Classification non supervisée : k moyennes et ACP

Chloé Clavel,
chloe.clavel@telecom-paristech.fr,
Sources : Isabelle Bloch, Jean-Marie Nicolas, Bernard Burtschy, Anne Sabourin

Telecom ParisTech, France

Plan du cours

Plan du cours

La classification non supervisée

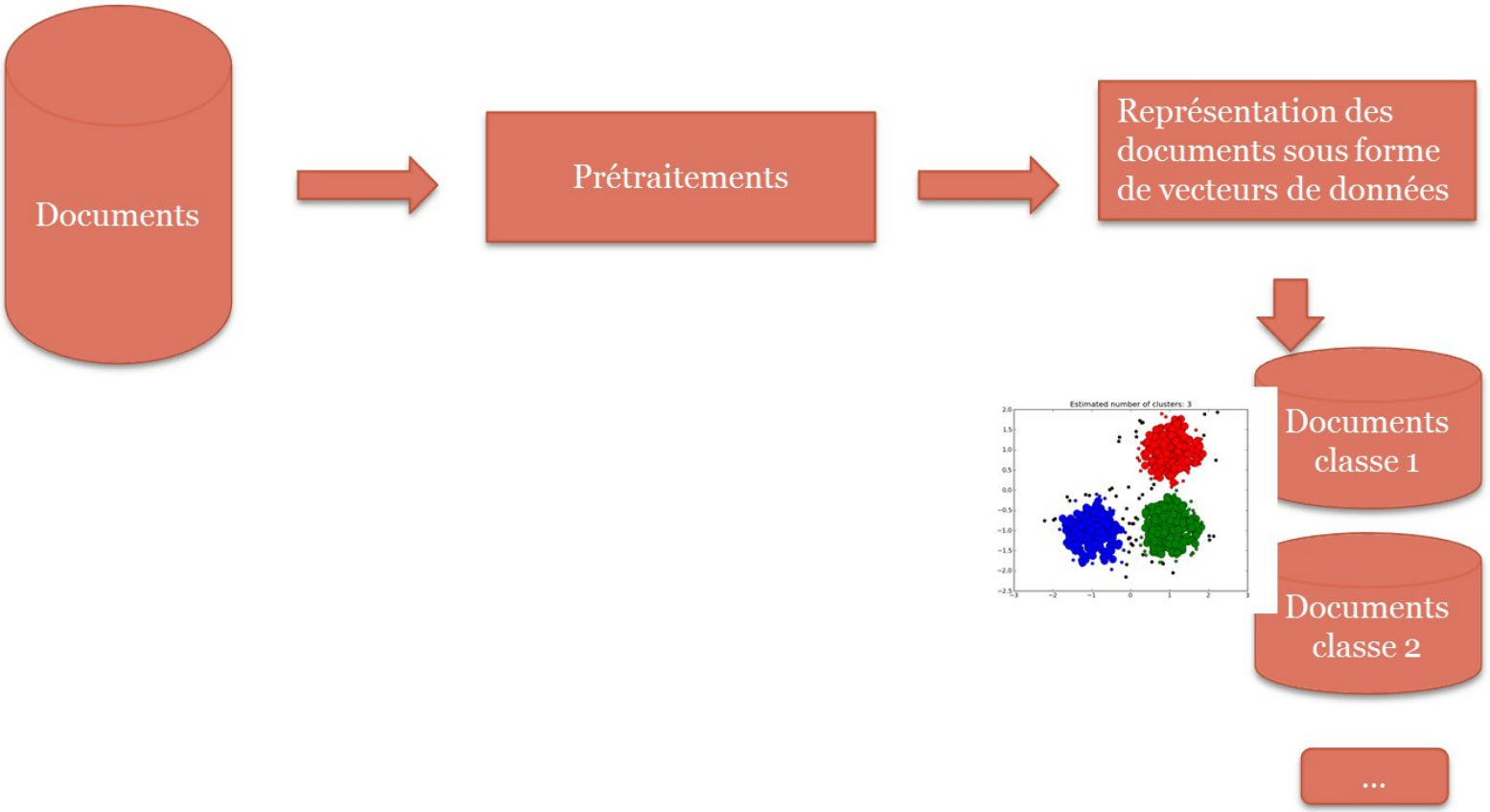
Méthodes de classification non supervisées

- Méthodes de regroupement ou de clustering
- problème de partitionnement des objets à classifier selon certains critères

Méthodes hiérarchiques *vs.* méthodes non hiérarchiques

- Méthodes hiérarchiques : graphes/arbres (voir polycopié - ne seront pas vues dans le cadre de ce cours)
- Méthodes non hiérarchiques : **k-moyennes**, ISODATA, boules optimisées, nuées dynamiques

Méthodes non hiérarchiques de classification non supervisée



Méthodes non hiérarchiques de classification non supervisée

- on cherche à regrouper les observations en différentes classes selon un **critère de regroupement**
- on recherche une description des classes par leur densité de probabilité :
  - on connait les formes des densités de probabilités mais pas les paramètres des lois
  - on cherche les paramètres des lois qui maximisent le critère de regroupement des observations selon ces classes

Méthodes non hiérarchiques de classification non supervisée

- problème d'estimation de paramètres comme vu dans le cours de classification bayésienne
  - en entrée : le nombre de classes, les probabilités a priori de chaque classe, et les formes de densité des probabilités conditionnelles d'appartenance des observations à chaque classe (ex : gaussienne, multinomiale, loi gamma)
  - ce que l'on cherche à estimer : les paramètres des lois de probabilités des classes
  - méthode classiquement utilisée : estimateur du maximum de vraisemblance

Estimation des paramètres des lois par le maximum de vraisemblance

Cas où : les observations sont distribuées dans chaque classe selon la loi normale dont on connait les variances.

- loi normale

$$p(x|\omega_i;\mu_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)}$$

- avec μ_i la moyenne (de même dimension que les échantillons)
- Σ_i la matrice de variance-covariance (supposée connue)
- d la dimension de l'espace des échantillons

Estimation des paramètres des lois par le maximum de vraisemblance

- On cherche les paramètres (ici les moyennes) qui maximisent (algorithme EM (Expectation-Maximization))

$$p(x|\omega_i; \mu_i)$$

- Après quelques calculs (voir poly, annulation de la dérivée de l'expression + application de la règle de Bayes), on obtient :

$$\hat{\mu}_i = \frac{\sum_{k=1}^{n} P(\omega_i|x_k; \hat{\mu})x_k}{\sum_{k=1}^{n} P(\omega_i|x_k; \hat{\mu})}$$

- Cette expression est implicite (les $\hat{\mu}_i$ sont des deux côtés de l'équation) : c'est un système d'équations couplées non linéaires dont la résolution est compliquée.

Estimation des paramètres des lois par le maximum de vraisemblance

Pour résoudre l'équation $\hat{\mu}_i = \frac{\sum_{k=1}^{n} P(\omega_i|x_k;\hat{\mu})x_k}{\sum_{k=1}^{n} P(\omega_i|x_k;\hat{\mu})}$, on utilise un schéma itératif :

- on part d'un regroupement initial et des moyennes associées à ce regroupement $\hat{\mu}(0)$
- on estime les moyennes du regroupement de l'itération suivante en fonction des estimations de l'itération précédente :

$$\hat{\mu}_i(j+1) = \frac{\sum_{k=1}^{n} P(\omega_i|x_k;\hat{\mu}(j))x_k}{\sum_{k=1}^{n} P(\omega_i|x_k;\hat{\mu}(j))}$$

- jusqu'à ce que $\hat{\mu}_i(j+1) = \hat{\mu}(j)$

Ne garantit pas l'obtention d'un minimum global (dépend de l'initialisation)

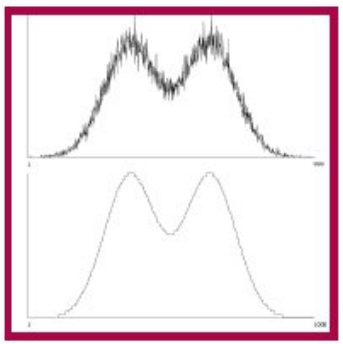
Estimation des paramètres des lois

Schéma itératif et principe général :

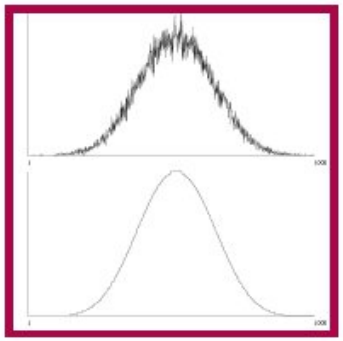
- initialisation :
  - considérer un ensemble K de clusters et initialiser les paramètres de la loi associée à chaque cluster
  - attribuer chaque observation à un cluster en fonction de sa probabilité d'appartenir à une classe (classe la plus probable) -> partitionnement initial
- itération :
  - Recalculer les paramètres du modèle sur la base des clusters du partitionnement courant
  - Redistribuer les observations dans les clusters à partir de ce nouveau modèle

Illustration dans le cas gaussien

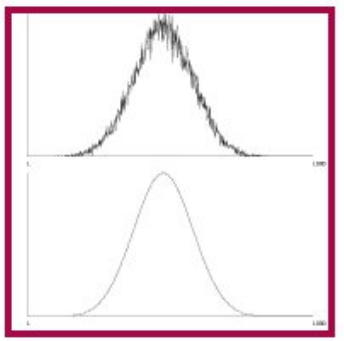
En entrée : 2 classes avec comme probabilités a priori $P(w1) = P(w2) = 0.5$, une forme de densité de probabilité gaussienne avec comme variance $\sigma = 8$



µ1 =50, µ2 =75



µ1 =50, µ2 =60



µ1 =50, µ2 =55

On cherche à estimer μ_1 et μ_2 en utilisant l'algorithme EM

Illustration dans le cas gaussien

On cherche à obtenir $\mu_1 = 50$ et $\mu_2 = 75, 60, 55$

μ_2	μ_1	μ_2	Nombre itérations
75	50,01	75,1	3
60	49,96	59,94	7
55	49,95	55,10	16

Le nombre d'itérations requises est plus élevé quand μ_1 et μ_2 sont proches

Plan du cours

Les approximations des k-moyennes

Rappel :

- La distance de Mahalanobis est la distance de l'observation au centre de la classe pondérée par la dispersion de la classe
  - centre de la classe i = moyenne des observations $\hat{\mu}_i$
  - dispersion de la classe i = variance des observations $\hat{\Sigma}_i$

$$d = (x_k - \hat{\mu}_i)^t \hat{\Sigma}_i^{-1} (x_k - \hat{\mu}_i)$$

- La probabilité a posteriori $P(\omega_i | x_k, \hat{\theta})$ est d'autant plus grande que la distance de Mahalanobis d est petite
- trouver $\hat{\mu}_i$ et $\hat{\Sigma}_i$ qui maximisent $P(\omega_i | x_k, \hat{\theta})$ revient à trouver les centres des classes qui minimisent d

Les approximations des k-moyennes

Première approximation :

- on remplace la distance de Mahalanobis par la distance euclidienne $\|x_k - \hat{\mu}_i\|^2$
- on cherche alors la moyenne $\hat{\mu}_i$ qui minimise la distance

Cette approximation ramène le problème du regroupement des observations en des classes en l'optimisation d'un critère lié à la distance au centre des classes (moyenne des observations sur les classes) : **l'inertie**

$$\sum_{i \in \{1,\ldots,C\}} \sum_{x_k \in \omega_i} \|x_k - \mu_i\|_2^2$$

Les approximations des k-moyennes

Deuxième approximation sur les lois a posteriori

- "Binarisation" du processus : x_k est considéré comme appartenant de manière certaine à la classe ω_m, et n'appartenant (de façon certaine aussi) à aucune des autres classes.
- la probabilité est égale à 1 si l'observation appartient à la classe ω_m et à 0 sinon

$$\hat{P}(\omega_i|x_k;\hat{\theta}) = \begin{cases} 1 \text{ si } i = m, \\ 0 \text{ sinon,} \end{cases} \quad (1)$$

- $\hat{\mu}_i(j+1) = \frac{\sum_{k=1}^{n} P(\omega_i|x_k;\hat{\mu}(j))x_k}{\sum_{k=1}^{n} P(\omega_i|x_k;\hat{\mu}(j))} \Rightarrow \hat{\mu}_i(j+1) = \frac{1}{|\omega_i(j)|} \sum_{x_k \in \omega_i(j)} x_k$

K-means - l'algorithme

L'algorithme de K-Means partitionne les points en K groupes disjoints $\{\omega_1, \ldots, \omega_K\}$ en minimisant la variance intra-classe. Le critère minimisé est appelé *inertie* :

$$\sum_{i \in \{1,\ldots,K\}} \sum_{x_k \in \omega_i} \|x_k - \mu_i\|_2^2$$

où les μ_i sont les centroides des classes :

$$\mu_i = \frac{1}{|\omega_i|} \sum_{k \in \omega_i} x_k, \forall i \in \{1, \cdots, K\} \ ,$$

et où $x_k \in \omega_{m_o}$ si :

$$m_o = argmin_{i \in \{1,\ldots,K\}} \|x_k - \mu_i\|_2 \ .$$

K-means - l'algorithme

L'apprentissage se fait en alternant deux étapes :

- une étape d'assignement où, sachant les (μ_i), on va calculer les labels de chaque point
- une étape de mise à jour des centroides sachant les labels.

On arrête l'algorithme quand l'inertie ne décroit plus beaucoup.
L'inertie est un critère non-convexe
$\Rightarrow$ la solution trouvée dépend de l'initialisation
$\Rightarrow$ lancer l'algorithme plusieurs fois avec des initialisations différentes pour ne garder que la solution avec l'inertie la plus faible.

K-means – l'algorithme

- **Etape 1 : Initialisation.** Choix de centres initiaux $\mu_i(1)$ arbitraires. Options :
  - équidistribués
  - tirés au hasard
  - les K échantillons choisis au hasard parmi les n
- **Etape 2 : Affectation.** À l'itération j, x est affecté à ω_i si :

$$\|x - \mu_i(j)\| = \min_{l=1}^{k} \|x - \mu_l(j)\|. \quad (2)$$

Tous les échantillons sont classés selon cette règle (du centre le plus proche).

K-means – l'algorithme

- **Etape 3 : Mise à jour des centres.** Calcul des nouveaux centres $\mu_i^{(j+1)}$ pour minimiser l'erreur quadratique :

$$J(\mu_i^{(j+1)}) = \sum_{x \in \omega_i} \|x - \mu_i^{(j+1)}\|^2$$

En annulant la dérivée de cette expression par rapport à μ_i, on obtient :

$$\frac{\partial J}{\mu_i} = -2 \sum_{x \in \omega_i} (x - \mu_i) = 0,$$

d'où la valeur optimale de μ_i pour l'itération $(j+1)$:

$$\mu_i^{(j+1)} = \frac{1}{n_i} \sum_{x \in \omega_i} x. \qquad (3)$$

K-means – l'algorithme

- **Etape 4 : Test de convergence.** Si $\forall i, \mu_i^{(j+1)} = \mu_i^{(j)}$, fin. Sinon, retour à l'étape 2.

Exemples

`./kmeansNLP.pdf`

Exercice d'application

`../Exercices/Exercises695Clus-solution.pdf`

Plan du cours

Analyse en composantes principales - ACP

Origine Karl Pearson (1901)
Autre dénomination : décomposition en valeurs singulières et
Décomposition de Karhunen-Loève (1958)

Analyse en composantes principales - ACP

Principe :

- approche non supervisée car les données sont traitées indépendamment de leurs classes
- représenter les données par des vecteurs de dimension réduite

Analyse en composantes principales - ACP

Les données :

- une matrice correspondant aux observations (N individus pour lesquels on dispose de p caractéristiques)

	$\mathbf{V_1}$		$\mathbf{V_j}$		$\mathbf{V_P}$
Obs. 1					
Obs. I					
.					
Obs n					
Moyenne	m_1		m_j		m_p

Analyse en composantes principales - ACP

- Exemple de données :
  - un individu = un document textuel et ses caractéristiques les mots qui le composent
  - un individu = une image et ses caractéristiques le niveau de gris de chacun de ses pixels

Analyse en composantes principales - ACP

Objectifs :

- Analyser les relations entre un grand nombre de variables
- Rechercher des variables de synthèse en nombres réduits (facteurs)
- En perdant le moins d'information

Analyse en composantes principales - ACP

Soient :

- N individus, p caractéristiques
- $V_k = (x_k^i)_{i=1..n}$ avec x_k^i la $k$ième caractéristique de l'individu i.
- X la matrice dans laquelle chaque ligne est constituée par un individu et chaque colonne représente une variable.

Analyse en composantes principales - ACP

	$\mathbf{V_1}$		$\mathbf{V_j}$		$\mathbf{V_P}$
Obs. 1					
Obs. I					
.					
Obs n					
Moyenne	m_1		m_j		m_p

Raisonnement sur les lignes :

- les observations = N lignes de X = N points dans l'espace $\mathbb{R}^p$.
- on peut identifier les groupes d'observations qui ont des mesures voisines (proximité des obs.)

Analyse en composantes principales - ACP

	$\mathbf{V_1}$		$\mathbf{V_j}$		$\mathbf{V_P}$
Obs. 1					
Obs. I					
.					
Obs n					
Moyenne	m_1		m_j		m_p

Raisonnement sur les colonnes :

- les caractéristiques = p col. de X = p points de l'espace $\mathbb{R}^N$
- chaque caractéristique peut être décrite par sa mesure sur les N observations et on peut observer la proximité entre les caractéristiques dans $\mathbb{R}^N$

Analyse en composantes principales - ACP

Objectif : Résumer les p variables en k synthèses (facteurs) $k \leq\leq p$ indépendantes

- Première contrainte : les facteurs sont des combinaisons linéaires des variables initiales $F_k = \sum_{j=1}^{p} \lambda_{kj} V_j$, soit par chaque individu i, $F_{k,i} = \sum_{j=1}^{p} \lambda_{kj} x_j^i$
- Deuxième contrainte : les facteurs ($F_k \in \mathbb{R}^n$) sont indépendants deux à deux $F_i \perp F_j$ quand $i \neq j$

Analyse en composantes principales - ACP

Le but de l'ACP est de visualiser dans un espace de plus petite dimension les proximités entre les observations et ainsi les corrélations entre les variables.

Les mesures utilisées dans l'ACP :

- Calcul de la distance entre deux observations (par exemple distance euclidienne) :

$$d^2(i,j) = \sum_{k=1}^{p} (x_k^i - x_k^j)^2$$

Analyse en composantes principales - ACP

Le but de l'ACP est de visualiser dans un espace de plus petite dimension les proximités entre les observations et ainsi les corrélations entre les variables.

Les mesures utilisées dans l'ACP :

- Mesure de la relation entre deux variables :
  - le produit scalaire :

$$\langle V_i, V_j \rangle = \sum_{k=1}^{N} x_i^k * x_j^k$$

Analyse en composantes principales - ACP

Les trois ACP :

- analyser les données brutes
- analyse centrée : on remplace chaque observation par ses écarts à la moyenne

$$x_k'^i = x_k^i - m_k$$

- analyse centrée réduite ou normée (**la plus courante**) :

$$x_k'^i = \frac{x_k^i - m_k}{\sigma_k}$$

Analyse en composantes principales - ACP

Les mesures utilisées dans l'analyse centrée et l'analyse centrée réduite :

- Calcul de la moyenne et de l'écart-type sur toutes les N observations pour la variable j :

$$m_j = \frac{1}{n} \sum_{i=1}^{N} x_j^i$$

et

$$\sigma_j^2 = \frac{1}{n} \sum_{i=1}^{N} (x_j^i - m_j)^2$$

Analyse en composantes principales - ACP

Les trois ACP :

- analyse centrée : on remplace chaque observation par ses écarts à la moyenne
  - le produit scalaire entre deux variables devient :

$$\langle V_i, V_j \rangle = \sum_{k=1}^{N} (x_i^k - m_i)(x_j^k - m_j) = N * Covariance(V_i, V_j)$$

Analyse en composantes principales - ACP

Les trois ACP :

- analyse centrée réduite ou normée (**la plus courante**) :

  $x'^i_k = \frac{x'^i_k - m_k}{\sigma_k}$

  - le produit scalaire devient :

$$\langle V_i, V_j \rangle = N * Correlation(V_i, V_j)$$

Rappel : $Correlation(V_i, V_j)$: $\frac{Covariance(V_i, V_j)}{\sigma_i \sigma_j}$

Analyse en composantes principales - ACP

Les étapes de l'ACP centrée réduite :

- ETAPE 1 : centrage et réduction des données ; soient m_k et σ_k, les moyenne et écart-type de la $k$ième variable, on notera $x_k'^i = \frac{x_k'^i - m_k}{\sigma_k}$ la donnée centrée réduite ;

Analyse en composantes principales - ACP

Objectif : visualiser dans un espace de plus petite dimension les proximités entre les observations et ***ainsi les corrélations entre les variables.***

Les étapes de l'ACP centrée réduite :

- ETAPE 2 calcul de la matrice de covariance des données centrées réduites (mesure les relations entre les variables) ;

$$\langle V_i, V_j \rangle = \sum_{k=1}^{N} (x_i^k - m_i)(x_j^k - m_j) = N * Covariance(V_i, V_j)$$

Analyse en composantes principales - ACP

- ETAPE 3 : Recherche du sous-espace de projection
  - Projection des observations (= nuage de $\mathbb{R}^p$) dans un espace de dimension plus petite d $\leq\leq p$
  - Le sous-espace optimal permet de visualiser les proximités entre les observations et les corrélations entre les variables.

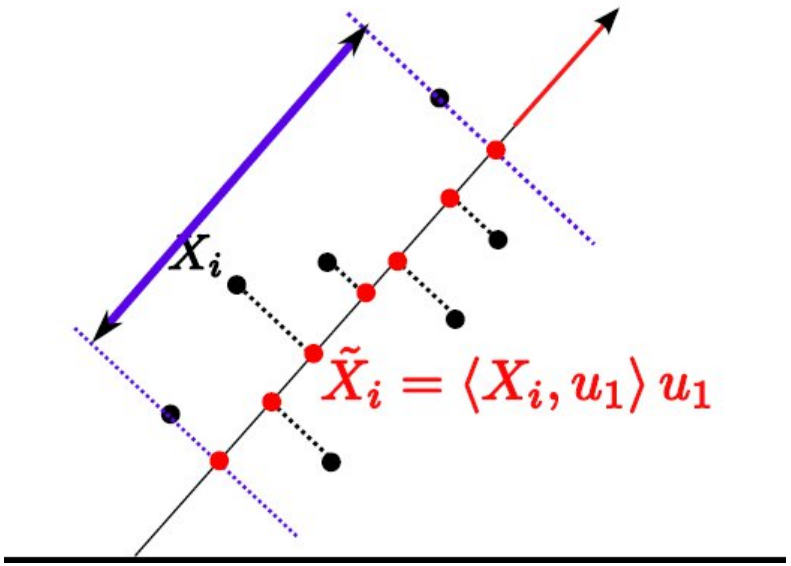
Analyse en composantes principales - ACP

On cherche donc l'espace de projection de dimension d qui déforme le moins les proximités entre les observations.
RAPPEL MATHEMATIQUE :

- un critère de déformation minimum de nuage par projection est celui où les distances entre les points projetés sont les plus voisines de celles entre les points initiaux.
- Ceci revient à rechercher les vecteurs propres de la matrice de covariance $X^t X$

Analyse en composantes principales - ACP

- ETAPE 3 : Recherche du sous-espace de projection
  - Le meilleur sous-espace de dimension 1 est caractérisé par un vecteur unitaire u_1 ($|| u_1 || = 1$) : le vecteur propre associé à la plus grande valeur propre de la matrice de covariance.
  - le vecteur qui «passe au mieux » à travers le nuage de points i.e. celui qui minimise la somme des carrés de la distance de chaque point à la droite.



Analyse en composantes principales - ACP

- ETAPE 3 (suite) : Recherche du sous-espace de projection
  - Passage aux dimensions supérieures dans l'espace de projection : calcul des valeurs propres λ_j et vecteurs propres u_j de la matrice de covariance ;

Analyse en composantes principales - ACP

- PROJECTION
  - vérification de l'ordre des vecteurs propres selon les valeurs propres croissantes
  - calcul des composantes principales $x_q'^i$ exprimées dans la base des vecteurs propres :

$$x_q'^i = x'^i u_q$$

en notant x'^i le vecteur ligne constitué par les observations de l'individu i.

Analyse en composantes principales - ACP

- PROJECTION
  - On obtient donc de nouvelles variables constituées par des combinaisons linéaires des anciennes.
  - Les composantes principales contiennent une quantité d'information proportionnelle à la valeur propre correspondante.
  - les valeurs propres mesurent l'influence de chaque facteur ou composante principale
  - On définit ainsi le pourcentage d'inertie par $\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$.

Analyse en composantes principales - ACP

Exemple d'utilisation

- dans le cas de l'analyse de documents textuels : décomposition de matrices selon leurs directions propres (ou singulières) pour conserver un maximum d'information sur un nombre minimum de dimensions. La décomposition en valeurs singulières de la matrice terme/document permet d'obtenir des thèmes dominants dans le corpus, chacun étant associé à un sous-espace singulier.
- dans le cas de l'analyse d'image (voir TP) : réduire le nombre de canaux nécessaires pour conserver l'essentiel de l'information contenue dans l'image

Plan du cours

Quelques références

Likforman-Sulem, Laurence, and Elisa Barney Smith. *Reconnaissance des formes-Théorie et pratique sous Matlab-Cours et exercices corrigés.* (2013).

Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.