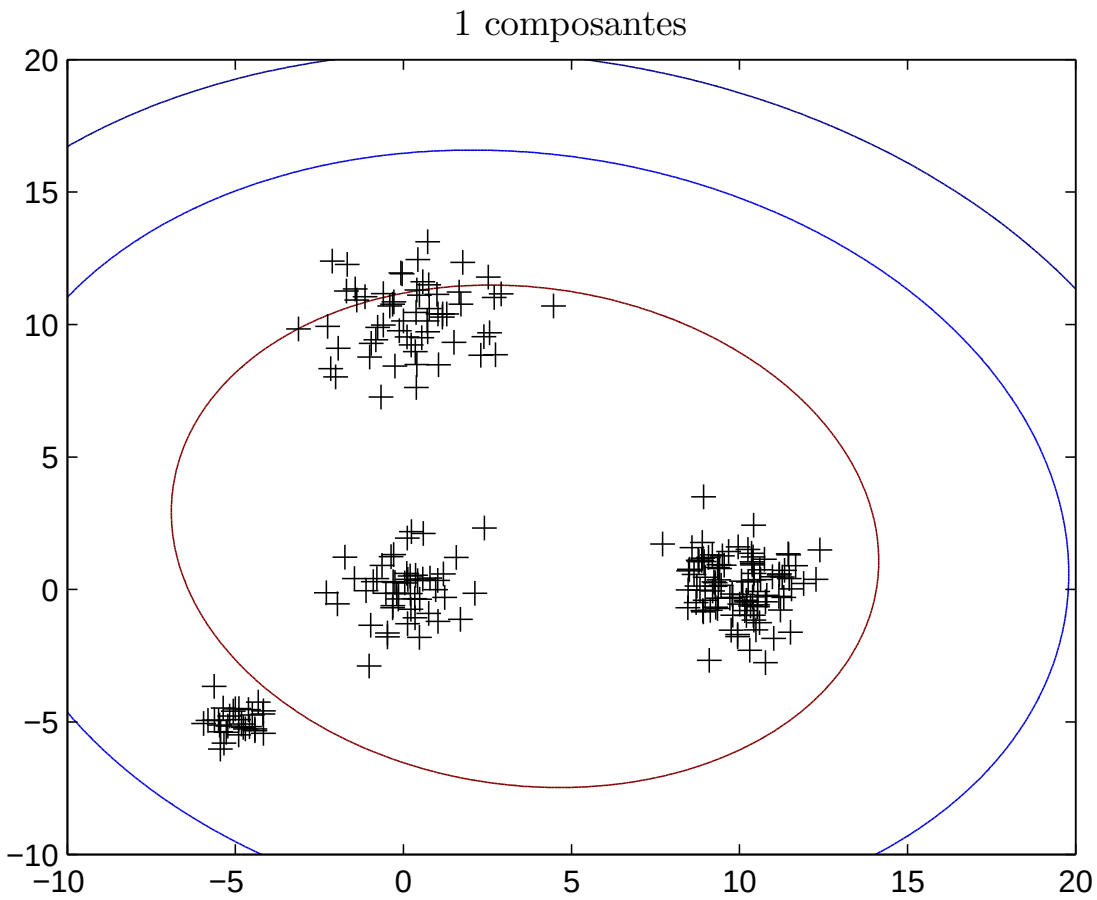
- Plus c est élevé, moins de points peuvent être affectés à chacune des Gaussiennes
- Recherche de modèles parcimonieux, c.a.d. limitant le nbre de composantes (donc le nbre de paramètres à estimer)

Critère BIC

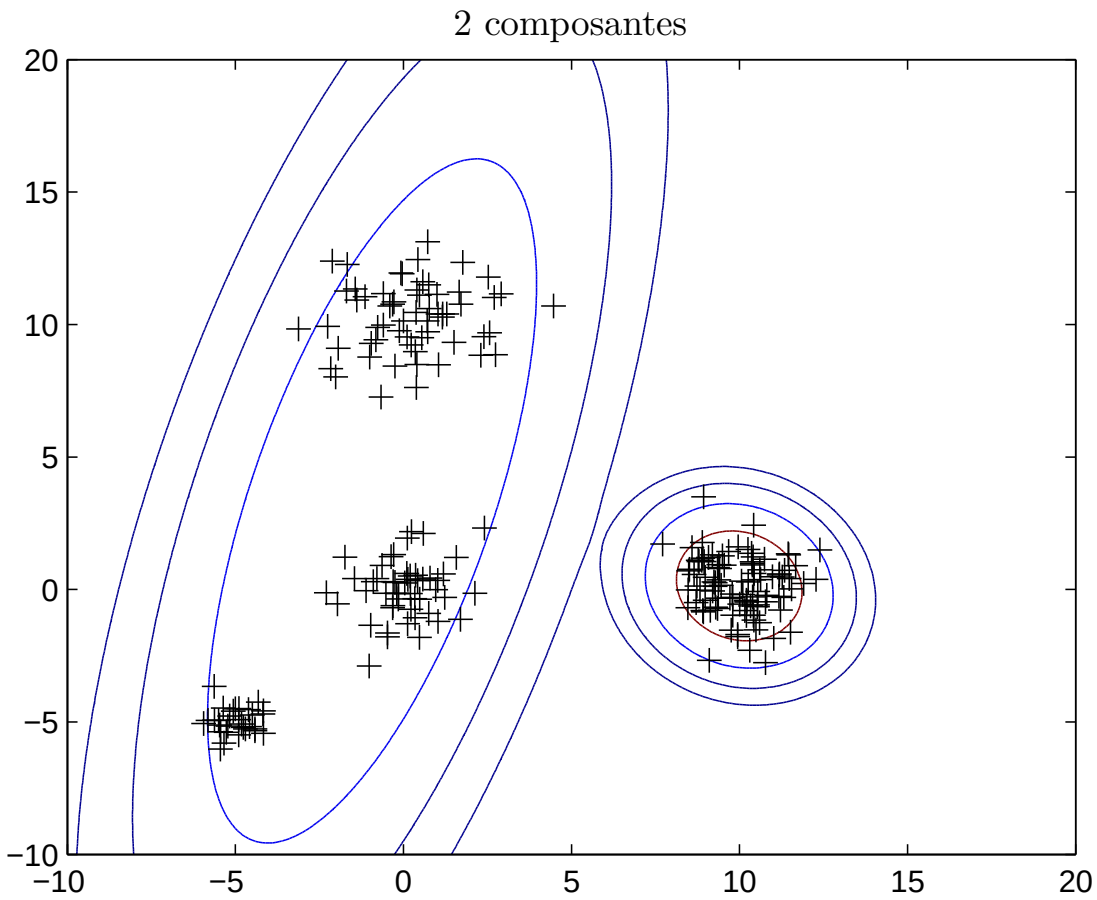
- Plus c grand, plus l est maximisé
- Contrebalancer la vraisemblance par un terme pénalisant la complexité

$$BIC = 2 * l(\theta) - |\theta|. \log(N)$$

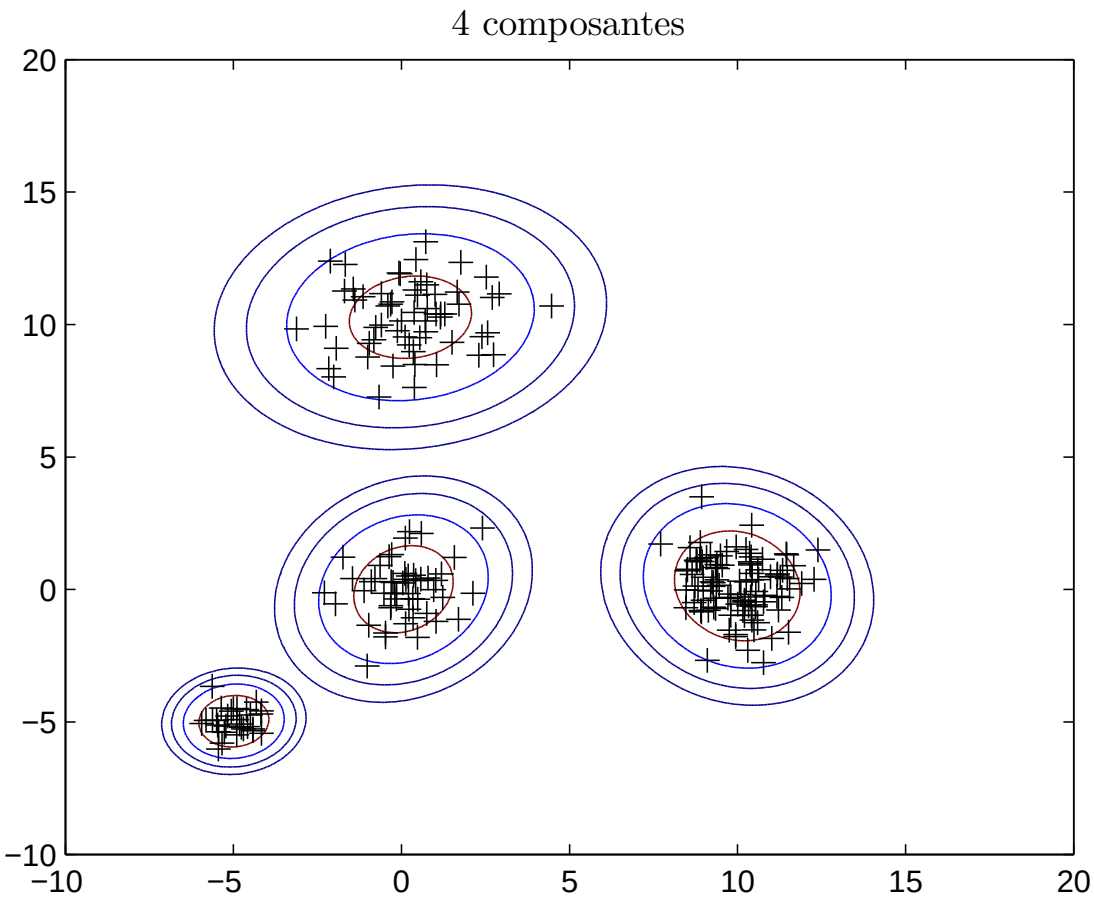
Exemple : mélange à 4 composantes



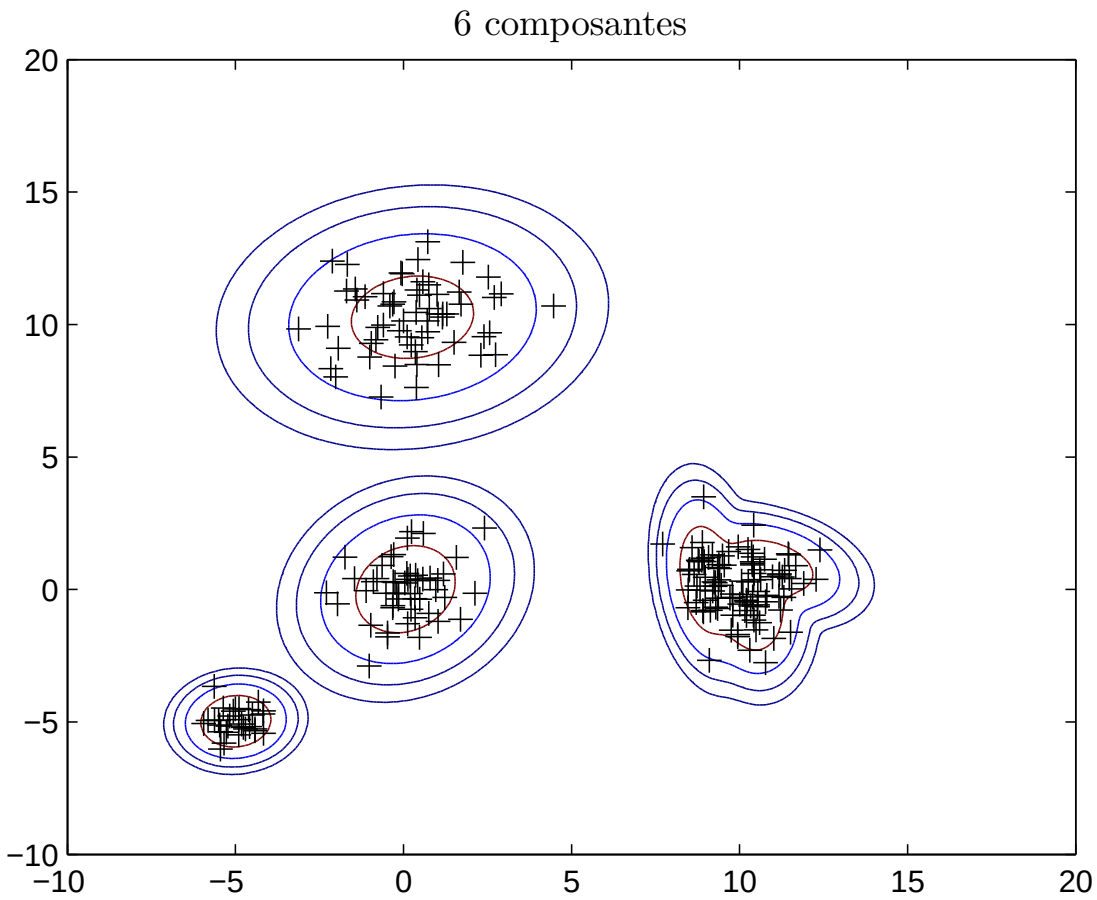
Exemple : mélange à 4 composantes



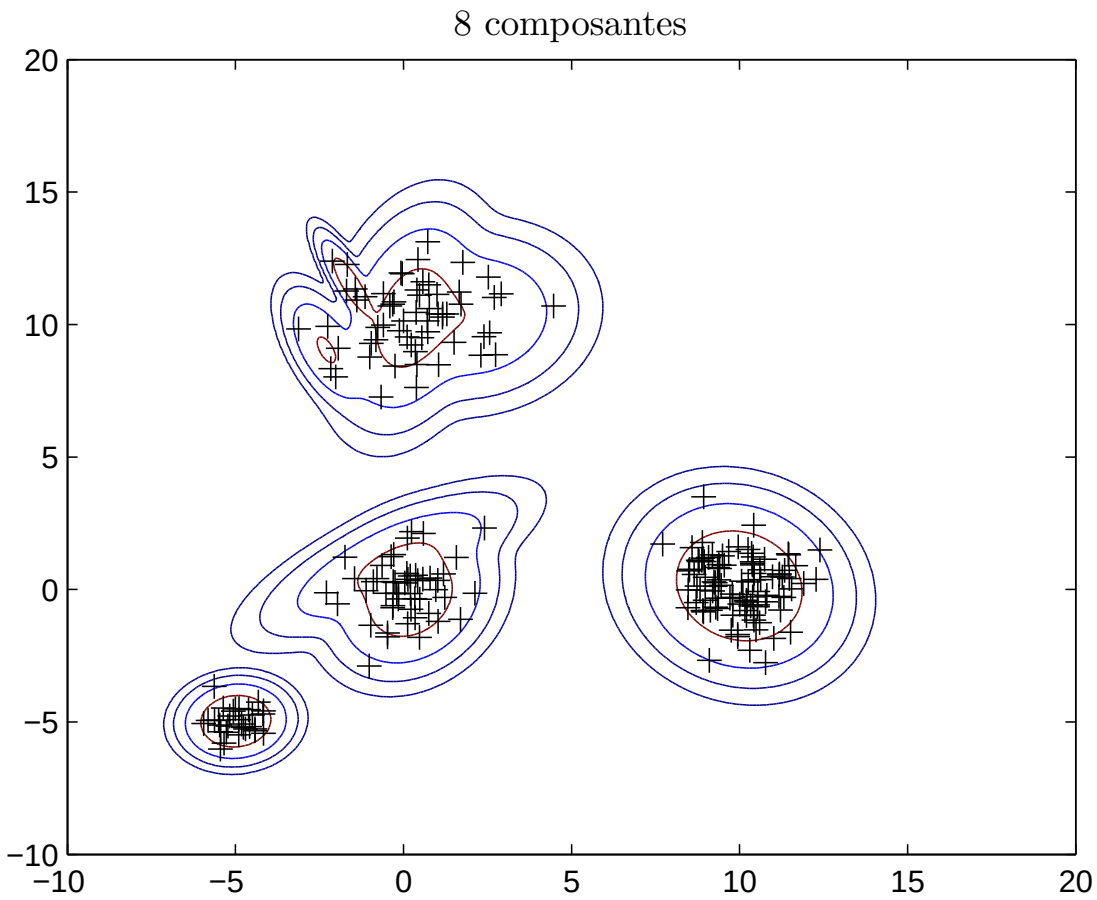
Exemple : mélange à 4 composantes



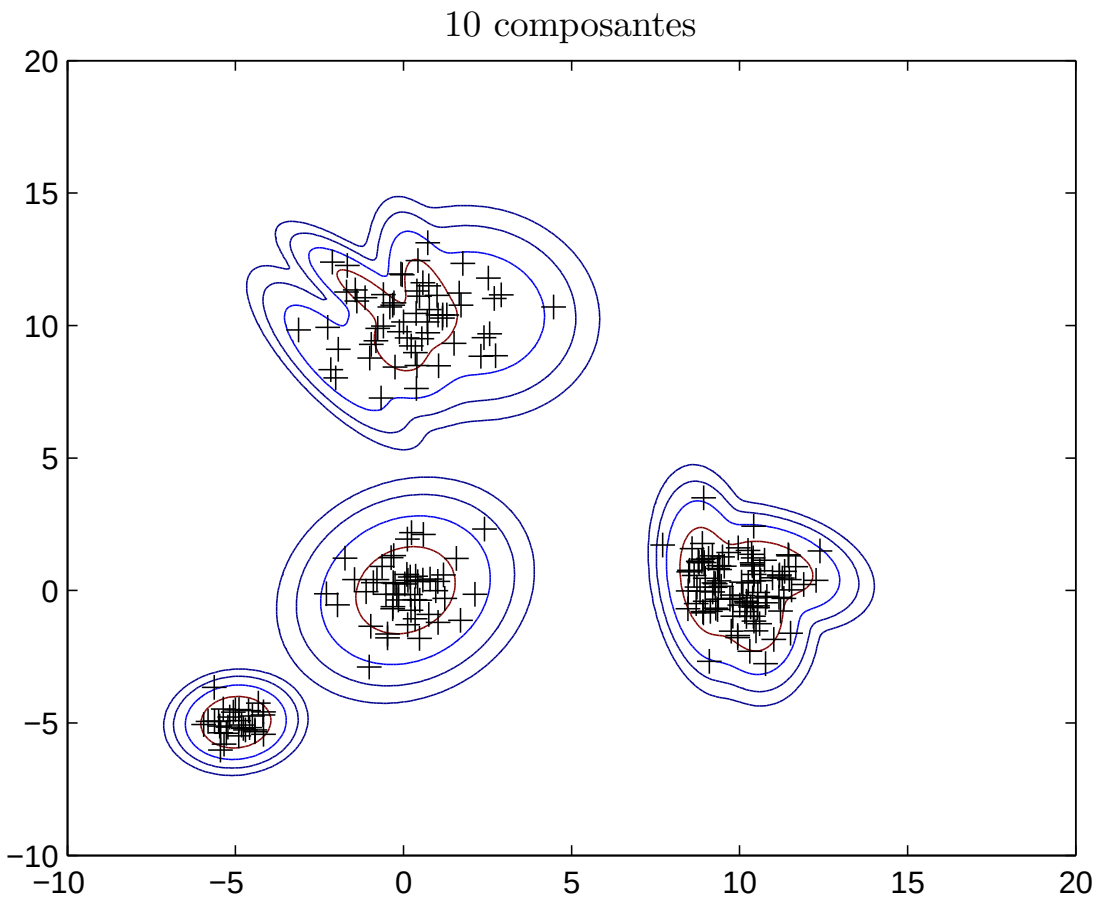
Exemple : mélange à 4 composantes



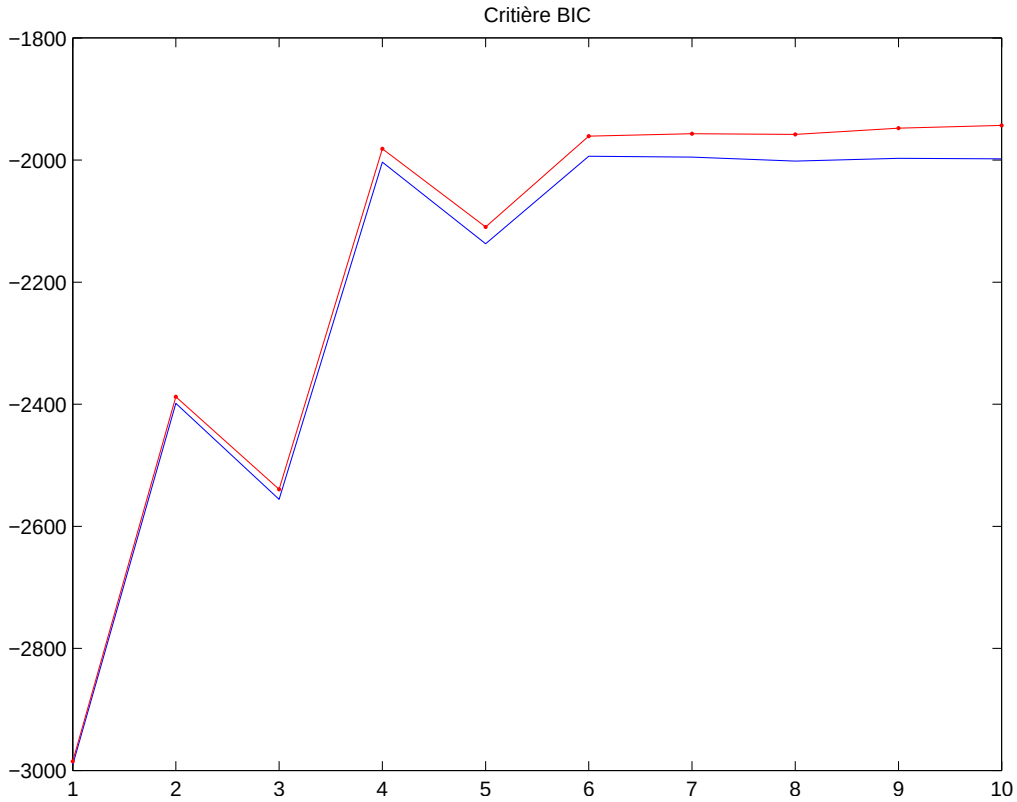
Exemple : mélange à 4 composantes



Exemple : mélange à 4 composantes



Exemple (suite)



- Nécessité de tester tous les modèles
- Dépend de la convergence de l'algo
- *Fine tuning* à la main !

Conclusions

Mélanges de Gaussiennes

- Le mélange de gaussiennes généralise les hypothèses de l'ACP et l'ADL
- Estimation explicite de la densité
- Également classification : si les composantes représentent des classes, un point i est associé à la classe k pour laquelle $p(k|x_i) \approx \pi_k \phi_k(x_i)$ est maximum par rapport à toutes les autres classes

Conclusions

Algorithme EM

- Algorithme itératif de maximisation de vraisemblance
- Utilisation dans de nombreux cas autres que mélanges de Gaussiennes
- Basé sur l'existence de variables cachées, dites *latentes*
- Probabilistic Latent Semantic Analysis (pLSA) $\rightarrow$ EM où les variables cachées sont les concepts latents