Apprentissage Non Supervise

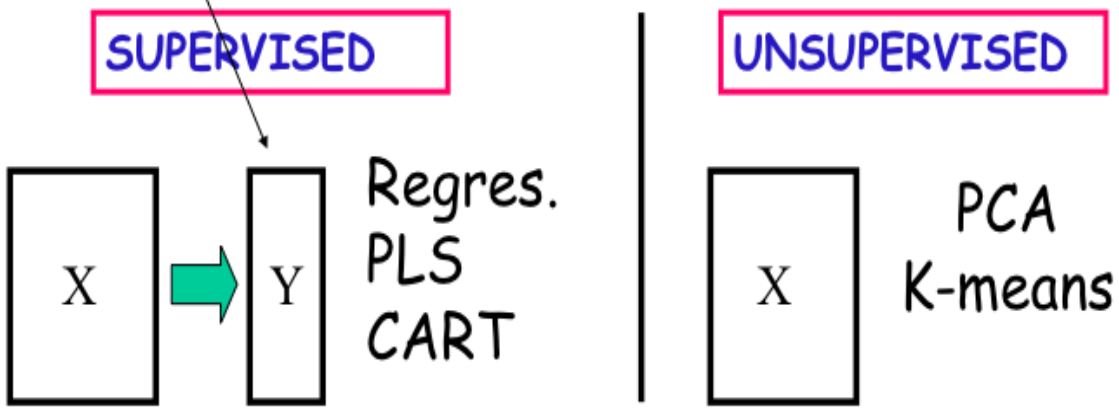
B. Ghattas & G. Oppenheim

Université d'Aix-Marseille - Université Paris-Sud

December 8, 2016

To be or not to be Supervised?

Psychological studies: when there is an information coming from a SUPERVISOR who says: It's good or It isn't.
Here the context is a little different:
there is a target



The last application I studied is a Supervised problem. Quality control of a Rubber/Metal suspension bushes for an engine mount (articulation élastique pour support moteur)

7

Les méthodes de cette séance

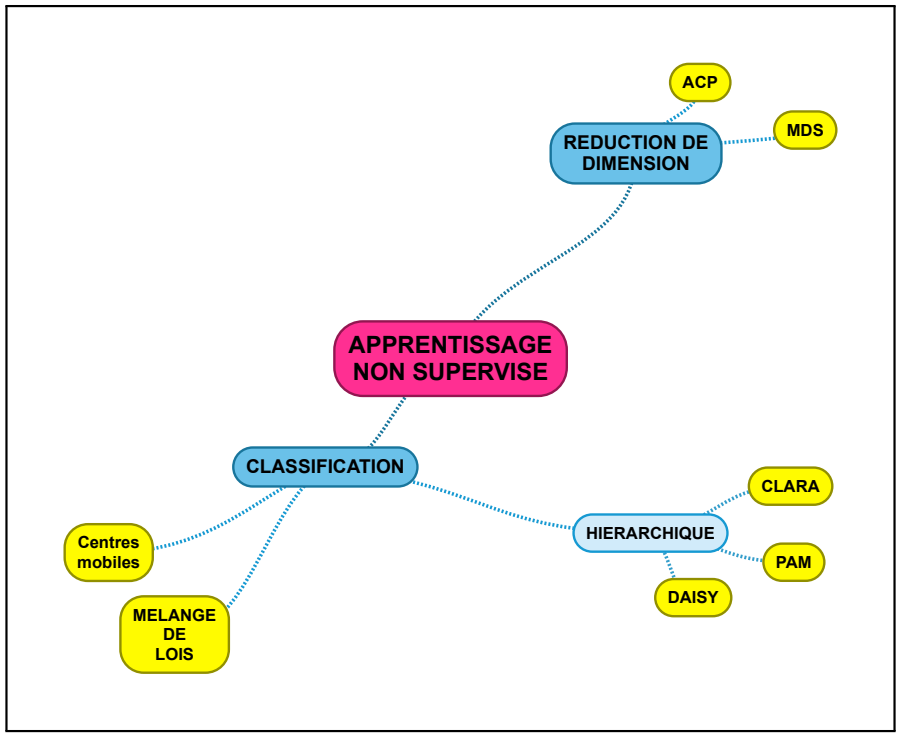


Figure: Les méthodes Non-Supervisées de cette séance

Exemple détaillé

Le cas étudié est un problème de cytométrie consistant à analyser des particules contenues dans un échantillon prélevés dans un environ marin.

An application: Flow cytometer dataset

- A collection of Phytoplankton cells was performed in a sailing yacht harbour in Marseille.
- Samples were collected every 30 min during one day, from 26-09-2011 at 17h00 to 27-09-2011 at 17h00.
- Five signals were recorded for each cell: FWS (ForWard Scatter), SWS (Side-Ward Scatter), FLY (Fluorescence Yellow), FLO (Fluorescence Orange) and FLR (Fluorescence Red).
- From each signal the following features were computed: average, maximum, length, inertia, fill factor, center of gravity and asymmetry.

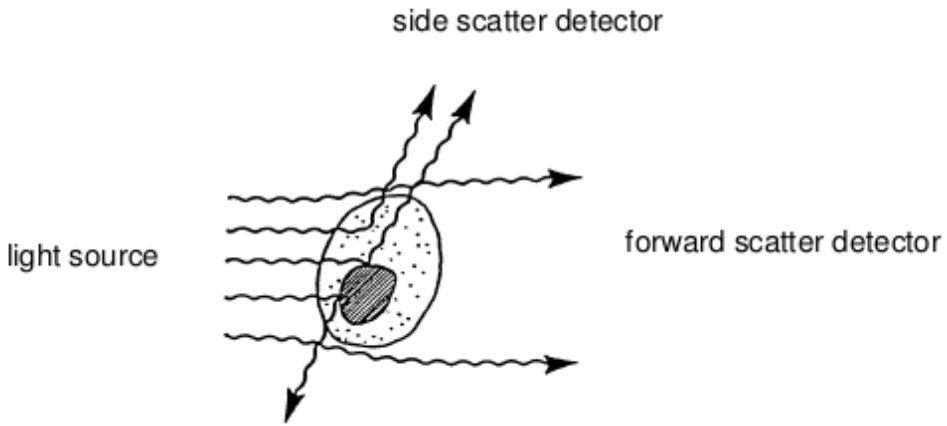
External factors

Some meteorological features were collected during those 2 days (26-27 September 2011) in order to characterize the evolution of Phytoplankton particles:

- - Temperature of the seawater (in Celsius řC).
- - Acidity/Basicity of the seawater (in pH units).
- - Oxidation-Reduction Potential (ORP): a measure of the cleanliness of the seawater and its ability to break down contaminants. (in mV).
- - Salinity of the water (in ppt Șparts per thousand).
- - Turbidity of the seawater: measures the cloudiness or haziness of the water. (in Nephelometric Turbidity Units (NTU)).
- - Chlorophyll (in ?g per liter ?g/l): a molecule that absorbs sunlight and uses its energy to synthesize carbohydrates from CO2 and water, its concentration in the seawater gives an insight of the distribution of Phytoplankton.

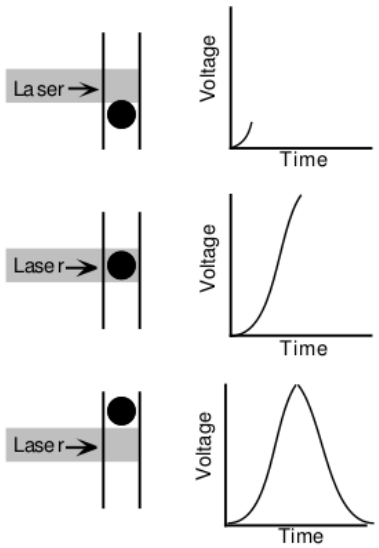
Data acquisition

Detection of the signal in the Flow Cytometer: Light scattering occurs when a particle deflects incident laser light. A detector in front of the laser beam measures forward scatter FS and several detectors to the side measure side scatter SS. Fluorescence detectors measure the fluorescence emitted from stained cells.



Signal creation

The highest point of the signal occurs when the particle is in the center of the beam and the maximum amount of scatter or fluorescence is achieved.



The Forward Scatter light FWS is proportional to the size of the cell and the Side-ward Scatter light is proportional to the internal complexity of the cell. The three Fluorescences (Yellow, Orange and Red) are used for the separation of the particles by a particular protein.

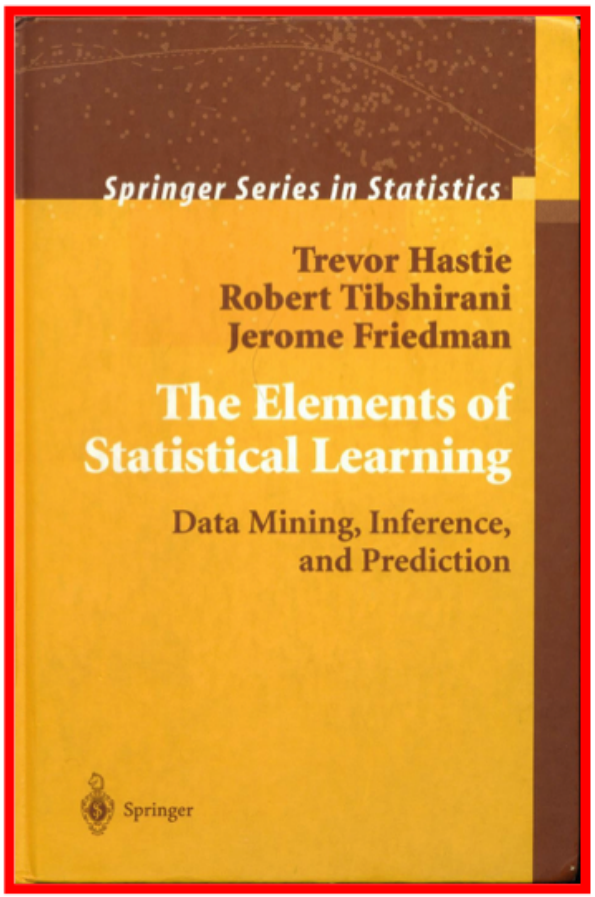
More about the dataset..

We have 39 samples, each consisting on almost 100.000 observations, with 50 variables corresponding to the features extracted from the signals.

Objective

- Cluster each sample. Choose an appropriate method, and an appropriate number of clusters.
- How can we model the differences between successive partitions ?
- How can we relate partitions' evolution to external factors like Temperature, turbidity or salinity ? Which ones are relevant to explain this evolution ?

Référence



2001

533p

19/nov/16 3

le cours est appuyé sur

- Le MOOC de Hastie sur le non-supervisé
- Le livre de Hastie-Tibshirani-Friedmann

Plan

Densité

Déterminer la densité.

- *Histogramme*
- *Densité*: noyaux

Plan

Clustering methods

Clustering aims to partition the data. Partitioning may be:

- *Hierarchical*: find successive groups splitting or joining previously established groups ("bottom-up", "top-down")
- *Non Hierarchical*: k-means. Density based methods: Mixture models, DBSCAN.

Some supervised methods are based also on partitioning the space of explanatory variables (CART, SVM, LDA).

Clustering methods

- Works often with dissimilarity matrix.
- Often needs the number k of clusters to be fixed in advance.
- There exist some heuristics to estimate k (pseudo T, pseudo F).
- We need also to estimate the quality of a partition and to compare partitions.

Often we need to deal with ...

- qualitative datasets, numeric, or mixed
- functional data, image data

The most widely used method: The k-means algorithm.

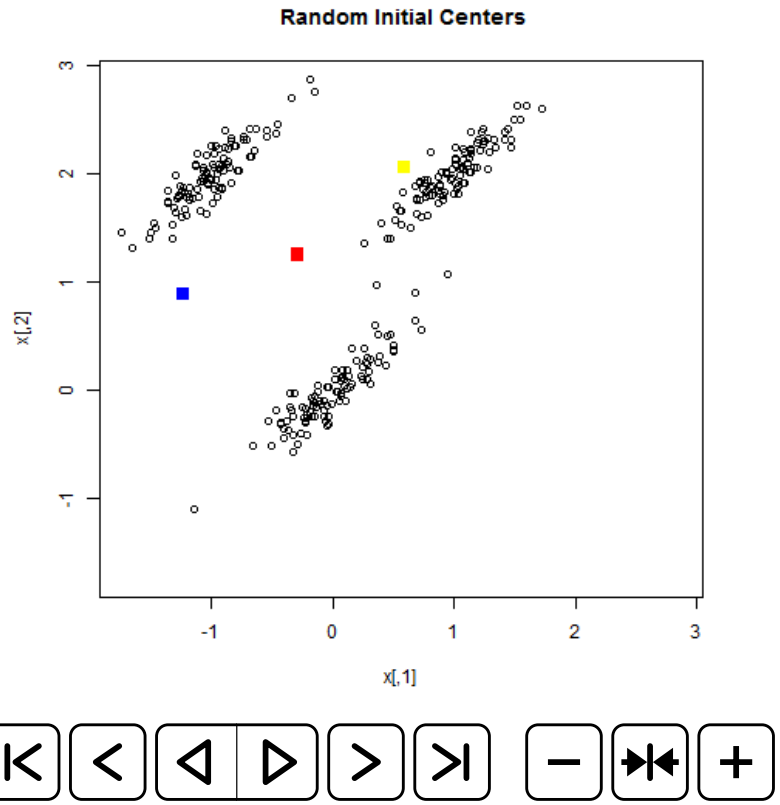
k-means

1. Choose arbitrary centers,
2. Repeat the two following steps until the centers do not change.
   1. assign each observation to the closest center,
   2. compute the new center of each class using the observations assigned to that class,
3. Repeat the previous steps 10 times and keep the partition $(G_{jk}, j = 1...)$ with the minimum within-cluster sum of squares given by,

$$\sum_{i=1}^{n}\sum_{j=1}^{k} \|\mathbf{x}_i - c_j\|^2 \mathbf{1}_{\{i \in G_j\}}$$

   where G_j is the j–th group and c_j is the corresponding center.
4. Create SRONG FORMS.

Kmeans



HastieKmeans1

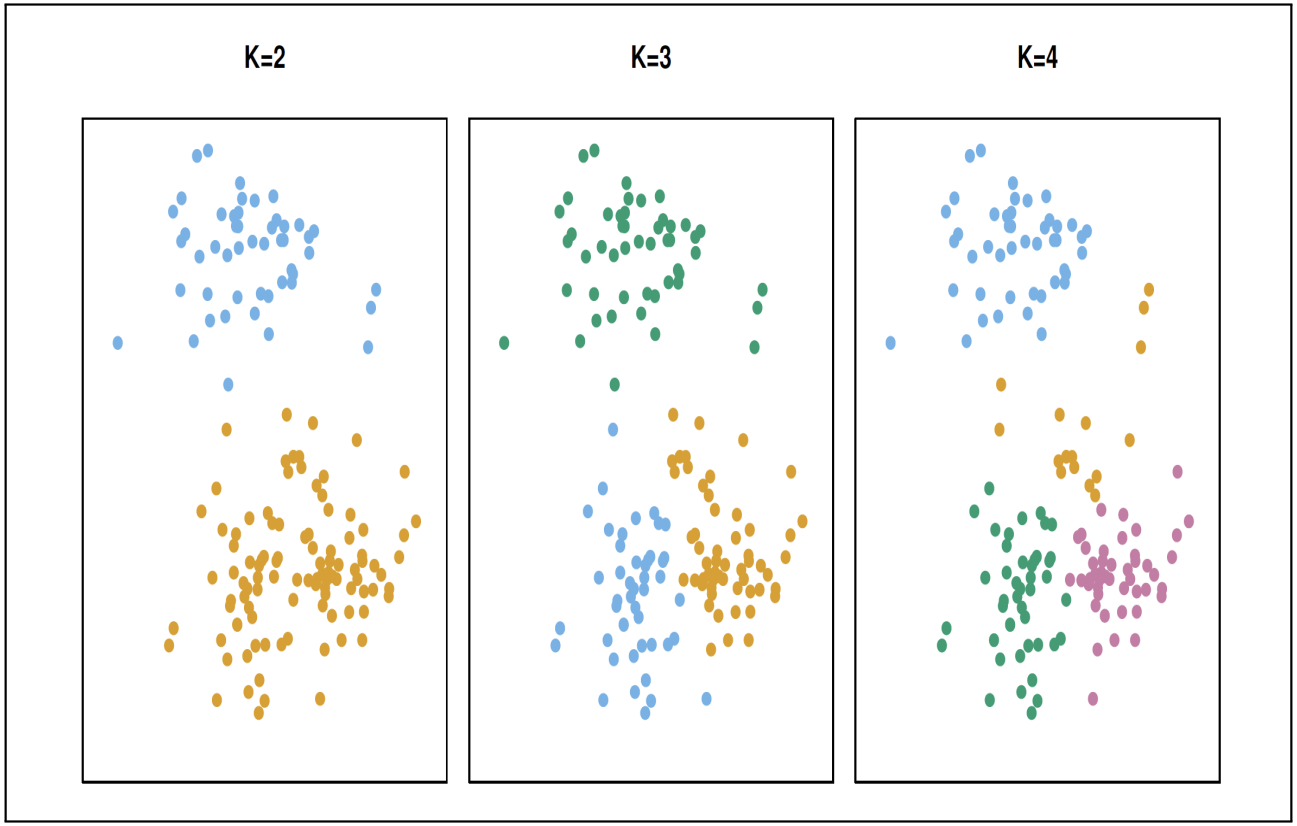


Figure: HastieKmeans1

HastieKmeans2

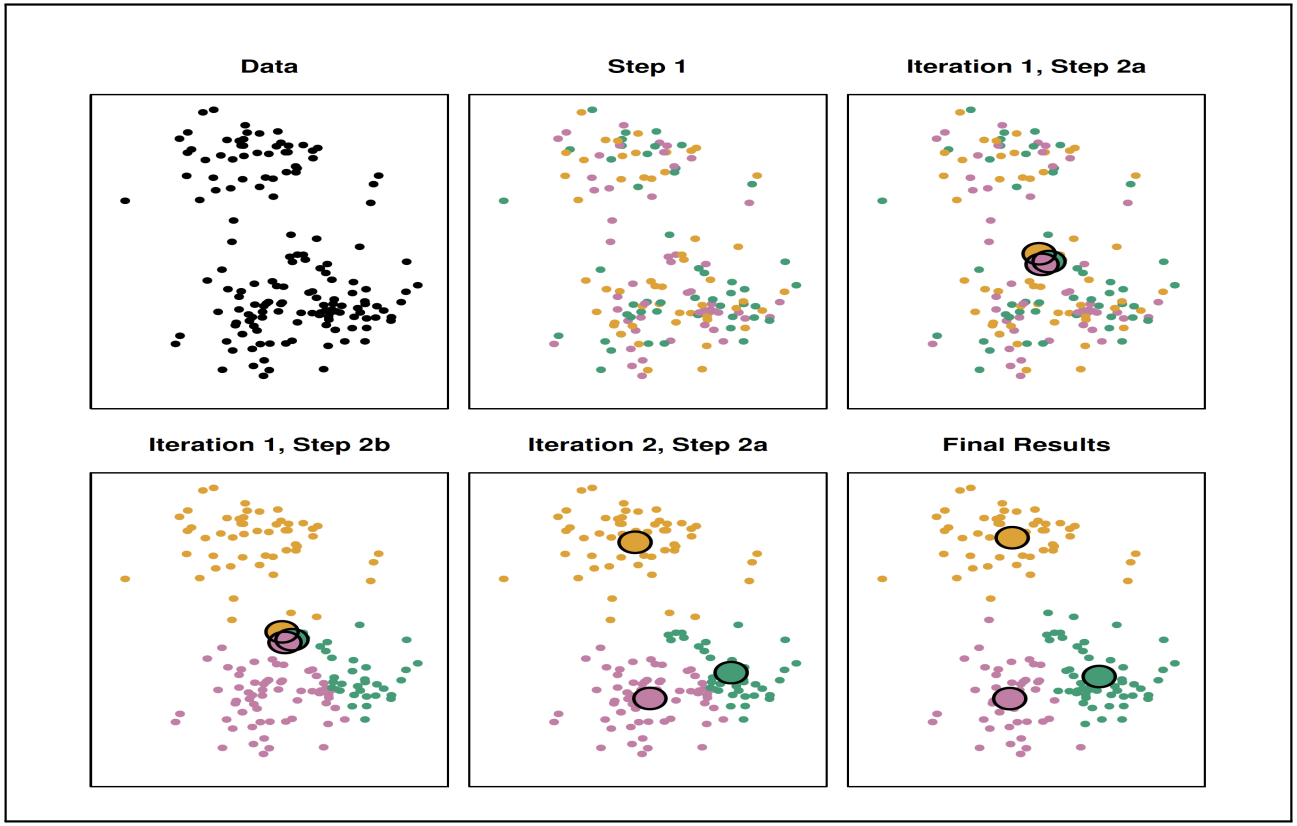


Figure: HastieKmeans2

HastieKmeans3

Figure: HastieKmeans3

k-modes/k-medians

Extensions of the well-known k-means algorithm [8].

- The k-modes algorithm [3] for categorical data that extends the k-means algorithm. It seeks to partition the observations into k groups such that the distance from the observations to the cluster modes is minimized. Simple-matching distance to assess the dissimilarity between pairs of observations is often used (Number of disagreements between pairs/Total number of pairs).
  See the *klaR* package [11] from **R**.
- The k-medians approach is recommended in [6] for addressing ordinal data. It is similar to the k-means algorithm except that it uses medians rather than means as centers for the clusters. Manhattan distance ($\sum_i |x_i - y_i|$) may be used to assess the dissimilarity between the observations and the cluster medians.
  See the *flexclust* package [6] from **R**.

Hierarchical Clustering

Hierarchical methods [9] aim to construct a dendrogram, given a distance d, and they can be

1. ascendant ("bottom-top") or
2. descendant ("top-down").

In ascendant hierarchical clustering, the hierarchy is constructed by iterative aggregation of pairs of nearest clusters. The distance between two clusters may be defined in several ways, the most popular being

- simple
- average
- complete linkage.

For ordinal data one may use the Manhattan distance as the dissimilarity measure between the observations, or the mutual information for nominal data.

The algorithm is implemented in the *hclust* function from **R**.

PAM, CLARA , DAISY

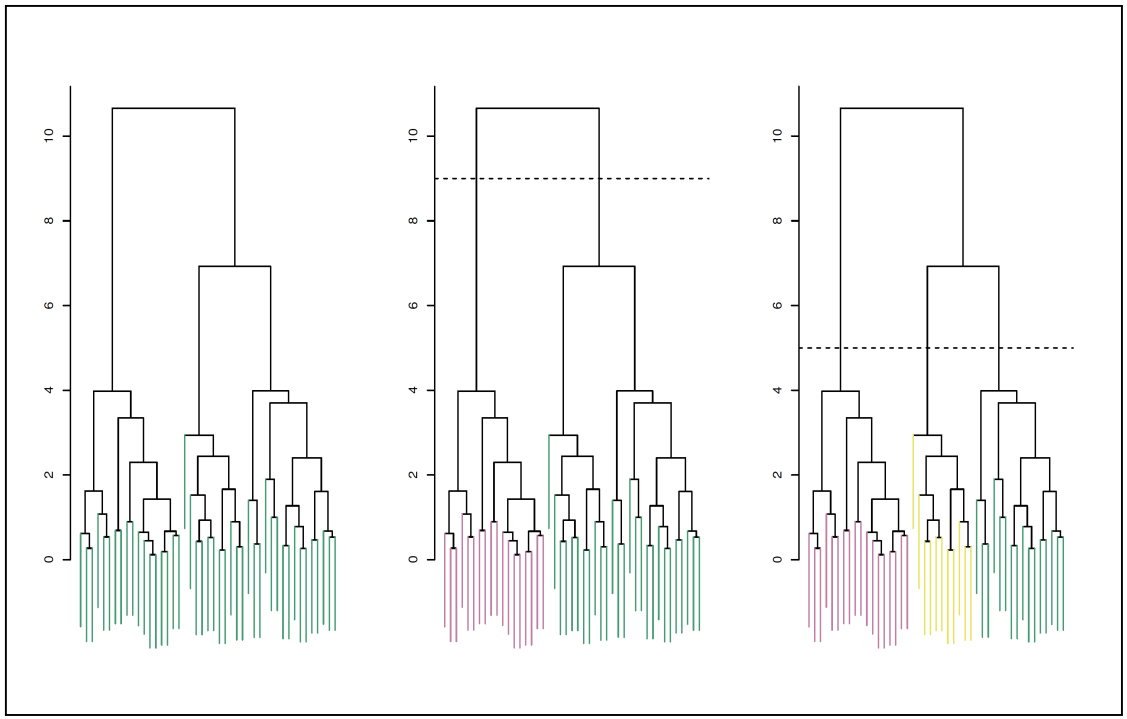


Figure: HastieCluster1

HastieCluster2

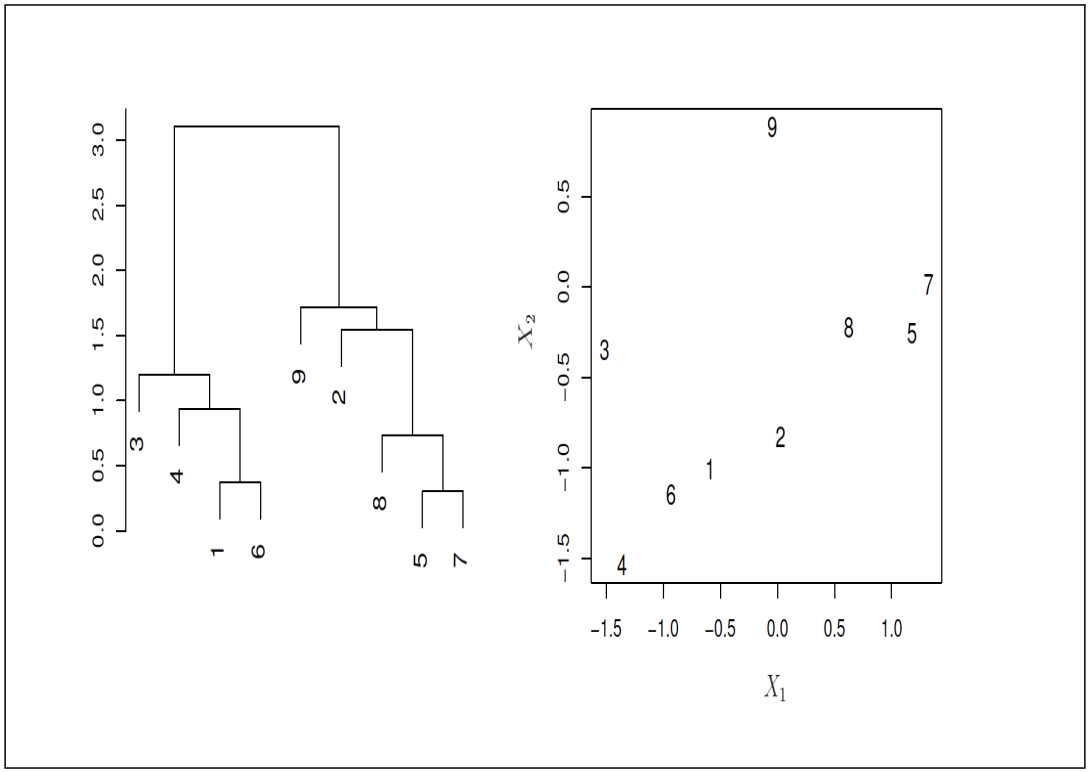


Figure: HastieCluster2

HastieCluster3

Types of Linkage

Linkage	*Description*
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities.
Average	Mean inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Figure: HastieCluster3

Category Utility ([3])

Consider a partition, $C = C_k$, $k = 1, ..., K$, found by a clustering algorithm based on given Variables X_j, $j = 1, ..., p$. The attributes are assumed nominal so that each X_j has a set of values or categories, x_{jl}. The category utility function scores partition C against the set of variables according to formula:

$$CU(C) = \frac{1}{K} \sum_{k=1}^{K} P(C_k) \left[\sum_j \sum_l P(X_j = x_{jl} | C_k)^2 - \sum_j \sum_l P(X_j = x_{jl})^2 \right]$$

Probability matching is random decision strategy using the empirical distribution, which is different from the bayesian majority vote prediction. The probability of winning when distribution is given by p_j, is $\sum_j p_j^2$

The matching error

We assess the performance using the matching error.
Let $y_1, ..., y_n$ be the class labels of each observation, and let $\hat{y}_1, ..., \hat{y}_n$ be the labels assigned to the n observations by a clustering algorithm. We denote by Σ the set of all possible permutations of the set of labels. The misclassification error rate, also called the "matching error" is defined as follows:

$$MCE = \min_{\sigma \in \Sigma} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}$$

DBSCAN

Density-based spatial clustering of applications with noise [4], a density-based, non-hierarchical algorithm.

- Fix two parameters: a reachability distance ϵ and a minimal number of points *MinPts*.
- Choose an arbitrary point P from the data and compute its ϵ-neighborhood, $V_\epsilon(P)$.
- If there are at least *MinPts* in $V_\epsilon(P)$, point P (said to be dense) and all its neighbors will form a cluster C, otherwise it is labeled as noise.
- Any point $P' \in V_\epsilon(P)$ which is dense, is also added to the cluster C, together with its own ϵ-neighborhood.
- Once no dense points are found, a new unvisited point is retrieved and processed to explore new clusters.

Manhattan distance is often used as the dissimilarity measure in the ordinal case and the mutual information in the nominal case. See the *dbscan* function of the *fpc* package [2] from **R**.

COBWEB ([5])

- A conceptual clustering approach based on the category utility (CU) measure.
- An incremental hierarchical algorithm is designed for qualitative data.
- Constructs a tree dynamically, inserting one individual at a time in the tree construction.
- At each individual insertion, four options are available options: inserting the individual to an existing cluster, creating a new cluster, merging two nodes, or splitting one node.
- Each option gives rise to a different partition whose CU is computed: the option maximizing the CU is selected.
- Loop over observations from the dataset, and stop when a minimal value of the category utility can not be exceeded.

The *Cobweb* function [1] of the *RWeka* package from **R** can be used.

DIVCLUS-T

DIVCLUS-T [2] is a top-down hierarchical clustering method. It is monothetic, i.e subsets of observations in the dataset are split using single variable. Its goal is to optimize the same criterion as for classical CAH using Ward's method. It is designed for quantitative, qualitative and mixed data. To compute the splitting criterion, i.e the within-class inertia, DIVCLUS-T uses the euclidean distance in the quantitative case while it uses the Khi-2 distance in the qualitative case. The main advantage of this method, in comparison with other hierarchical methods, is that in addition to providing a dendrogram, each node is labeled by its binary splitting rule. So the dendogram provided by DIVCLUS-T can be read exactly like a binary decision tree.
The algorithm is available with the *divclust* function of the *divclust* package from **R**.

Plan

Images

1. faire des classes d'images en utilisant toutes les méthodes de clusterisation non supervisé pour faire des classes
2. SEGMENTATION D'IMAGES: Couper l'image en zones homogènes.

Clustering de pixels à base de densité 1

Histogram-based methods (Wikipedia sur la segmentation d'image à base d'histogramme)

- Sont des méthodes très efficaces ne demandant qu'une seule passe au travers des pixels, pour calculer l'histogramme
- Ses pics et ses vallées servent à localiser les classes se basant soit sur la couleur soit sur l'intensité
- Peut être utilisé itérativement pour créer des classes plus fines
- Les pics et vallées peuvent être difficiles à déterminer
- Utilisables pour des multi-frames servant en vidéo pour suivre des mouvements

Dual Clustering

Méthode elle aussi basée sur l'histogramme des pixels (Wikipedia sur la segmentation d'image à base d'histogramme)
La méthode combine 3 charactéristiques de l'image La partition de l'image est construite sur les pics et les vallées de l'histogramme L'homogénéité de la partition est contrôlée par la compacité des clusters et la taille des gradients sur les bords des clusters.
Deux espaces sont introduits

- Le 1ř contient l'histogramme de brightness $H = H(B)$, qui permet de définir le seuil T de clusterisation en 2 classes (blanc,noir)
- le 2ř est l'espace dual de l'image originelle $B = B(x, y)$ à 3-dimensions. qui contient la bitmap-de-l'image est $b = \Phi(x, y)$, where $\Phi(x, y) = 0$, if $B(x, y) < T$, and $\Phi(x, y) = 1$, if $B(x, y)?T$.]

Une statistique mesure la compacité du bitmap.

Il reste à définir de bons bords en choisissant T.
Le seuil T étant donnée, on pose $M(T) = G/(k - L)$ has to be calculated where

- k est lé différence entre la brightness de la classe et celle du background,
- L est la longueur du bord, G la moyenne du gradient sur le bord.on the borders.

$T^{opt} = argmax_T M(T)$ est le seuil optimal en 2 classes.

LC simulation, Model M1

We set: $K = 3$, $p = 9$, $n \in \{100, 300, 500, 1000\}$, each variable has $m = 5$ levels, and clusters are equally sized.
Three clusters are defined, each characterized by a high frequency of one level: Level 1 (resp. 3 and 5) is the most frequent for cluster 1 (reps. 2 and 3). The other levels are uniformly distributed.
For example, The distribution for each variable in cluster one is given by:

$$\begin{aligned} P(X_{.j} = 1) &= q \\ P(X_{.j} = l) &= \frac{1-q}{m-1} \qquad \text{for} \qquad l \neq 1 \end{aligned}$$

We fix $q = 0.8$. This means, for each variable, the distribution of the five levels is: $(0.8, 0.2, 0.2, 0.2, 0.2)$ for the first cluster, $(0.2, 0.2, 0.8, 0.2, 0.2)$ for the second and $(0.2, 0.2, 0.2, 0.2, 0.8)$ for the third.
The same is done for the other clusters.

A tree model, M3

$p = 3$, $k = 4$. Each variable X_j, $j \in \{1, ..., p\}$, has $m = 6$ levels. Each level is coded as an integer, and we distinguish odd and even levels. The partition used for the simulation is shown in figure 9. Clusters are defined as follows:

- $C1$: x_1 and x_2 have odd levels, and x_3 is arbitrary
- $C2$: x_1 has odd levels, x_2 has even levels, and x_3 is arbitrary
- $C3$: x_1 has even levels, x_3 has odd levels, and x_2 is arbitrary
- $C4$: x_1 and x_3 have even levels, and x_2 is arbitrary

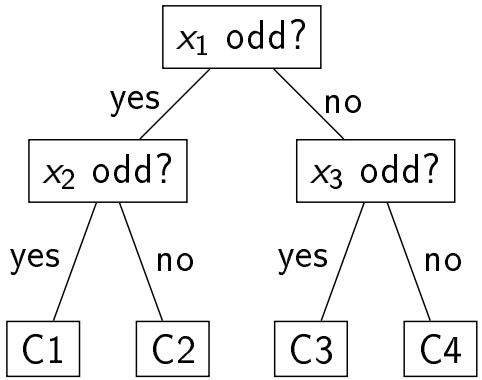


Figure: Tree structure used for data simulation model M3.

Another tree model, M4

$p = 3$, $k = 4$. Here, each variable $X_{.j}$, $j \in \{1, ..., p\}$, has $m = 4$ levels. The only difference with M3is that variable levels are not uniformly distributed in each cluster. Here, we consider a parameter p_0 that controls the non-uniformity of the distribution of levels, for example $p_0 = 0.8$. Clusters are defined as follows:

- $C1$: x_1 and x_2 have odd levels with $P(x_1 \in \{1, 3\}) = P(x_2 \in \{1, 3\}) = p_0$, and x_3 is arbitrary
- $C2$: x_1 has odd levels, x_2 has even levels with $P(x_1 \in \{1, 3\}) = P(x_2 \in \{2, 4\}) = p_0$, and x_3 is arbitrary
- $C3$: x_1 has even levels, x_3 has odd levels with $P(x_1 \in \{2, 4\}) = P(x_3 \in \{1, 3\}) = p_0$, and x_2 is arbitrary
- $C4$: x_1 and x_3 have even levels with $P(x_1 \in \{2, 4\}) = P(x_3 \in \{2, 4\}) = p_0$, and x_2 is arbitrary

Les méthodes de cette partie

- ACP
- ICA comme extention
- Multidimensional Scaling de tableau de distances

Plan

Quelle est la question? ACP/ICA (1)

Des mesures (p) sont faites à l'aide de descripteurs sur des individus (n). Certaines sont utiles, d'autres superflues.

- ACP: construire un petit nombre de variables NOUVELLES qui discriminent au mieux les descriptions des individus. Elles seront appelées les Composantes principales
- ICA comme extension: séparer les sources.

Quelle est la question? MDS (2)

- Multidimensional Scaling de tableau de dis-similarités. La connaissance de départ est un indice de dis-similarité entre les couples de descripteurs initiaux, en général des distances (plus la distance est grande plus les différences sont grandes). Appelons *DIS* ces informations.
  Le problème est de construire de nouveaux points (appelés $(i^{repr}, i = 1...n)$ admettant une description simple couplés aux individus $(i, i = 1...n)$, et pour lesquels il sera facile de lire sur des figures la dissimilarité DIS^{repr}. On veut que ces nouveaux points possèdent la propriété $DIS^{repr} - DIS^{repr}$ soit globalement petits.

PCA.Hastie 1

PCA produces a low-dimensional representation of a dataset.
It finds a sequence of linear combinations of the variables that

1. have maximal variance
2. mutually uncorrelated.

PCA also serves as a tool for data visualization and understanding
The 1rst principal component of a set of features
$X1, X2, \ldots, Xp$ is the normalized linear combination of the features

$$Z1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$$

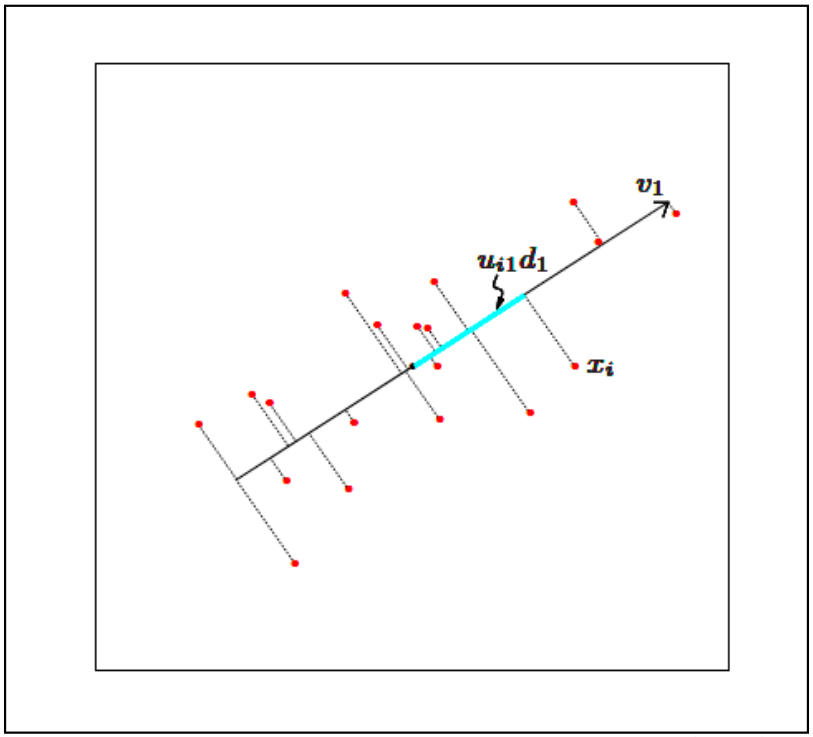
that has the largest variance. Moreover $\sum_{j=1}^{p} (\phi_{j1})^2 = 1$
The $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})'$ are the loadings defining the 1rst principal component loading vector,

Mettre un tableau $X_{n,p}$; Nuage

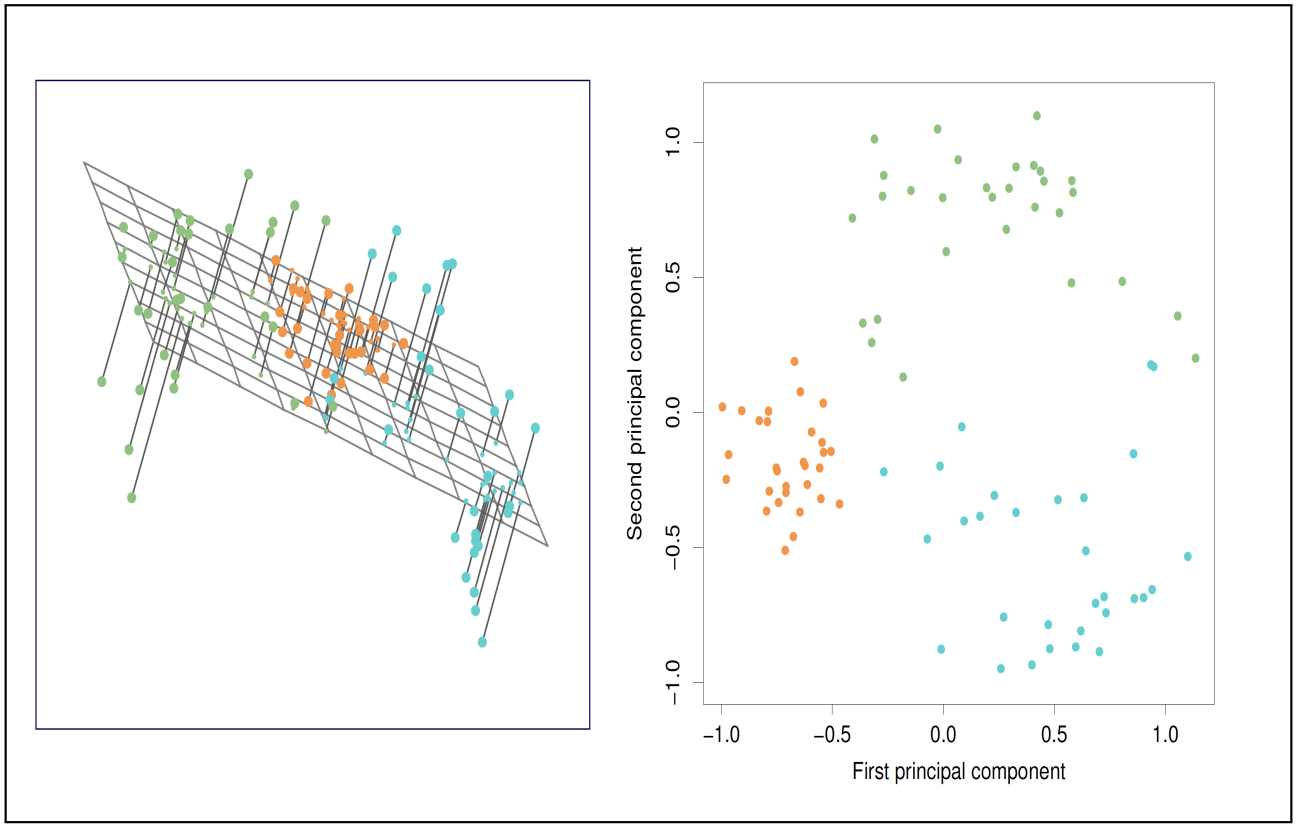
1. $X_{n,p}, n = 8, p = 2$
2. Tableau centré $scale(X)$ moyenne des colonnes$= 0$, var des colonnes $= 1$.
3. Tableau de correlation $R_{p,p} = (1/n)X'X$;
4. $R = \begin{pmatrix} \boxed{\begin{matrix} 1 & r \\ r & 0 \end{matrix}} \end{pmatrix}$
5. $V = (1/\sqrt{2}) \begin{pmatrix} \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} \end{pmatrix}$
6. $RV_1 = (1 + r)V_1; RV_2 = (1 - r)V_1;$
7. $CompoPrincip_1 = XV_1, CompoPrincip_2 = XV_2$ sont de moyenne nulle
8. $var(CompoPrincip_1) = (1/n)XV_1, CompoPrincip_2 = XV_2$

Nuage projeté 1D

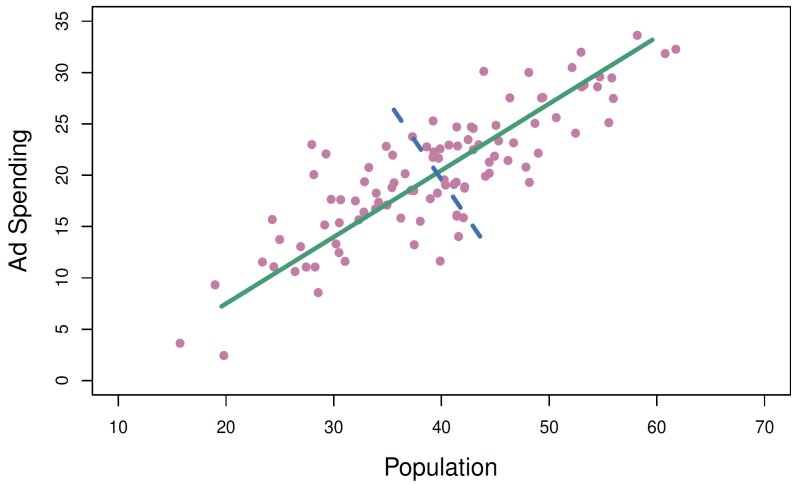
$a_i = Projection_{[a]}(X_i)$ Variance et Inertie
$var(x) = (1/n)\sum_{i=1}^{n}(x_i - \overline{x})^2$



Nuage projeté 2D



PCA Un axe. Intuition



The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.

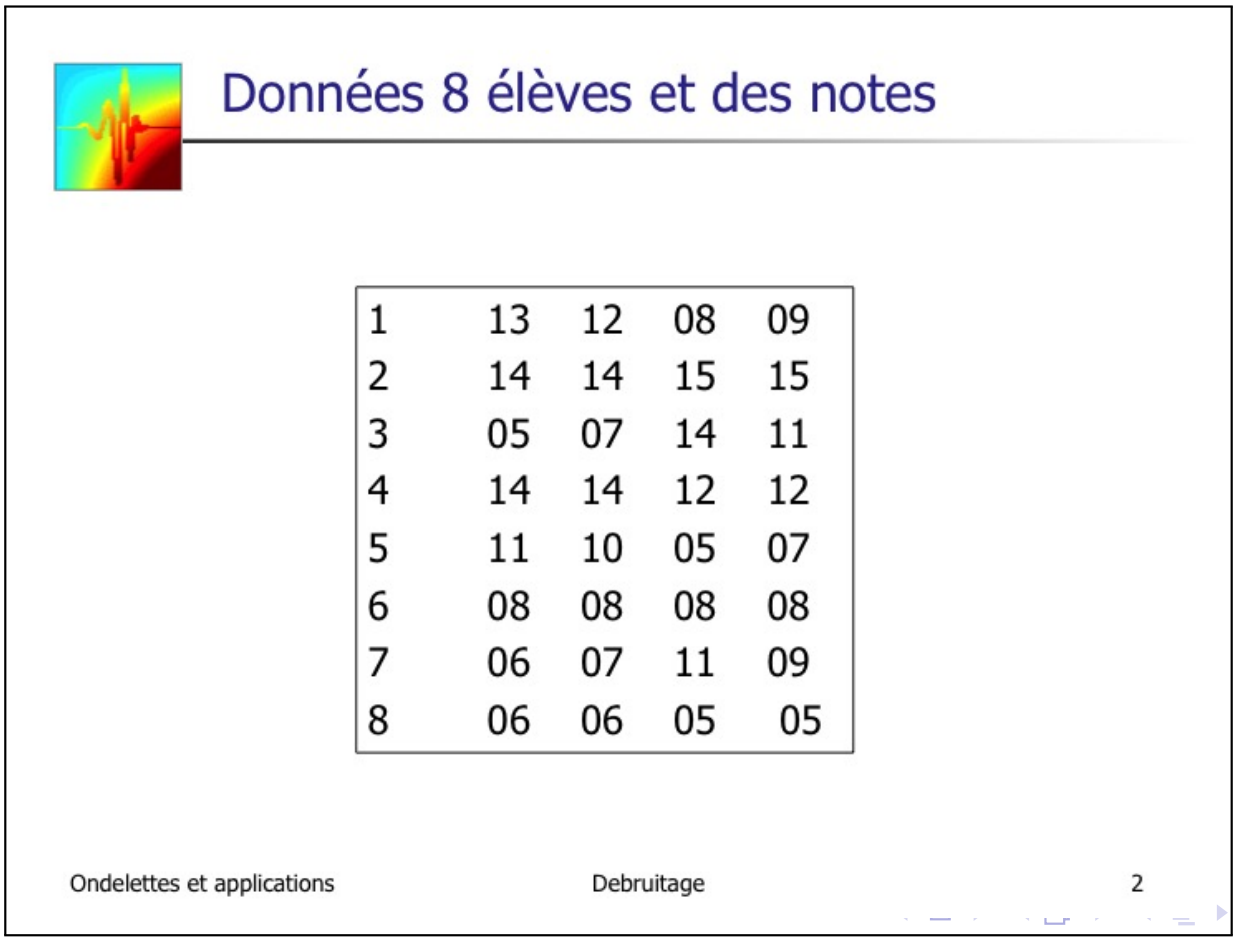
Figure: HastiePCA7

8 élèves, 4 matières de dimension 2 (1)

On a enregistré les notes de 8 élèves dans 4 matières
mathématique (V1), physique (V2), Français (V3), latin (V4).
Une ACP centrée réduite recherche les associations entre les matières.

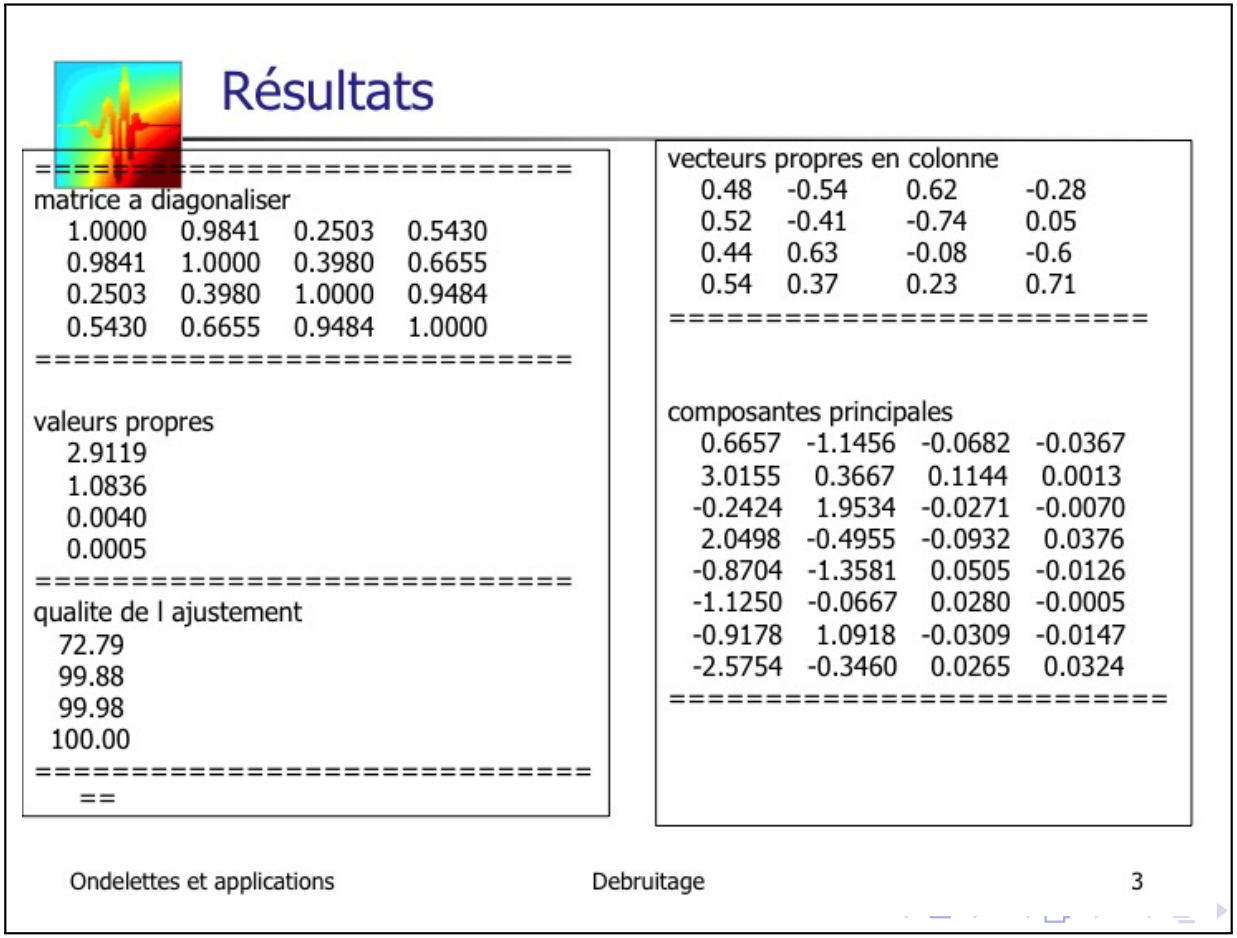
- Le premier axe est celui du facteur général, regroupement des 4 notes de chacun des élèves. Les coefficients sont ceux de la première colonne du tableau des vecteurs propres. Sur ce premier axe, l'ordre des élèves est celui des moyennes.
- Le second axe oppose les matières scientifiques V1 et V2 à V3 et V4, variables littéraires. Les élèves ayant de bons résultats dans les matières littéraires s'opposent à ceux dont les résultats sont bons dans les matières scientifiques.
- La dimension du problème est égale à 2.

(2)


Données 8 élèves et des notes
1 13 12 08 09
2 14 14 15 15
3 05 07 14 11
4 14 14 12 12
5 11 10 05 07
6 08 08 08 08
7 06 07 11 09
8 06 06 05 05
Ondelettes et applications
Debruitage
2

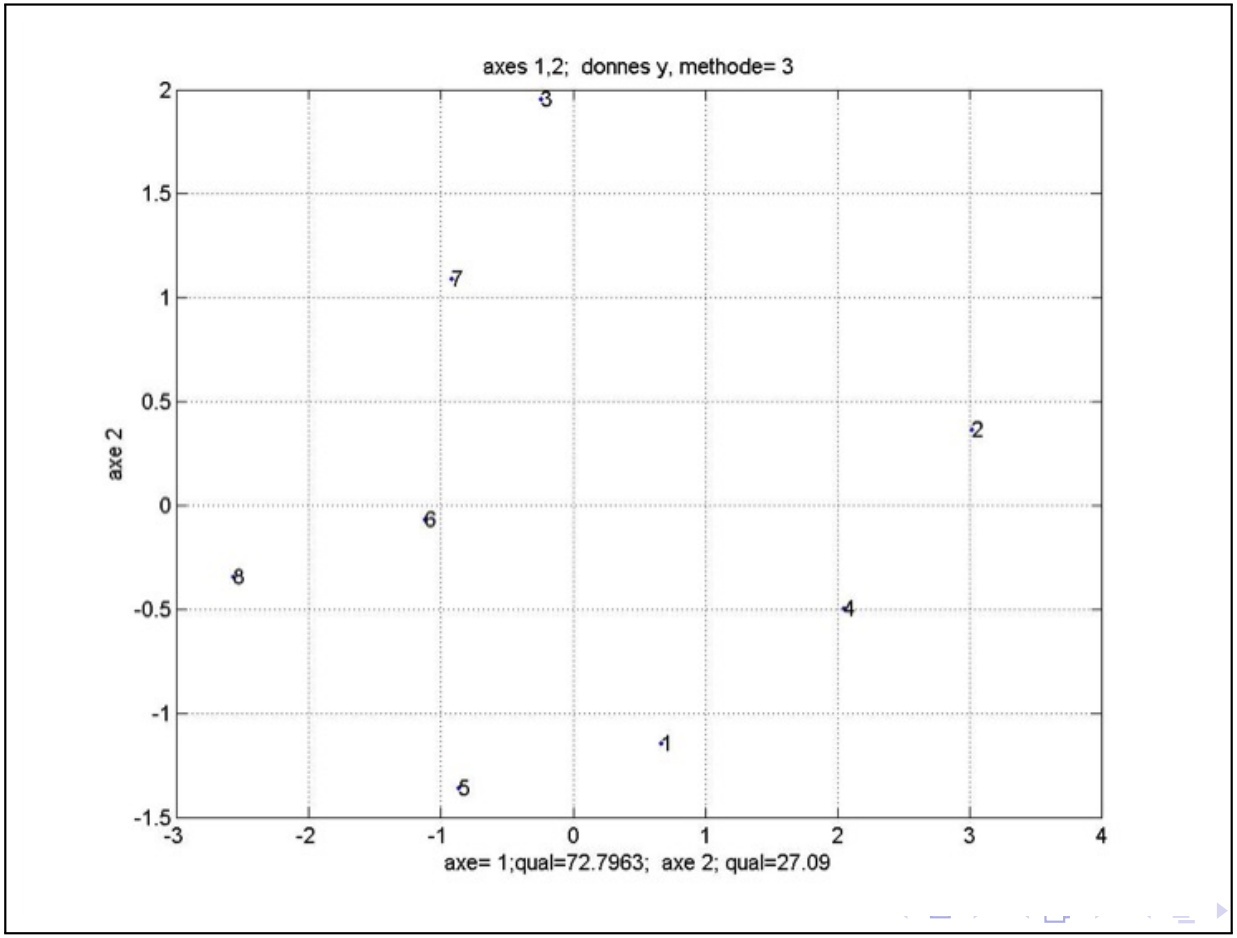
Exemple: Notes

(3)



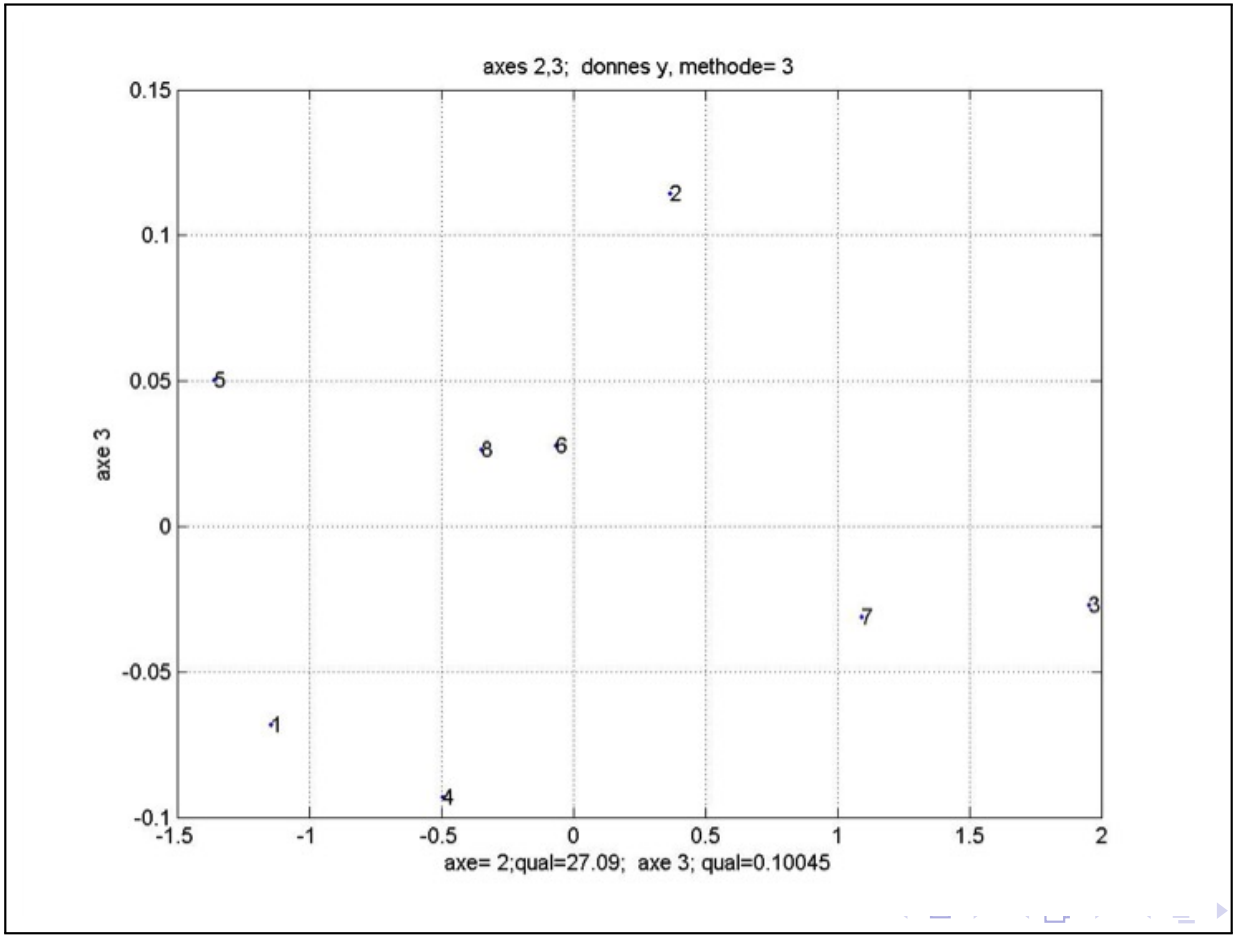
Exemple: Notes

(4)



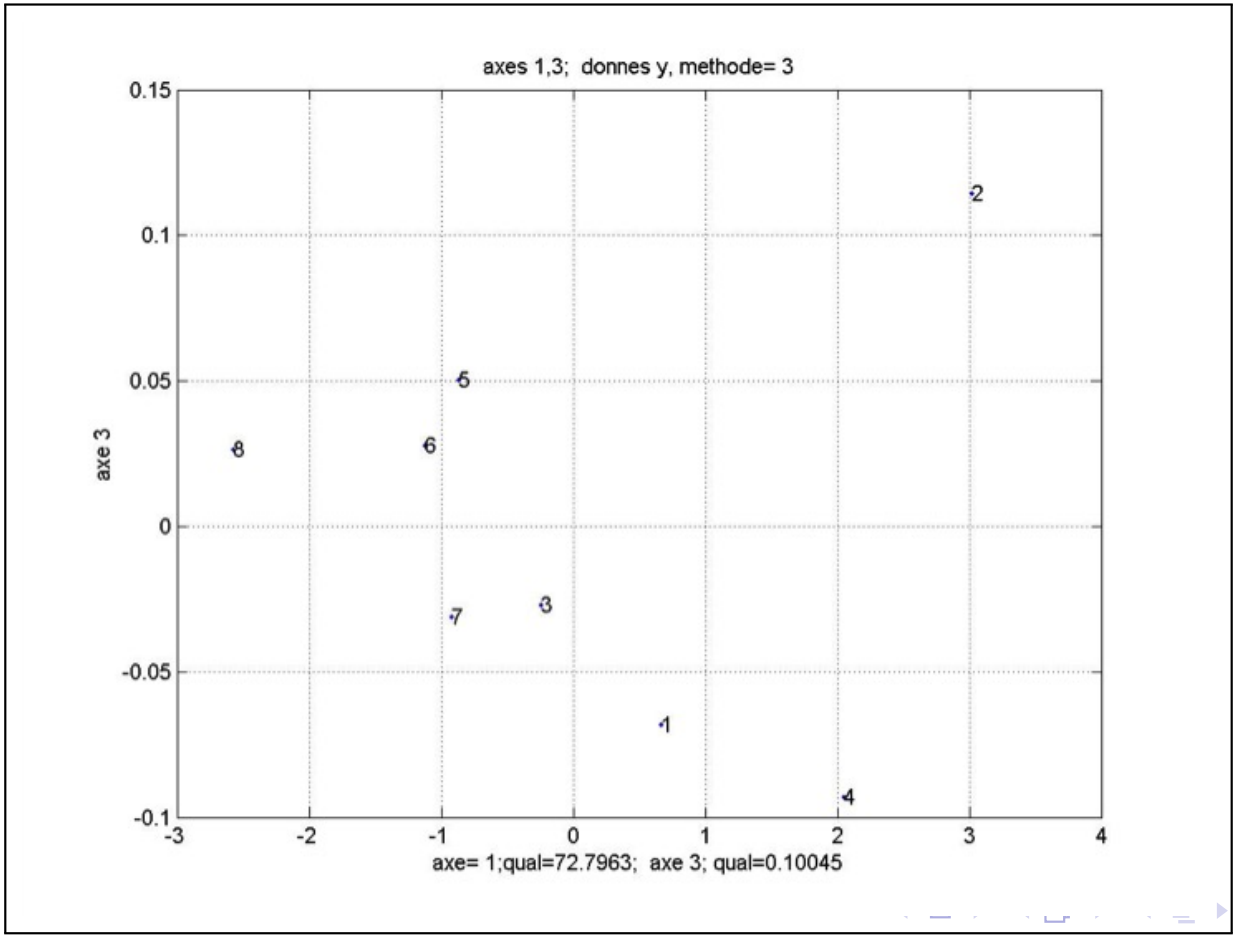
Exemple: Notes

(5)



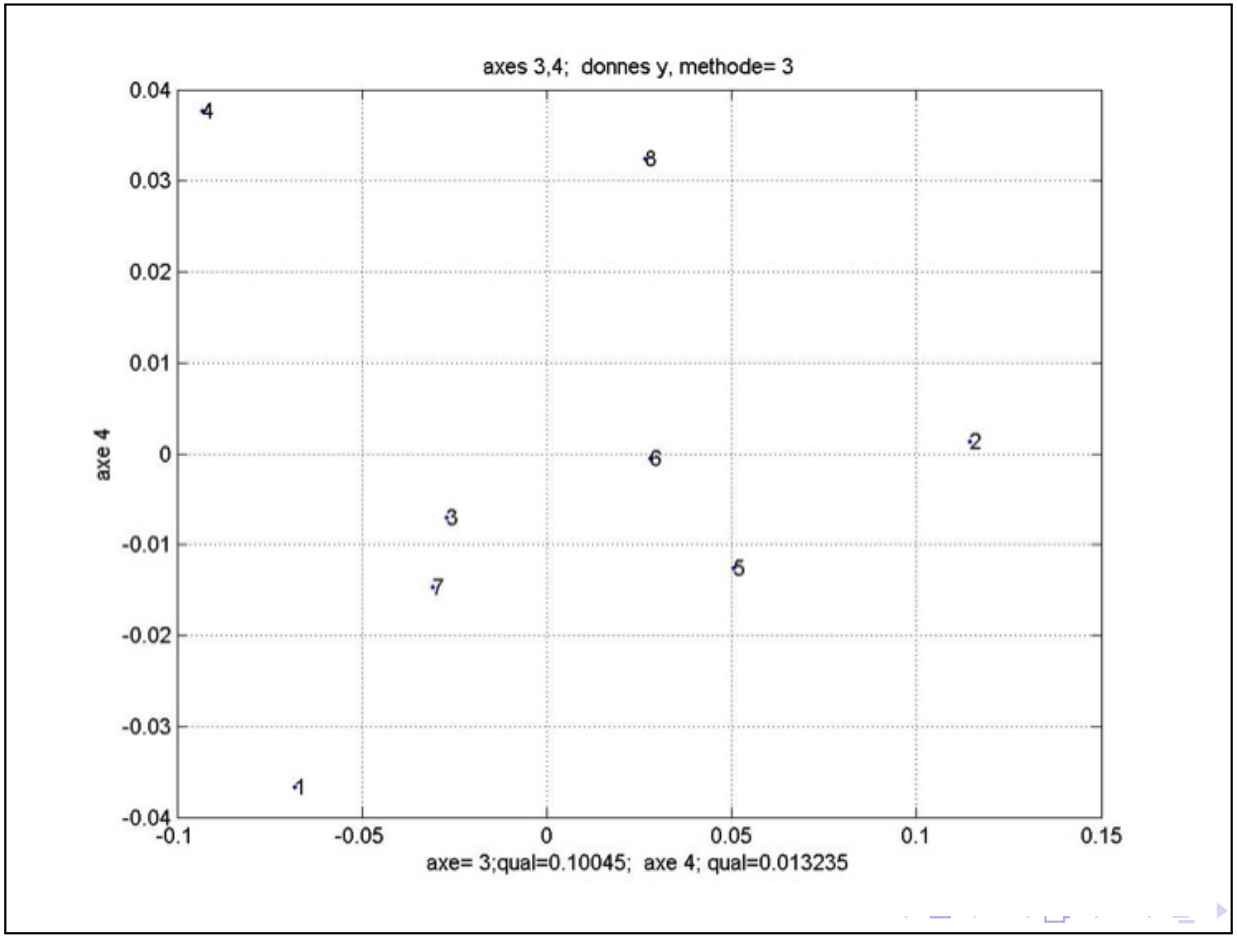
Exemple: Notes

(6)



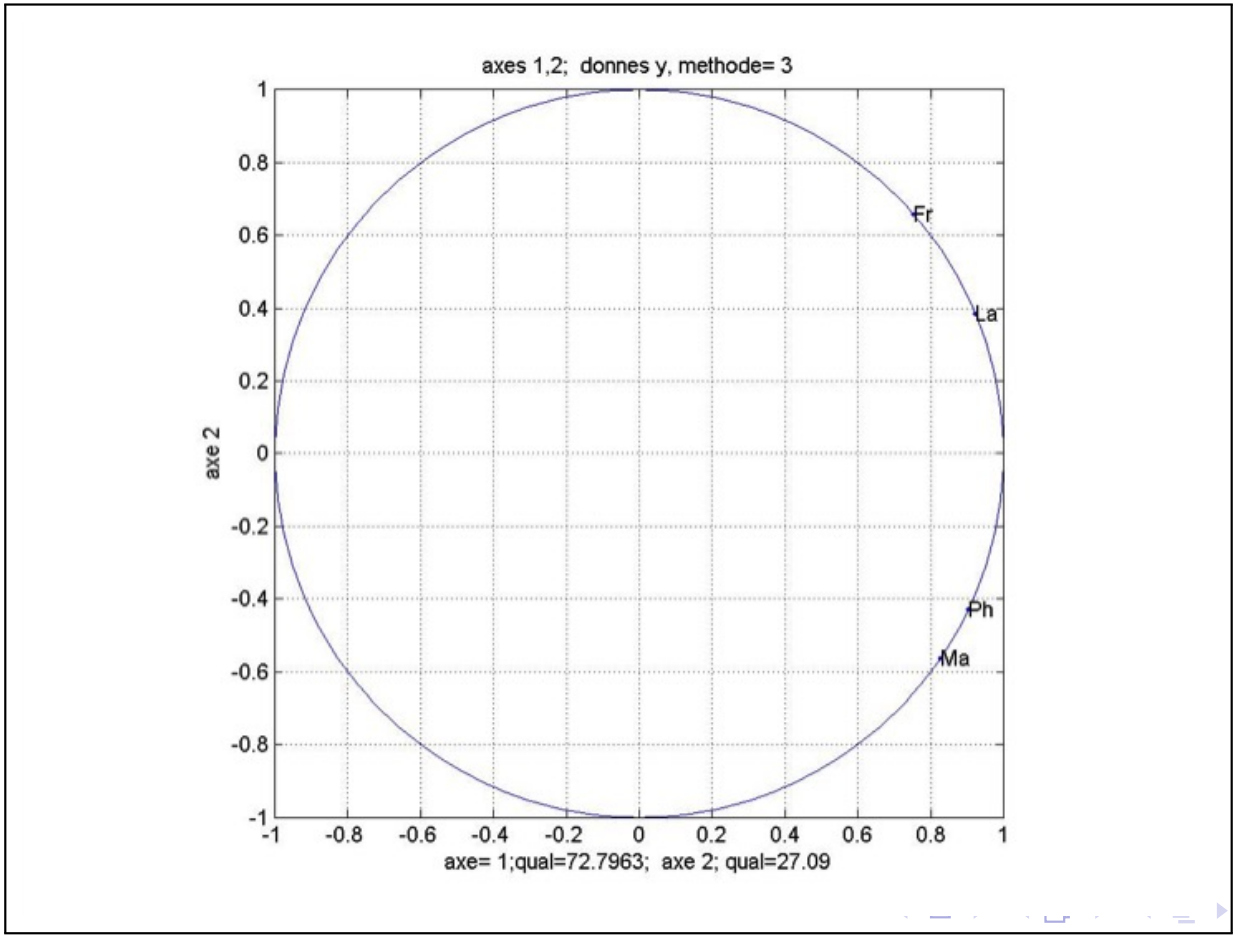
Exemple: Notes

(7)



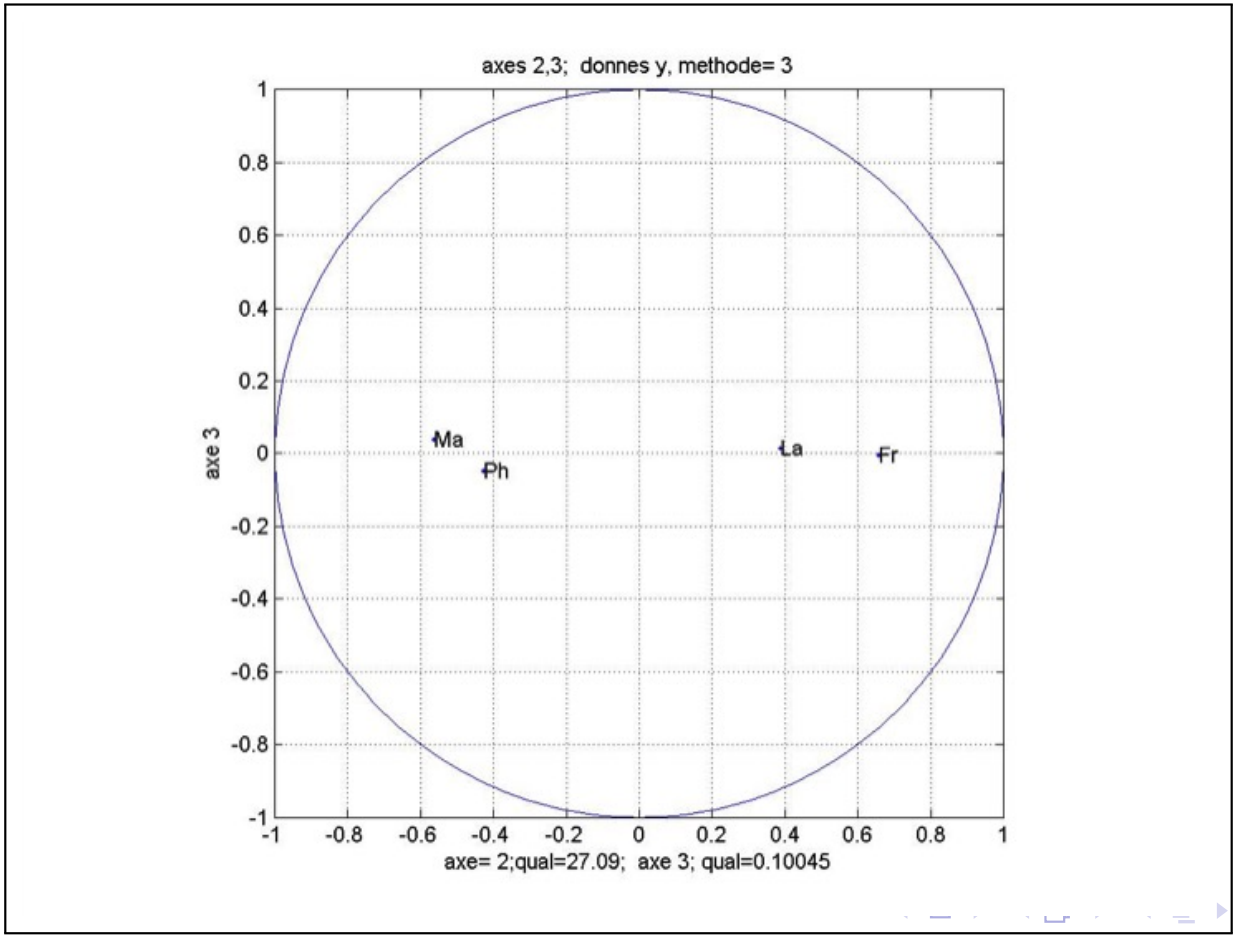
Exemple: Notes

(8)



Exemple: Notes

(9)



Les données des chiens

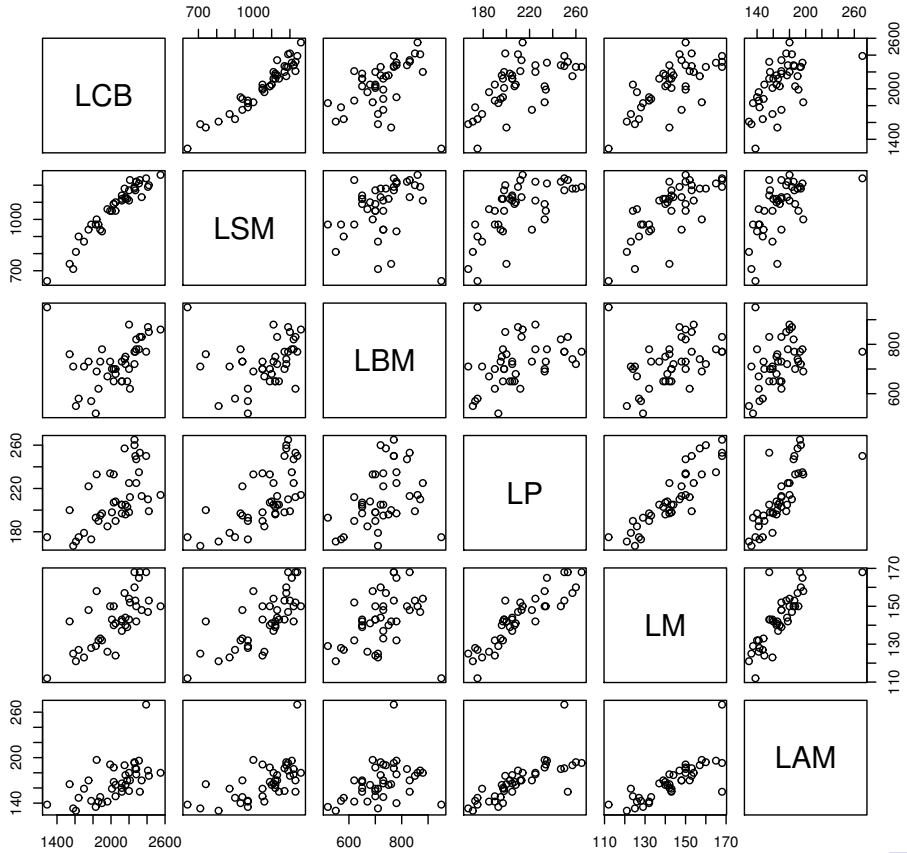
Les ossements de cranes de 42 canidés, identifiés comme chiens et loups, ont été mesurés. Les variables sont : Les lignes correspondent aux **individus** et les colonnes aux **variables**
Le nombre d'individu est noté n et le nombre de colonnes est noté p

- LCB : longueur condylo-basale
- LSM : longueur de la mâchoire supérieure
- LBM : largeur bi-maxillaire
- LP : longueur de la carnassière supérieure
- LM : longueur 1ř molaire supérieure
- LAM : largeur 1ř molaire supérieure

TYPE	LCB	LSM	LBM	LP	LM	LAM
BULL-DOG 1	1290	640	950	175	112	138
BULL-DOG 2	1540	740	760	200	142	165
BOXER	1580	710	710	167	125	133
SAINT-BERNARD	2200	1110	880	225	154	180
BULL-MASSIF	1900	930	780	197	132	140
DOGUE ALLEMAND 1	2410	1190	870	210	147	183
DOGUE ALLEMAND 2	2420	1200	850	199	153	176
SETTER 1	2010	1050	700	198	143	159
SETTER 2	1960	1060	670	185	126	142

Repris de Cours d'ACP. 2010. Nicolas Durrande - durrande@emse.fr. Script en R Cours 1 ACP. Repris de Durande 2010. Source : M.JAMBU - Classification

Exemple sur les données des chiens



Distances et Inertie

Distance et le produit scalaire euclidien et inierties sont définis par

$$\begin{aligned} d^2(x_i, x_j) &= ||x_i - x_j||^2 = \sum_{k=1}^{p} (x_i^k - x_j^k)^2 \\ \langle x_i, x_j \rangle &= \sum_{k=1}^{p} x_i^k x_j^k = X_i' X_j \end{aligned}$$

Inertie par rapport à un point a: $I_a = \sum_{i=1}^{n} \frac{1}{n} d^2(x_i, a)$
On a : $I_a = I_g + ||a - g||^2$ L'inertie du nuage par rapport à un sev $E \subset R^p$: $I_E = \sum_i \frac{1}{n} d^2(x_i, E) = \sum_i \frac{1}{n} d^2(x_i, P_E(x_i))$
Soit C La matrice de covariance empirique des variables

$$\begin{aligned} \mathrm{C} &= \sum_{i=1}^{n} \frac{1}{n} (x_i - g)^T (x_i - g) \\ \mathrm{C} &= PDP' \\ f_1 &= e_1 \mathrm{P}, ||f_1|| = 1 \\ I_{\Delta_1^\perp} &= e_1 \mathrm{P} \mathrm{D} \mathrm{P}^T e_1^T = f_1 \mathrm{D} f_1^T = \sum_{i=1}^{p} \lambda^i (f_1^i)^2 \\ &\leq max(\lambda^i) \sum_{i=1}^{p} (f_1^i)^2 = max(\lambda^i) \end{aligned}$$

Le maximum de $I_{\Delta_1^\perp}$ est atteint pour $f_1 = (1 \ 0 \ \ldots \ 0)$. $e_1 = f_1 P'$ donc e_1 est le premier vecteur propre de C.
Le premier VP de C est appelé le 1^{er} **axe factoriel** ou la 1^{ere} **direction principale**.
Les valeurs des projections des données suivant ses axes s'appellent les **composantes principales**
On cherche ensuite un axe Δ_2 porté par e_2, orthogonal à Δ_1 tel que l'inertie du nuage par rapport au plan $(g, \vec{e_1}, \vec{e_2})$ soit minimale.

$$\begin{aligned} \Delta_2 &= argmax(I_{(\Delta_1 \oplus \Delta_2)^\perp}) \\ &= argmax(\sum_i \frac{1}{n} d^2(x_i, (\Delta_1 \oplus \Delta_2)^\perp)) \\ &= argmax(\sum_i \frac{1}{n} d^2(x_i, g) - d^2(x_i, \Delta_1) - d^2(x_i, \Delta_2)) \end{aligned}$$

théorème :

Les k premières directions principales correspondent aux VP de la matrice de covariance associés aux k plus grandes vp. Elles sont deux à deux orthogonales.

Les plans des 2 axes factoriels sont les plans factoriels. Les variables associées aux axes sont le CP.

Retour sur l'exemple des chiens-loups

Peut-on interpréter ces résultats ?

- Quels sont les liens entre les nouvelles variables et les anciennes ?
- Les nouvelles variables ont elle un sens ?

On va pour cela s'intéresser à la corrélation entre composantes principales et variables.

Retour sur l'exemple des chiens-loups

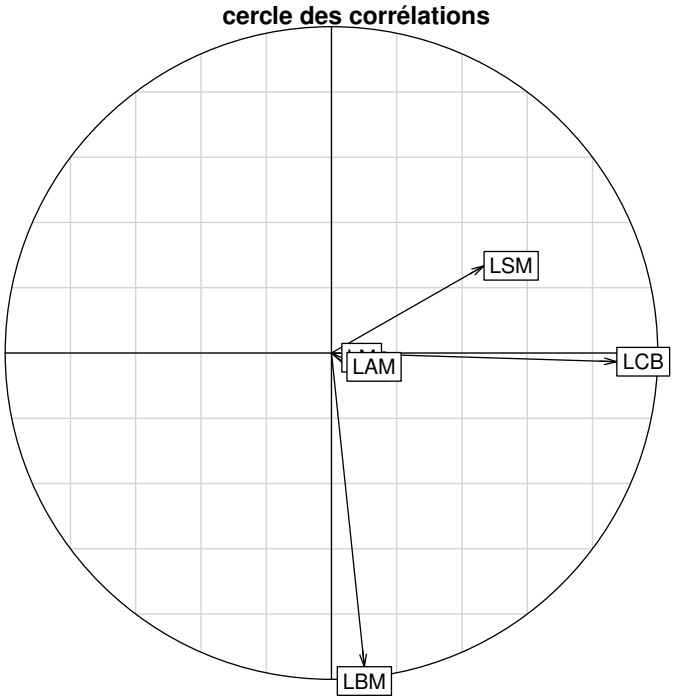
Notons $X_c = X - g$ la matrice des données centrées. Les composantes principales Y^1, Y^2, ..., Y^p sont donc les colonnes de la matrice $Y = X_c \mathrm{P}$. On en déduit que la corrélation entre X^i et Y^j vaut

$$corr(X^i, Y^j) = \frac{\sqrt{\lambda_j}}{\sqrt{var(X^i)}} p_i^j$$

On peut alors représenter les résultats que l'on vient d'obtenir sur un graphique appelé **cercle des corrélations**.

Retour sur l'exemple des chiens-loups

Cercle des corrélations



Problèmes posés

- Les résultats dépendent de la variance des variables de départ
- Si les variables non homogènes, il y a un problème de sens physique

Pour s'affranchir de ces deux problèmes, on va utiliser la méthode d'**ACP normée**.

ACP normée

Centrer et réduire une variable=soustrait sa moyenne et on la divise par son écart type : $X_{cr} = \frac{X - mean(X)}{sd[X].}$ Cela répond aux deux problèmes que l'on avait.

- Les résultats ne dépendent pas de la variance des variables
- Les données sont adimensionnées

Centrer-réduire les variables correspond à un changement de métrique.
$\tilde{x} = x - mean(x)$

$$d^2(x_i, x_j) = \sum_{k=1}^{p} \frac{(\tilde{x}_i^k - \tilde{x}_j^k)^2}{var(x^k)}. \quad (1)$$

Si on reprend la démonstration de l'ACP, les axes principaux sont les vecteurs propres de la **matrice R de corrélation**.

- L'inertie totale est égale à p
- Le cercle des corrélations représente les variables dans la base des directions propres

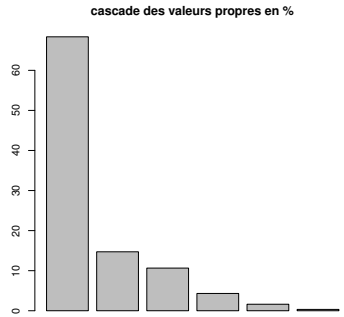
l'ACP centrée-réduite assure que les variables présentant de faibles variances ne sont pas "écrasées" par les autres.

Retour sur l'exemple des chiens-loups

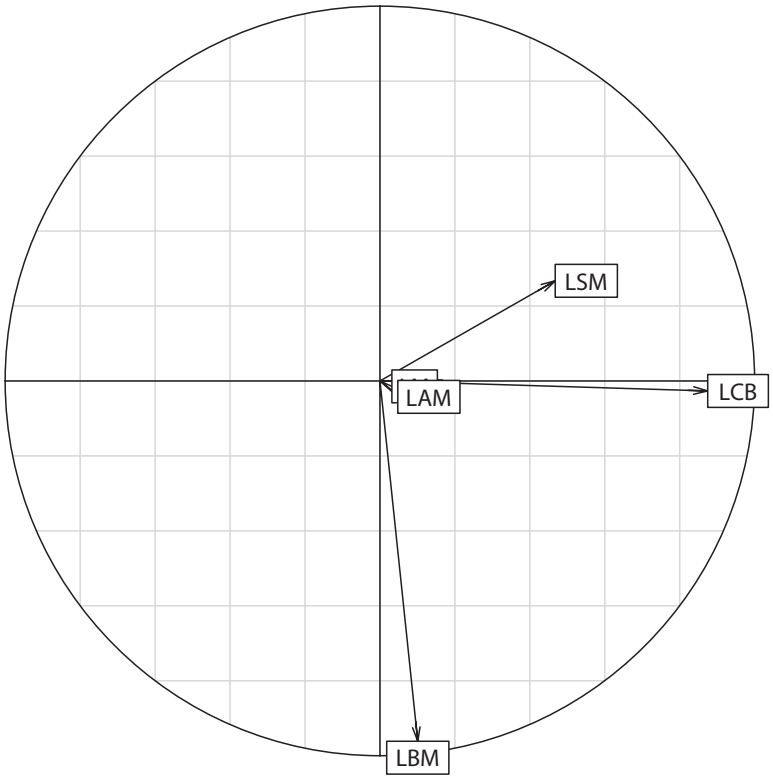
$$\mathbf{Cov} = \begin{pmatrix} 76467 & 40311 & 8899 & 4463 & 2718 & 4050 \\ 40311 & 23020 & 2803 & 2632 & 1524 & 2249 \\ 8899 & 2803 & 8522 & 896 & 441 & 816 \\ 4463 & 2632 & 896 & 689 & 320 & 499 \\ 2718 & 1524 & 441 & 320 & 186 & 268 \\ 4050 & 2249 & 816 & 499 & 268 & 621 \end{pmatrix}$$

La matrice de corrélation :

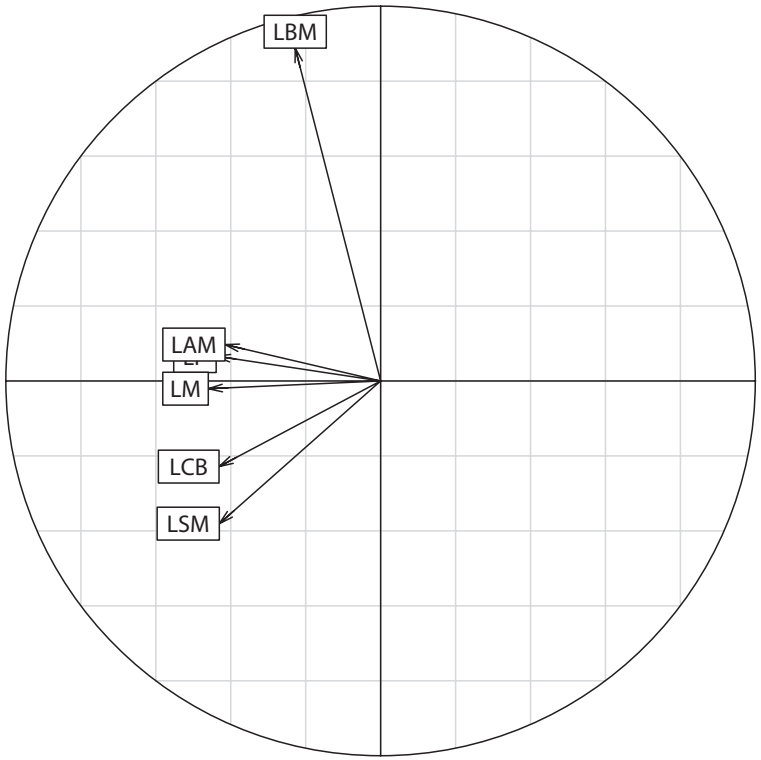
$$\mathbf{Corr} = \begin{pmatrix} 1.00 & 0.96 & 0.34 & 0.61 & 0.71 & 0.58 \\ 0.96 & 1.00 & 0.20 & 0.66 & 0.73 & 0.59 \\ 0.34 & 0.20 & 1.00 & 0.36 & 0.35 & 0.35 \\ 0.61 & 0.66 & 0.36 & 1.00 & 0.89 & 0.76 \\ 0.71 & 0.73 & 0.35 & 0.89 & 1.00 & 0.78 \\ 0.58 & 0.59 & 0.35 & 0.76 & 0.78 & 1.00 \end{pmatrix}$$



Directions trouvées



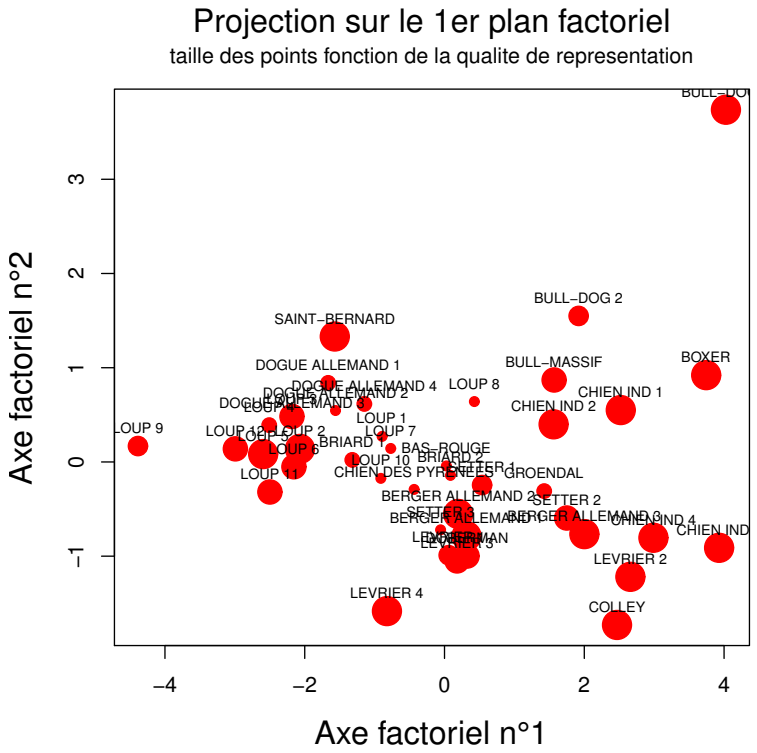
Cercle des corrélations ACP classique



Cercle des corrélations ACP normée

Retour sur l'exemple des chiens-loups

Dans l'espace des individus



Exemple 2

Conditions de travail et au niveau de vie dans différentes villes du monde.

On s'intéresse aux conditions de travail et au niveau de vie dans différentes villes du monde. On étudie $n = 46$ villes sur lesquelles on mesure $p = 3$ variables quantitatives :

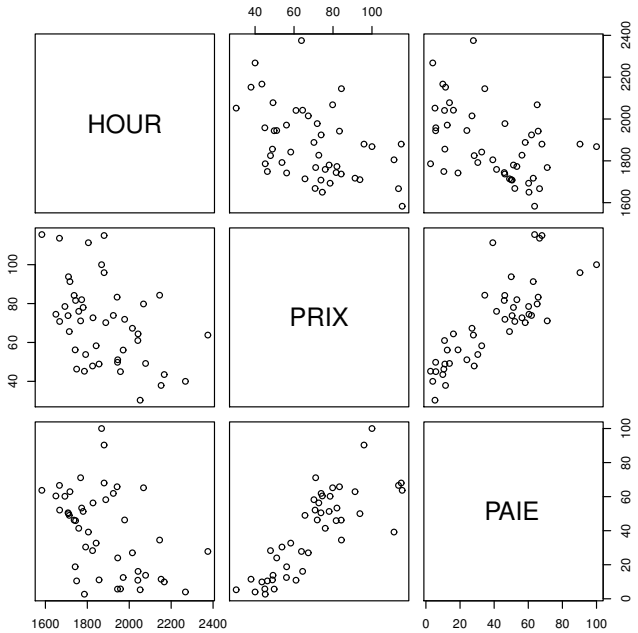
- HOUR : nb moyen d'heures de travail dans 12 activités
- PRIX : indice des prix sur la base de 112 produits et services
- PAIE : index des salaires horaires dans 12 activités, déductions faites

source : *Prices and earnings around the globe Economic Research Department*, Union Bank of Switzerland, Zurich.

Ville	HOUR	PRIX	PAIE
Amsterdam	1714	65.6	49.0
Athens	1792	53.8	30.4
Bogota	2152	37.9	11.5
Bombay	2052	30.3	5.3
Brussels	1708	73.8	50.5
Buenos Aires	1971	56.1	12.5
Caracas	2041	61.0	10.9
Chicago	1924	73.9	61.9
Copenhagen	1717	91.3	62.9
Dublin	1759	76.0	41.4
Dusseldorf	1693	78.5	60.2
Frankfurt	1650	74.5	60.4
Geneva	1880	95.9	90.3
...	...	...	...
Zurich	1868	100.0	100.0

Exemple 2

Diagramme pairs et matrice de corrélation

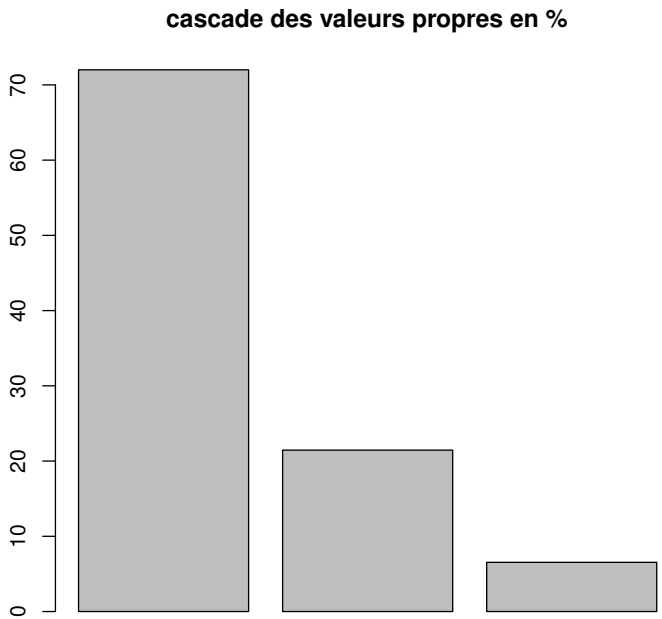


Matrice de corrélation :

	HOUR	PRIX	PAIE
HOUR	1.00	-0.45	-0.45
PRIX	-0.45	1.00	0.80
PAIE	-0.45	0.80	1.00

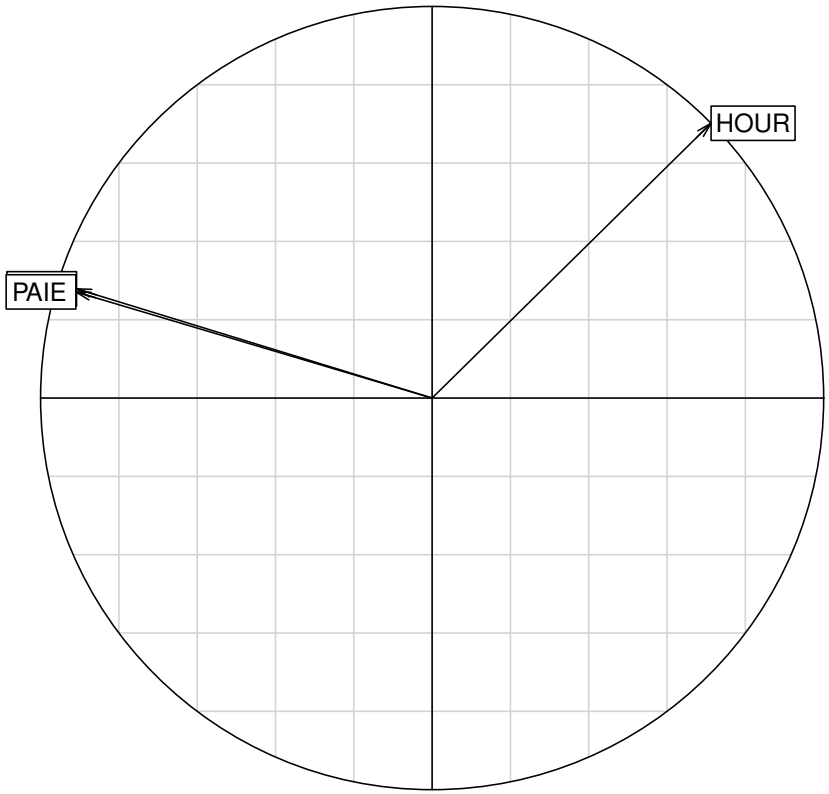
VP et vp

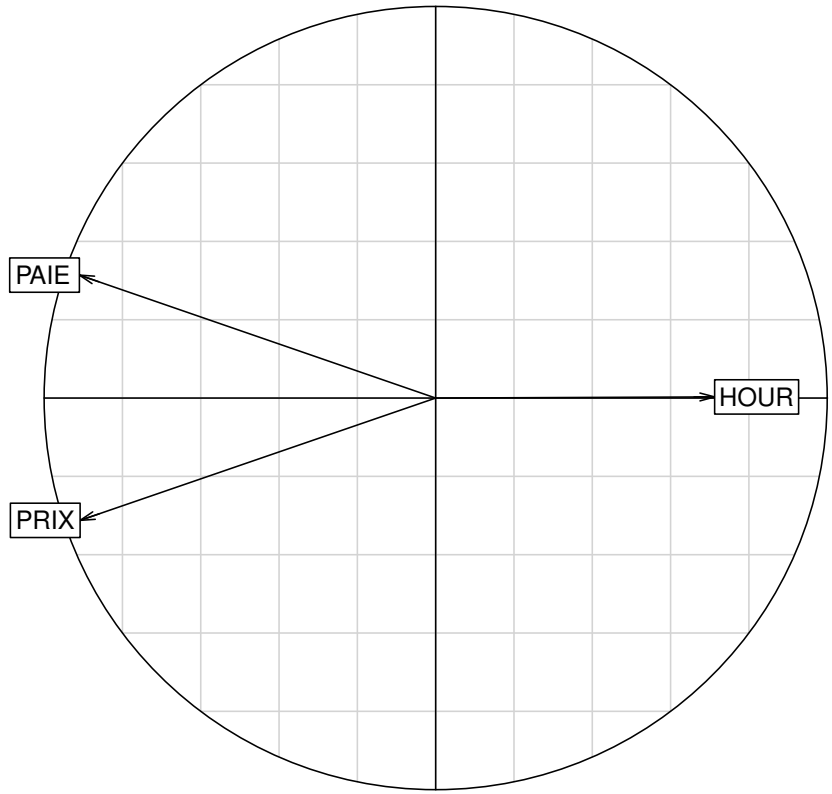
Valeurs propres et vecteurs propres



	VP1	VP2	VP3
HOUR	0.48	0.87	0.00
PRIX	-0.61	0.34	-0.70
PAIE	-0.61	0.33	0.70

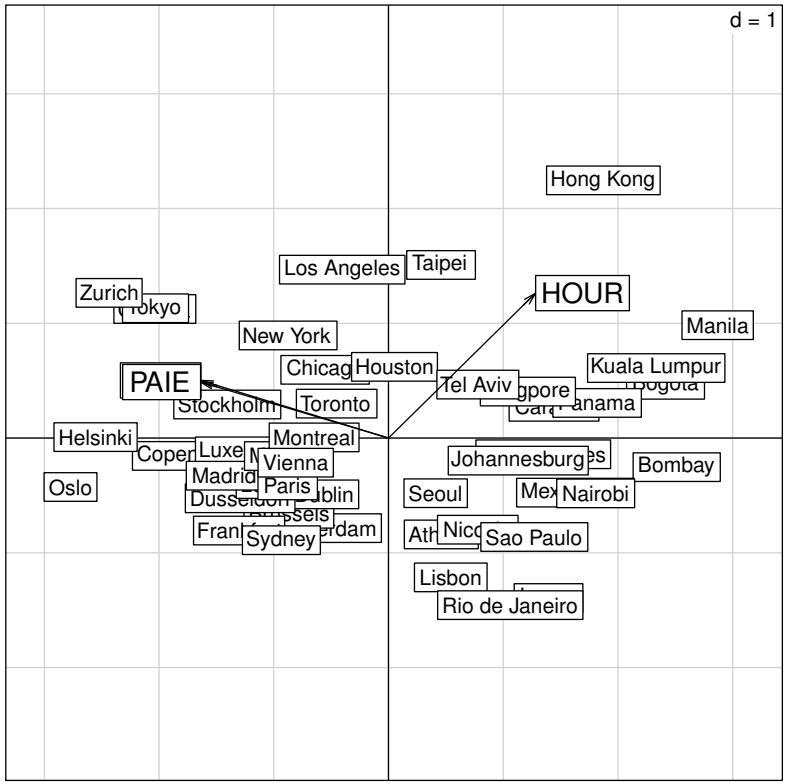
Exemple 2

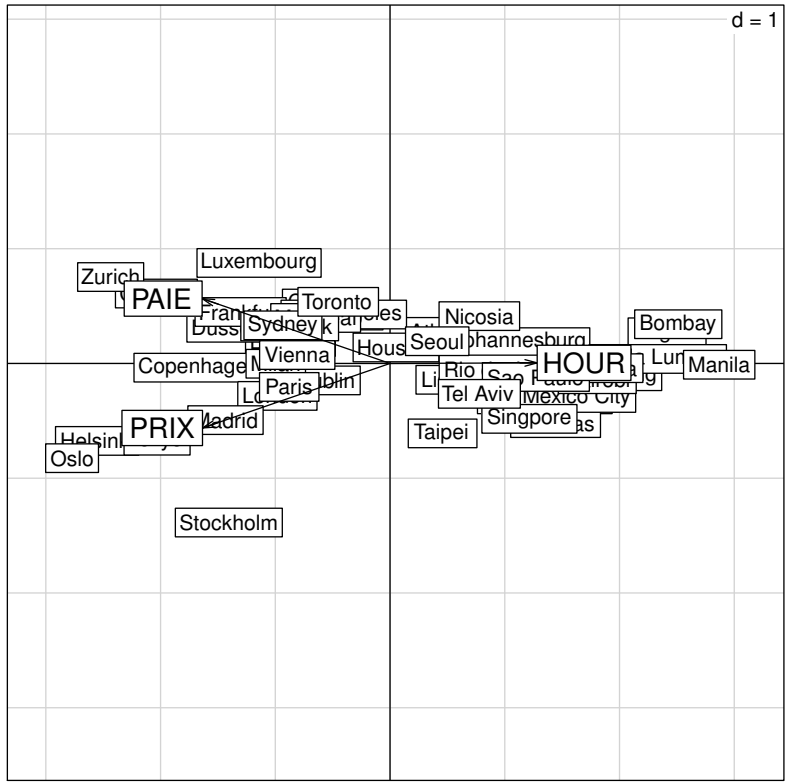




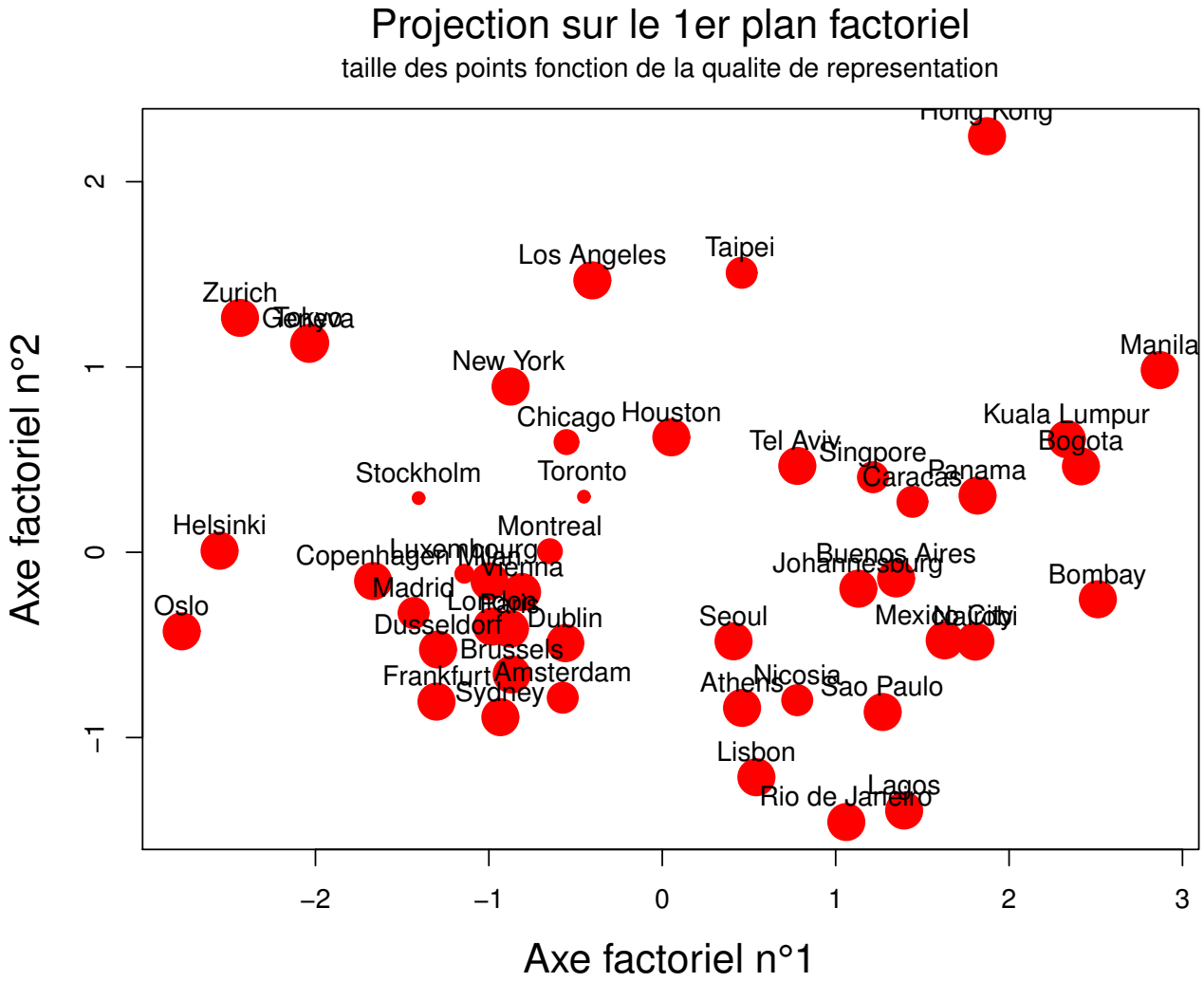
Peut on interpréter les directions que l'on a trouvées ?

Exemple 2


d = 1
Hong Kong
Los Angeles
Taipei
Zurich
HOUR
Manila
New York
Chicag
Houston
Tel Aviv
Kuala Lumpur
PAIE
Stockholm
Toronto
Helsinki
Montreal
Vienna
Johannesburg
Bombay
Madrid
Paris
Oslo
Seoul
Nairobi
Sydney
Sao Paulo
Lisbon
Rio de Janeiro


d = 1
Luxembourg
Zurich
PAIE
Toronto
Nicosia
Bombay
Sydney
Seoul
Vienna
HOUR
Manila
Copenhage
Rio
Paris
Tel Aviv
Singpore
PRIX
Taipei
Oslo
Stockholm

Exemple 2


Projection sur le 1er plan factoriel
taille des points fonction de la qualite de representation
Axe factoriel n°2
Axe factoriel n°1
Hong Kong
Los Angeles
Taipei
Zurich
Tokyo
Manila
New York
Chicago
Houston
Kuala Lumpur
Bogota
Tel Aviv
Singpore
Caracas
Panama
Stockholm
Toronto
Helsinki
Montreal
Copenhagen
Luxembourg
Vienna
Buenos Aires
Johannesburg
Bombay
Madrid
Oslo
Dublin
Seoul
Mexico City
Brussels
Frankfurt
Amsterdam
Sydney
Athens
Nicosia
Sao Paulo
Lisbon
Rio de Janeiro
Lagos

Plan

ICA: Analyse en Composantes Indépendantes

- voir *https* : *//fr.wikipedia.org/wiki/Analyse_en_composantes_indpendantes*
- Etudie le meme problème que l'ACP mais remplace l'obligation de n'etre pas corrélée par l'obligation d'être indépendant qui est plus forte.
- ACP sans sous-bassement gaussien ni quadratique
- Basé sur l'entropie mutuelle
- L'illustration classique de la séparation de sources est le problème suivant: Lors d'une soirée, on dispose de P microphones dans une salle où N personnes discutent par groupes de tailles diverses.
  - Chaque microphone enregistre la superposition des discours des personnes à ses alentours et
  - le problème consiste à retrouver la voix de chaque personne débarrassée des autres voix considérées comme parasites.
- $x = As + \epsilon$ où les inconnues sont les sources s.

ICA: Analyse en Composantes Indépendantes

- [Cardoso J.F] Il s'agit de méthodes traitant des observations multivariées) afin d'en extraire des composantes linéaires qui soient aussi indépendantes que possible. Cette simple idée s'est révélée très fructueuse pour le traitement des signaux multi-capteurs dans de nombreux domaines : réseaux d'antennes pour les télécommunications, prise de son, signaux biomédicaux multi-électrodes, et de manière plus générale dans tous les cas où un système de plusieurs capteurs, fournissant des signaux cohérents, est à l'écoute d'un ensemble discret de sources de signal que l'on cherche à extraire des observations et que l'on peut, pour des raisons physiques, supposer mutuellement statistiquement indépendantes.

ICA: Analyse en Composantes Indépendantes.Image

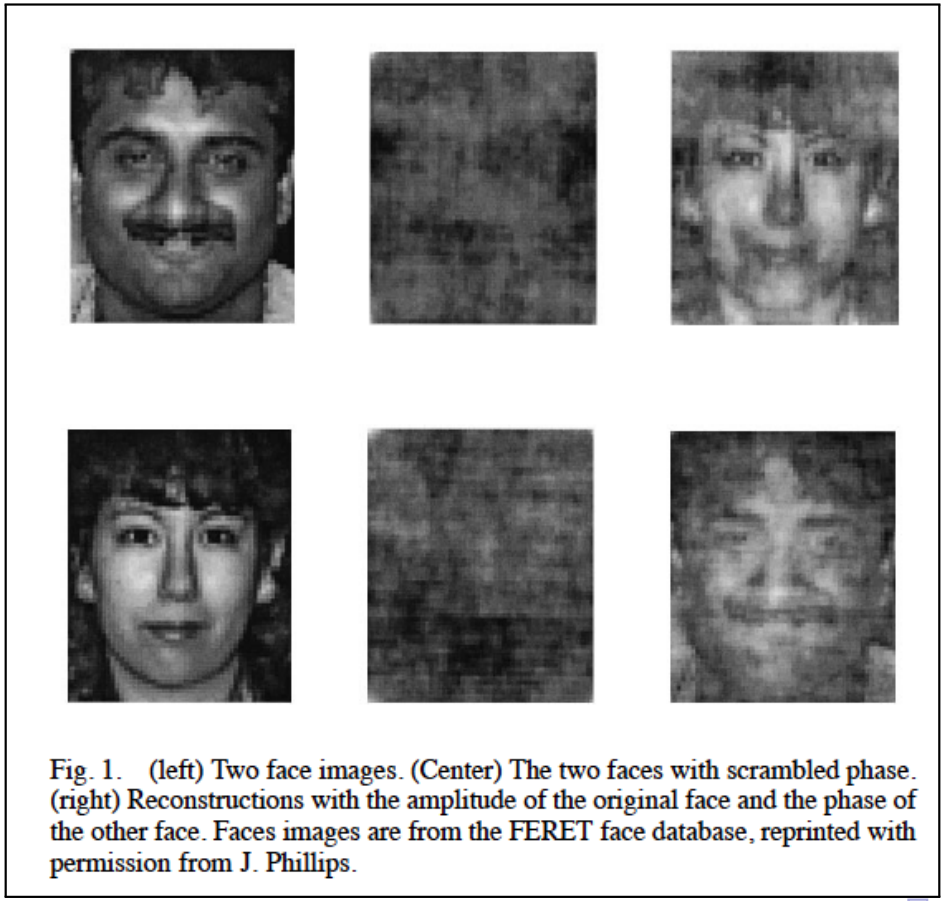


Fig. 1. (left) Two face images. (Center) The two faces with scrambled phase. (right) Reconstructions with the amplitude of the original face and the phase of the other face. Faces images are from the FERET face database, reprinted with permission from J. Phillips.

Plan

MDS 1

But de MDS

On se donne des indices $\delta_{ii'}, 1 \leq i, i' \leq N$

On veut représenter N points ,

dans un espace de petite dimension

de sorte que les distances euclidiennes de la représentation soient aussi proches que possible des indices $d_{ii'}, 1 \leq i, i' \leq N$

$$d_{ii'} \approx \delta_{ii'}$$

49

Critères pour le calcul. CRIT1

Trouver dans un espace de petite taille $z_i \in R^k$

Least squares ou Kruskall-Shepard scaling

STRESS function $S(z_1, \ldots z_n) = \left[\sum_{1 \leq i \neq j \leq n} (d_{ij} - \| z_i - z_j \|)^2 \right]^{1/2}$ Soit minimum

$ArgMin_{z_1, \ldots z_n} \; S(z_1, \ldots z_n)$

Par un algorithme de descente du gradient

50

MDS 2

Critère sur produits scalaires CRIT2

Nécessité des produits scalaires et pas seulement des distances

Classical scaling. Les données sont les points x_i

similarité $s_{ij} = < x_i - \bar{x}, x_j - \bar{x} >$

$$S(z_1,....z_n) = \left[\sum_{1 \le i \ne j \le n} (s_{ij} - < z_i - \bar{z}, z_j - \bar{z} >)^2 \right]^{1/2}$$

Solution: soit la diagonalisation $S_{n,n} = V' D V_{n,n}$, $z_1,....z_n$ sont les lignes de

$Z = V_k D_k^{1/2}$

$V_{n,k}$ $D_{n,k}$

51

isoMDS en R

- isoMDS(MASS) **Kruskal's Non-metric Multidimensional Scaling**
- One form of non-metric multidimensional scaling

- isoMDS(d, y = cmdscale(d, k), k = 2, maxit = 50, trace = TRUE, tol = 1e-3, p = 2)

Arguments

- d distance structure of the form returned by dist, or a full, symmetric matrix. Data are assumed to be dissimilarities or relative distances, but must be positive except for self-distance. Both missing and infinite values are allowed.
- y An initial configuration. If none is supplied, cmdscale is used to provide the classical solution, unless there are missing or infinite dissimilarities.
- k The desired dimension for the solution, passed to cmdscale.
- maxit The maximum number of iterations.
- trace Logical for tracing optimization. Default TRUE.
- tol Convergence tolerance.
- p Power for Minkowski distance in the configuration space.

52

MDS 3

Critères pour le calcul CRIT3

Sammon mapping (conserve les petites valeurs)

$$S(z_1,\ldots z_n) = \left[\sum_{1\le i\ne j\le n} \frac{(d_{ij} - \| z_i - z_j \|)^2}{d_{ij}}\right]^{1/2}$$

Favorise les petites distances ce qui permet la représentation des propriétés locales mieux que dans les méthodes concurrentes

Exemple de R (complété)

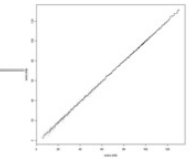
Swiss Fertility and Socioeconomic Indicators (1888)
Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

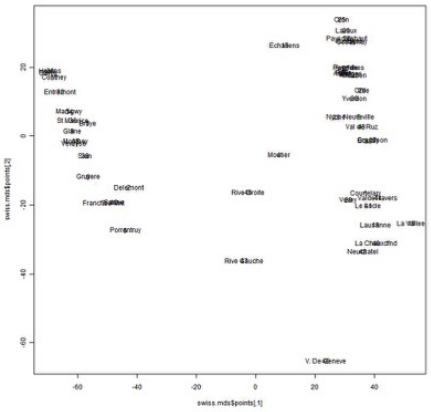
A data frame with 47 observations on 6 variables, each of which is in %

[,1] Fertility *Ig*,	'common standardized fertility measure'
[,2] Agriculture	% of males involved in agriculture as occupation
[,3] Examination	% draftees receiving highest mark on army examination
[,4] Education	% education beyond primary school for draftees.
[,5] Catholic	% 'catholic' (as opposed to 'protestant').
[,6] InfantMortality	live births who live less than 1 year.

53

Essai de isoMDS avec une distance euclidienne entre les points de l'espace initial





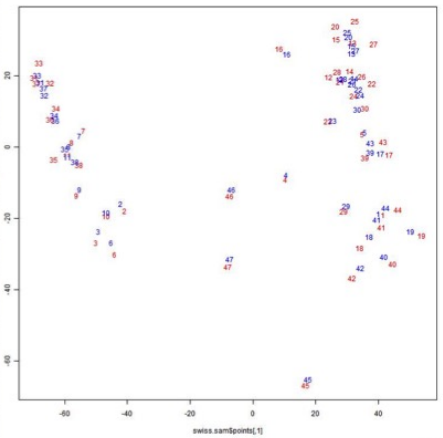
```
data(swiss)
wiss.x          <- as.matrix(swiss[, -1])
swiss.x         <- as.matrix(swiss)
swiss.dist      <- dist(swiss.x)
swiss.mds       <- isoMDS(swiss.dist,k=3,p=2)
plot(swiss.mds$points, type = "n");
text(swiss.mds$points, labels =
as.character(1:nrow(swiss.x)))
text(swiss.mds$points, labels = noms)

swiss.sh        <- Shepard(swiss.dist, swiss.mds$points)
plot(swiss.sh, pch = ".");
lines(swiss.sh$x, swiss.sh$yf, type = "S")

#---------------------- sammon
plot(swiss.sam$points, type = "n");
text(swiss.sam$points, labels =
as.character(1:nrow(swiss.x)))
text(swiss.$points, labels = noms)

#-----------------------------
# comparaison des deux analyses
plot(swiss.sam$points, type = "n");
text(swiss.sam$points, labels =
as.character(1:nrow(swiss.x)),col='red')
text(swiss.mds$points, labels =
as.character(1:nrow(swiss.x)),col='blue')
```

54

MDS 4

Sammon et MDS

1	Courtelary	40	-19	42	-19
2	Delemont	-42	-16	-41	-18
3	Franches-Mn	-49	-24	-50	-27
4	Moutier	11	-7.8	10	-9
5	Neuveville	36	4.2	35	3.7
6	Porrentruy	-45	-27	-44	-30
7	Broye	-55	3.2	-54	4.7
8	Glane	-59	0.18	-58	1.3
9	Gruyere	-55	-12	-56	-14
10	Sarine	-47	-18	-47	-19
11	Veveyse	-59	-2.8	-59	-2.3
12	Aigle	28	19	24	20
13	Aubonne	31	26	32	29
14	Avenches	32	19	31	21
15	Cossonay	31	29	26	30
16	Echallens	11	26	8.4	28
17	Grandson	41	-1.8	43	-2.1
18	Lausanne	37	-25	34	-28
19	La Vallee	50	-24	54	-25
20	Lavaux	30	31	26	34
21	Morges	32	19	28	18
22	Moudon	34	16	38	18
23	Nyone	25	7.5	24	7.2
24	Orbe	34	15	32	14
25	Oron	30	32	33	35
26	Payerne	32	18	35	20
27	Paysd'enhaut	33	27	39	29
28	Rolle	29	19	27	21
29	Vevey	30	-17	29	-18
30	Yverdon	33	11	36	11
31	Conthey	-68	18	-70	20
32	Entremont	-66	14	-65	18
33	Herens	-69	20	-68	24
34	Martigwy	-63	9	-63	11
35	Monthey	-60	-0.54	-63	-3.5
36	St Maurice	-63	7.4	-65	7.7
37	Sierre	-67	17	-69	18
38	Sion	-57	-4.1	-55	-4.8
39	Boudry	37	-1.4	36	-3
40	La Chauxdfnd	42	-31	45	-33
41	Le Locle	39	-20	41	-22
42	Neuchatel	34	-34	32	-37
43	Val de Ruz	37	1.2	41	1.5
44	ValdeTravers	42	-17	46	-17
45	V. De Geneve	18	-65	17	-67
46	Rive Droite	-6.8	-12	-7.5	-14



55

Critère CRIT4

NON METRIQUE Kruskal-Shepard scaling

travaille sur les rangs: calculés sur les rangs

d_{ij} calculés sur les rangs

soit une fonction $\theta()$ croissante

STRESS function $S(z_1,....z_n) = \dfrac{\sum_{1\le i\neq j\le n} (\theta(d_{ij}) - \| z_i - z_j \|)^2}{\sum_{1\le i\neq j\le n} d_{ij}^{\,2}}$

à θ fixé, un algorithme de descente du gradient

à z_i donnés, une regression monotone (isotonic) permet le calcul de θ croissant.

On intère jusqu'à stabilisation

56

MDS 5

Exemple

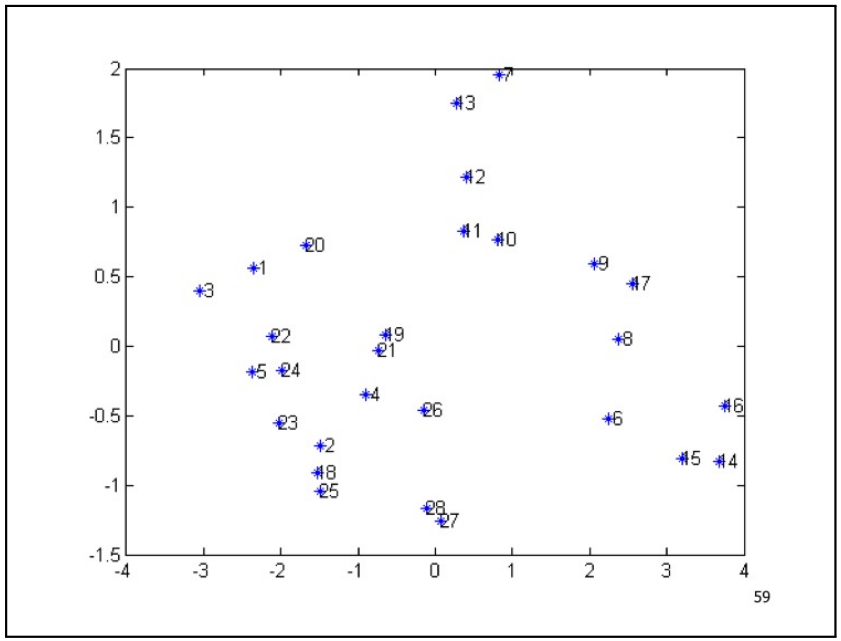
Voitures mds

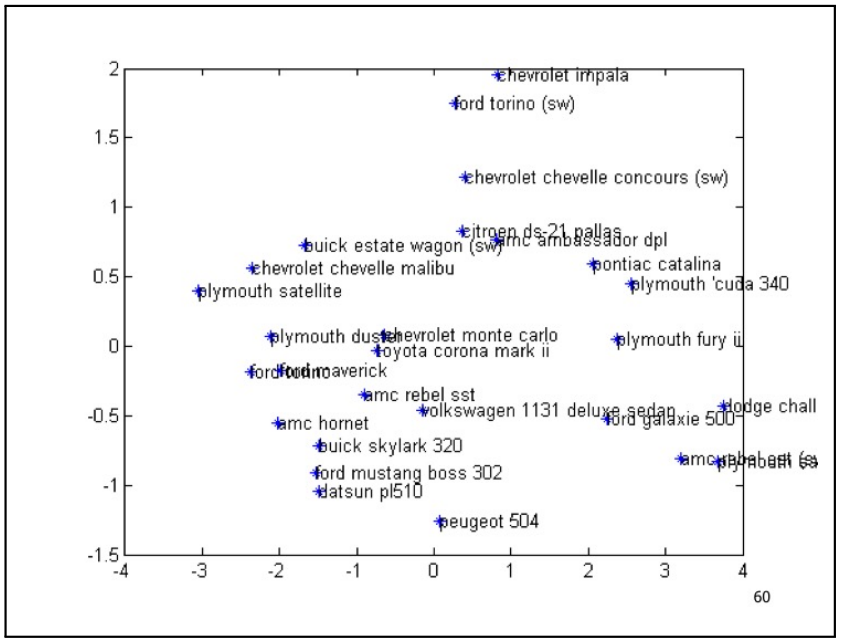
57

```
% Demos stats en MATLAB mdscale
load carbig
X = [MPG, Acceleration, Displacement, Weight, Horsepower];
models77 = find((Model_Year==77));


h = glyphplot(X(1:9,:),  'glyph', 'star', 'varLabels',varNames, 'obslabels',
    Model(1:9,:));


dissimilarity = pdist(zscore(X(models77,:)))
Y = mdscale(dissimilarity,2)
```

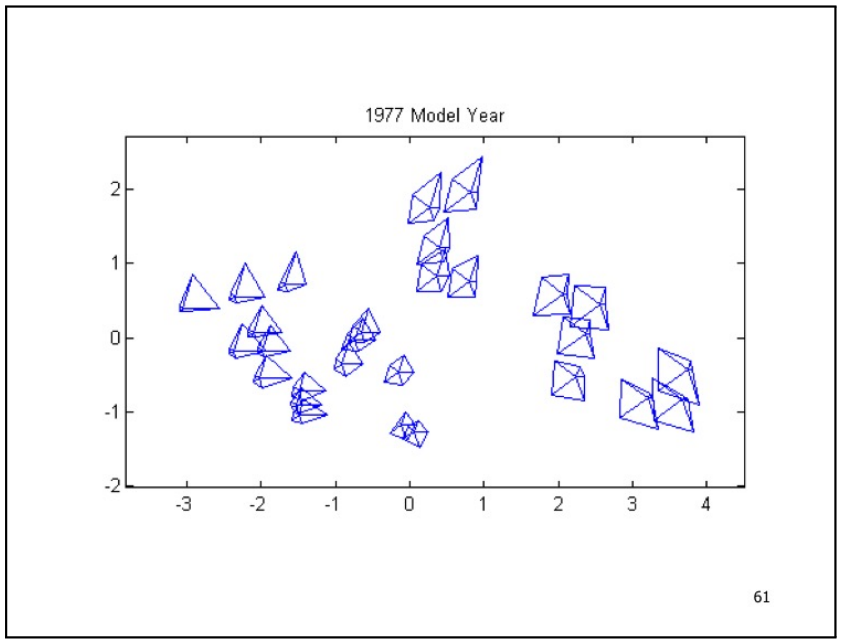
58

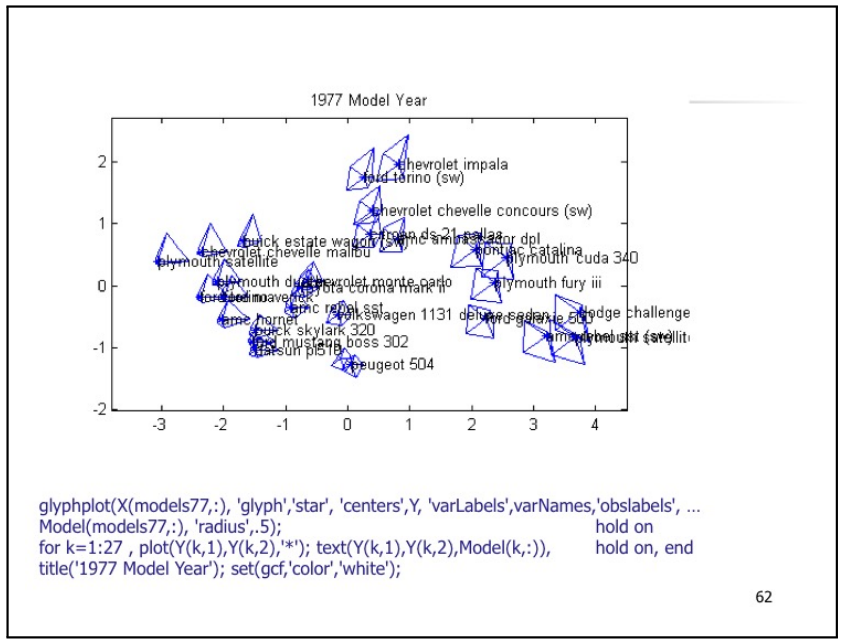
MDS 6


59


chevrolet impala
ford torino (sw)
chevrolet chevelle concours (sw)
citroen ds-21 pallas
buick estate wagon (sw)
amc ambassador dpl
chevrolet chevelle malibu
pontiac catalina
plymouth satellite
plymouth 'cuda 340
plymouth duster
chevrolet monte carlo
toyota corona mark ii
plymouth fury iii
ford maverick
amc rebel sst
volkswagen 1131 deluxe sedan
dodge chall
amc hornet
ford galaxie 500
buick skylark 320
ford mustang boss 302
datsun pl510
peugeot 504
60

MDS 7





ACC

Analyse en composantes curvilgnes

Plan

Similarité ou dissimilarité

Objets entre lesquels une similarité est calculée

- Vecteur, Matrice, Cube, Courbe, Fonction, Surface, Image, Espace vectoriel, Opérateurs
- Probabilité, Histogramme, Densité, Fonction de répartition, Spectre de signal
- Ensembles, Partitions, Graphes dont arbres
- Génomes

Exemple $X = [0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1]$,
$Y = [1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1]$
$(n_{10} + n_{01})/(n_{00} + n_{11})$
Discordances/ Concordances

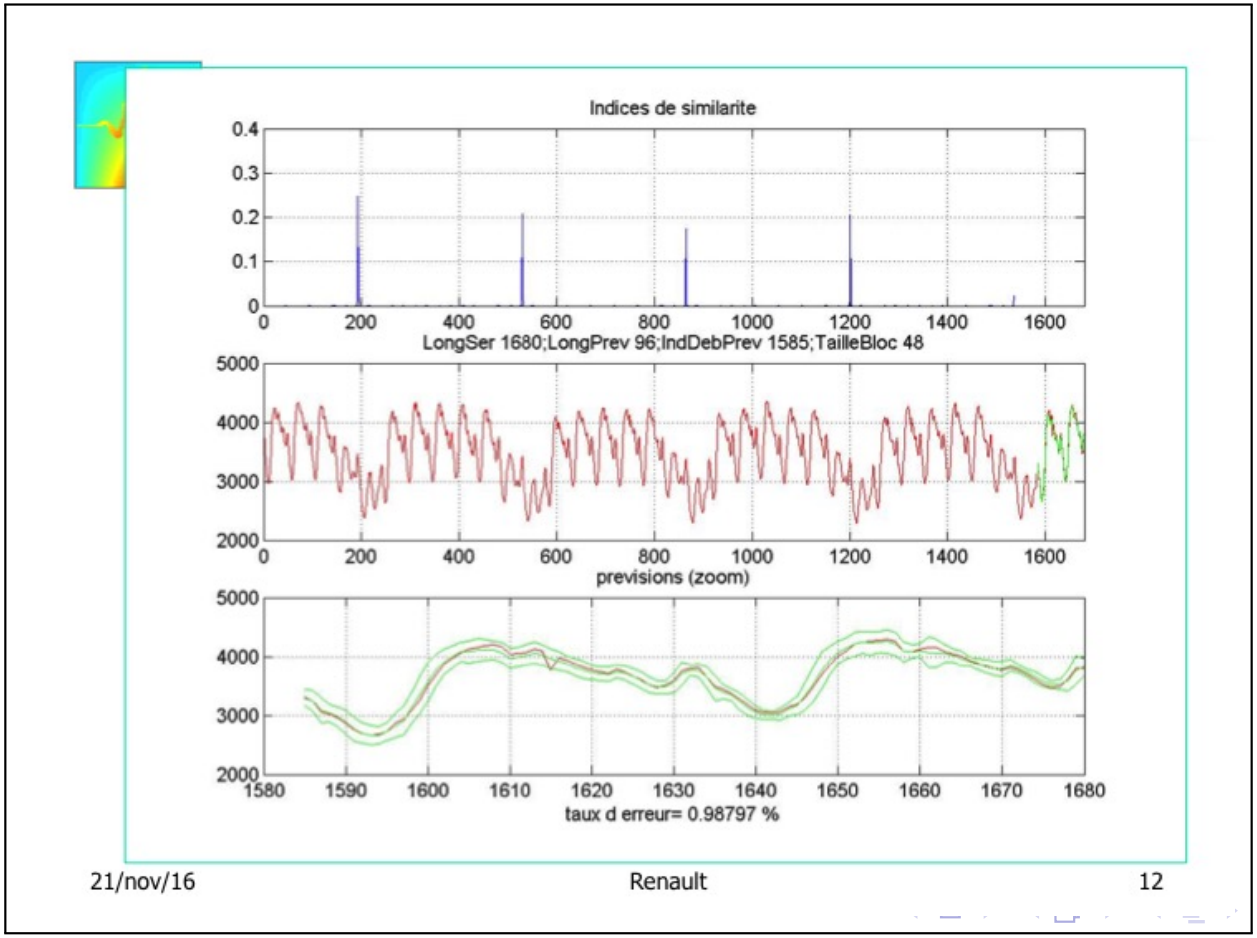
Semblable? quand?

Deux éléments sont semblables si l'indice de dissimilarité est petit
A partir d'un indice et d'un seuil

- De façon experte
- Si on dispose de la loi de variation: Test d'égalité à 0
- $U = r\sqrt{[n-2]/(1-r^2)}$, U distribué comme t à $n?2$ ddl
- $V = (1/2)log([1+r]/[1-r])$ distribué comme $N(0, 1/\sqrt{n-3})$

1. Prévoir par similarité les analogues. Un exemple
- Indices de similarité un gros catalogue
  - distances
- Similarité de partitions
- Entropie
- Multidimensional Scaling
- Apprentissage de variété: distances géodésiques. ISOMAP.
- corrélations 2 a 2

Prevision par similarité 1



Prevision par similarité 2

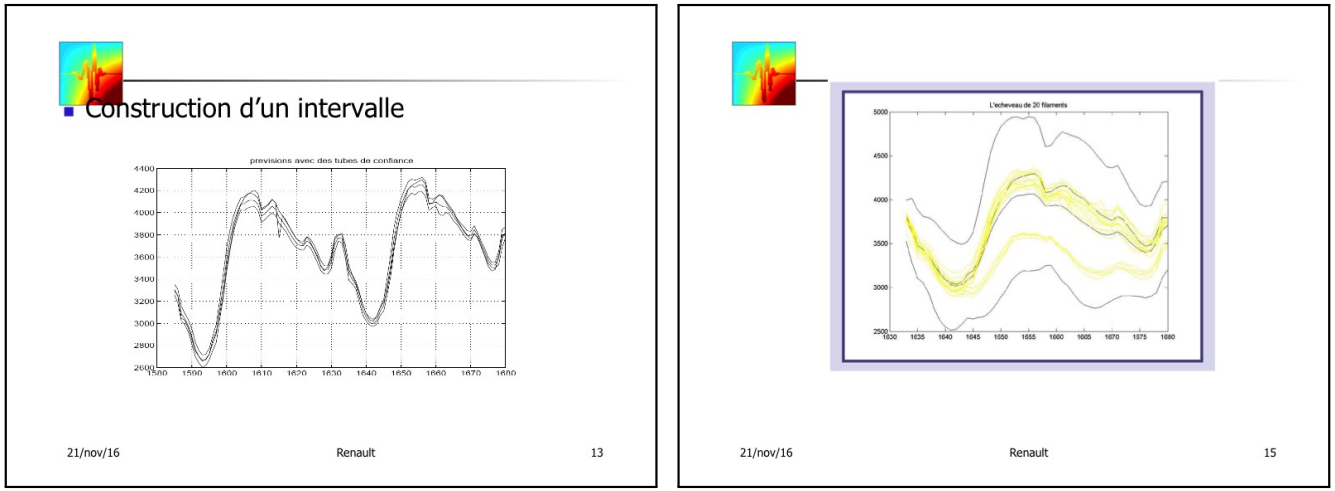
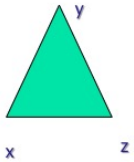


Figure: Que se passe t-il dans l'intervalle de confiance?

Distance

- Distances
  - d(x,y)=0 <=> x=y
  - Symétrique d(x,y)=d(y,x)
  - Transitif $d(x,z) \le d(x,y) + d(y,z)$

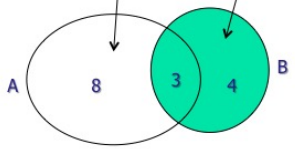


22

Similarité des graphes, une longue histoire

- Distances entre ensembles (puis relations)
  - Une vraie distance le nombre d'éléments de la différence symétrique

$$\Delta(A,B) = (A \cap \overline{B}) \cup (\overline{A} \cap B); d(A,B) = \#\Delta(A,B)$$



$d(A,B) = 12$

21/nov/16 Renault 23

Distance

Distance Euclidienne

- Il y a un produit scalaire <x,y>

$< x, y >$ forme bilinéaire positive

$\|x\|$ longueur $|x\| = \sqrt{<x,x>}$

d(x,y) distance $\|x-y\|$

Perpendiculaires

Notion d'angle

(donc corrélation)

24

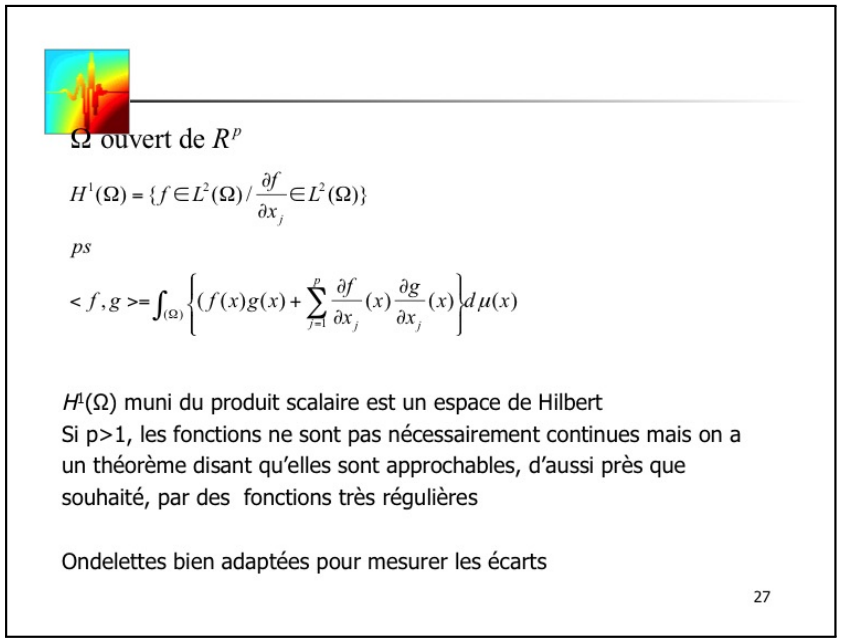
Perpendiculaire, Pythagore angle, cos corrélation

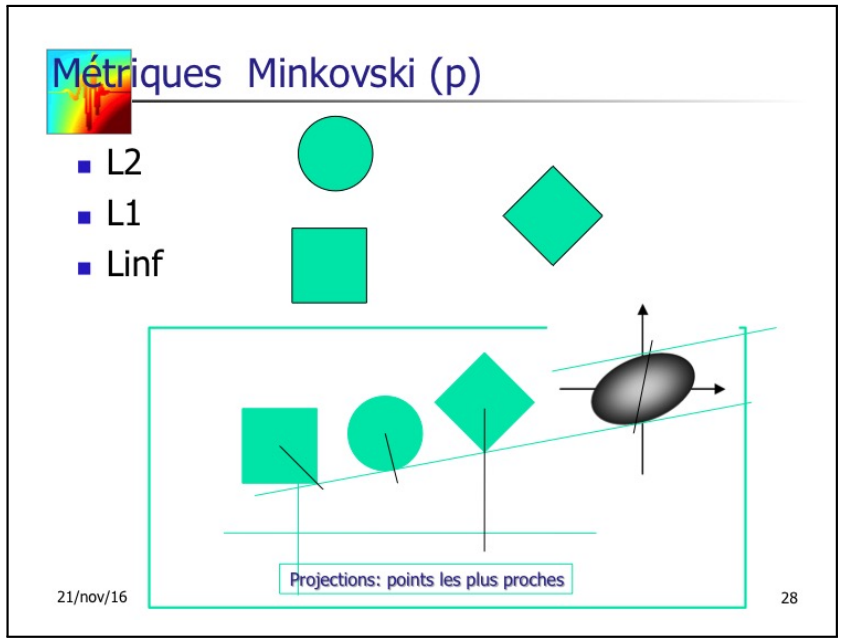
Espace euclidien	Espace normé	Espace métrique	Espace topologique
ps			
norme	norme		
distance	distance	distance	
topologie	topologie	topologie	topologie

- Utilité: mesurer, proche ou lointain
- Structurer l'espace: boules
- Convergence: proximité en mouvement d'une position

25

Distance


Ω ouvert de R^p
$H^1(\Omega) = \{f \in L^2(\Omega) / \frac{\partial f}{\partial x_j} \in L^2(\Omega)\}$
ps
$< f, g >= \int_{(\Omega)} \left\{ (f(x)g(x) + \sum_{j=1}^{p} \frac{\partial f}{\partial x_j}(x) \frac{\partial g}{\partial x_j}(x) \right\} d\mu(x)$
$H^1(\Omega)$ muni du produit scalaire est un espace de Hilbert
Si p>1, les fonctions ne sont pas nécessairement continues mais on a un théorème disant qu'elles sont approchables, d'aussi près que souhaité, par des fonctions très régulières
Ondelettes bien adaptées pour mesurer les écarts
27


Métriques Minkovski (p)
L2
L1
Linf
Projections: points les plus proches
21/nov/16
28

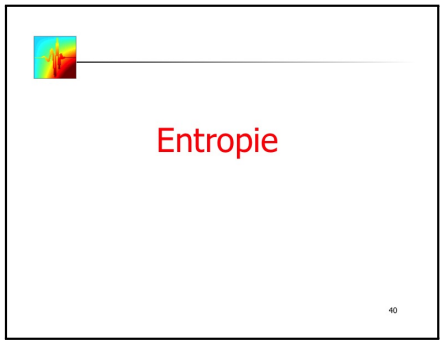
Variation totale

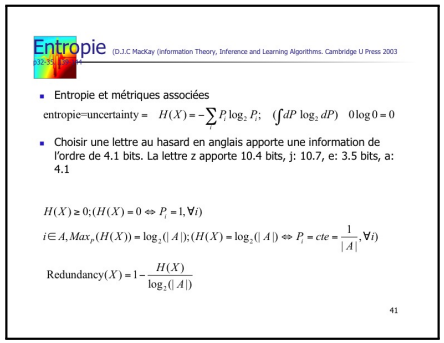
La notion (norme sur un bon espace) est utile dans l'analyse d'image et des fonctions

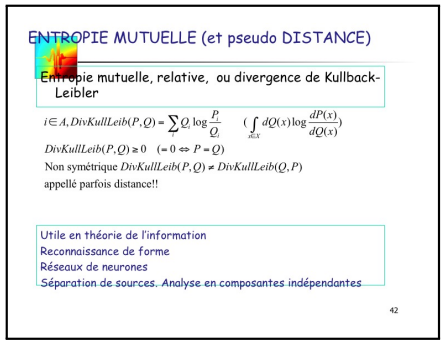
$$\begin{aligned} VT(P,Q) &= sup_{A \in evenement} |P(A) - Q(A)| \\ &= (1/2) \int |p(x) - q(x)| dx \\ &= (1/2) \sum_x |P(x) - Q(x)| dx \end{aligned}$$

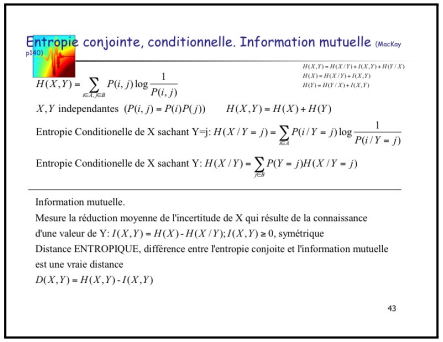
Apparait en applications dans des domaines aussi variés que le traitement d'image, le contrôle optimal, le calcul des variations, et plus largement l'analyse numérique

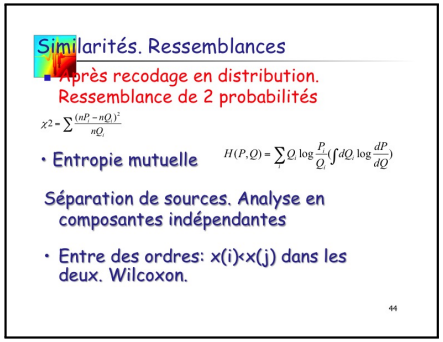
Entropie


Entropie
40


Entropie
(D.J.C MacKay (Information Theory, Inference and Learning Algorithms. Cambridge U Press 2003
Entropie et métriques associées
entropie=uncertainty = $H(X) = -\sum_i P_i \log_2 P_i$; $(\int dP \log_2 dP)$ $0 \log 0 = 0$
Choisir une lettre au hasard en anglais apporte une information de l'ordre de 4.1 bits. La lettre z apporte 10.4 bits, j: 10.7, e: 3.5 bits, a: 4.1
$H(X) \geq 0; (H(X) = 0 \Leftrightarrow P_i = 1, \forall i)$
$i \in A, Max_P(H(X)) = \log_2(|A|); (H(X) = \log_2(|A|) \Leftrightarrow P_i = cte = \frac{1}{|A|}, \forall i)$
$\text{Redundancy}(X) = 1 - \frac{H(X)}{\log_2(|A|)}$
41


ENTROPIE MUTUELLE (et pseudo DISTANCE)
Entropie mutuelle, relative, ou divergence de Kullback-Leibler
$i \in A, DivKullLeib(P,Q) = \sum_i Q_i \log \frac{P_i}{Q_i}$ $(\int_{x \in X} dQ(x) \log \frac{dP(x)}{dQ(x)})$
$DivKullLeib(P,Q) \geq 0$ $(= 0 \Leftrightarrow P = Q)$
Non symétrique $DivKullLeib(P,Q) \neq DivKullLeib(Q,P)$
appellé parfois distance!!
Utile en théorie de l'information
Reconnaissance de forme
Réseaux de neurones
Séparation de sources. Analyse en composantes indépendantes
42


Entropie conjointe, conditionnelle. Information mutuelle (MacKay
$H(X,Y) = \sum_{x \in X, y \in A} P(i,j) \log \frac{1}{P(i,j)}$
$H(X,Y) = H(X/Y) + I(X,Y) + H(Y/X)$
$H(X) = H(X/Y) + I(X,Y)$
$H(Y) = H(Y/X) + I(X,Y)$
X, Y independantes $(P(i,j) = P(i)P(j))$ $H(X,Y) = H(X) + H(Y)$
Entropie Conditionelle de X sachant Y=j: $H(X/Y=j) = \sum_{i \in A} P(i/Y=j) \log \frac{1}{P(i/Y=j)}$
Entropie Conditionelle de X sachant Y: $H(X/Y) = \sum_{j \in A} P(Y=j) H(X/Y=j)$
Information mutuelle.
Mesure la réduction moyenne de l'incertitude de X qui résulte de la connaissance d'une valeur de Y: $I(X,Y) = H(X) - H(X/Y); I(X,Y) \geq 0$, symétrique
Distance ENTROPIQUE, différence entre l'entropie conjoite et l'information mutuelle est une vraie distance
$D(X,Y) = H(X,Y) - I(X,Y)$
43


Similarités. Ressemblances
Après recodage en distribution.
Ressemblance de 2 probabilités
$\chi^2 = \sum \frac{(nP_i - nQ_i)^2}{nQ_i}$
Entropie mutuelle $H(P,Q) = \sum_i Q_i \log \frac{P_i}{Q_i} (\int dQ_i \log \frac{dP}{dQ})$
Séparation de sources. Analyse en composantes indépendantes
Entre des ordres: x(i)<x(j) dans les deux. Wilcoxon.
44

Plan

Les données sont des échantillons de réaction de cellules dont on veut étudier les évolutions dans le temps. Les cellules sont différentes aux différents instants et leur nombre n_t change avec le temps.

- Réfléchir à la façon de définir la qualité de résultats dans les situations non supervisées.
- Réfléchir à la façon d'expliquer les classes et donner les raisons pour lesquelles des individus se retrouvent dans une meme classe.
- Comparer des classifications.
- Evaluer l'importance es individus par suppression.
- Construire des classes de cellules.

Bibliography

- Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) *Classification and regression trees*. Editions Chapman & Hall/CRC, Monterey, CA.
- Chavent, M., Lechevallier, Y. and Briant, O. (2007) *DIVCLUS-T: a monothetic divisive hierarchical clustering method*. Computational statistics and data analysis. 52(2), 687–701.
- Corter, J. E. and Gluck, M. A. (1992). *Explaining basic categories: Feature predictability and information*. Psychological Bulletin, 111(2), 291-303.
- Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996) *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 226–231.
- Fisher, D. H. (1987) *Knowledge Acquisition Via Incremental Conceptual Clustering*. Machine Learning 2, 2:139-72. doi:10.1023/A:1022852608280.
- Fraiman, R., Ghattas, B. and Svarc, M. (2013) *Interpretable clustering using unsupervised binary trees*. Advances in data analysis and classification, **7**, 125–145.
- Ghattas, B., Svarc, M. and Fraiman, R. (2013) ***R**-package for interpretable clustering using binary trees*. http://lumimath.univ-mrs.fr/ghattas/CUBT.html.
- Gluck, M. A. and Corter, J. E. (1985). *Information, uncertainty and the utility of categories*. In Proceedings of the Seventh Annual Conference of the Cognitive Science Society. Hillsdale NJ: Lawrence Erlbaum.

Bibliography...

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I., H. (2009). *The WEKA Data Mining Software: An Update.* SIGKDD Explorations, Volume 11, Issue 1.
- Hennig C. (2014) *fpc: Flexible procedures for clustering.* R package version 2.1-9, http://CRAN.R-project.org/package=fpc.
- Huang, Z. (1998) *Extensions to the k-modes algorithm for clustering large data sets with categorical values.* Data Mining and Knowledge Discovery, **2(3)**, 283–304.
- Hubert, L. and Arabie, P. (1985) *Comparing partitions.* Journal of Classification, **2(1)**, 193–218.
- Kodinariya, T. M. and Makwana, P. R. (2013) *Partitioning Clustering algorithms for handling numerical and categorical data: a review*, arXiv, *arXiv:1311.7219v1*.
- Leisch, F. (2006) *A tool box for K-centroids cluster analysis.* Computational Statistics and Data Analysis, **51(2)**, 526–544.
- Liu, B., Xia, Y., and Yu., P. S. (2000) *Clustering Through Decision Tree Construction.* In Proceedings of the Ninth International Conference on Information and Knowledge Management, 20-29. CIKM Š00. New York, NY, USA: ACM. doi:10.1145/354756.354775.
- MacQueen, J. (1967) *Some methods for classification and analysis of multivariate observations.* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Editions L. M. Le Cam & J. Neyman, **1**, 281–297.
- Murtagh, F. (1985) *Multidimensional Clustering Algorithms.* COMPSTAT Lectures 4, Physica-Verlag, Vienna.
- Rokach, L. (2010) *A survey of Clustering Algorithms* In Maimon, O. and Rokach, L. (eds.). Data Mining and Knowledge Discovery Handbook, 269–298, Springer US.
- Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005) *klaR Analyzing German Business Cycles.* In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335–343, Springer-Verlag, Berlin.