

Classification

CEPE

Mars 2024

Plan

Introduction

Clustering

Partitionnement (Classification) avec k -means

Classification Ascendante Hiérarchique

DBSCAN : Méthode fondée sur la densité

Modèles de mélange

Dimension reduction

Analyse en composantes principales

Analyse des correspondances multiples

Objectif

Avant tout travail de modélisation, on se doit de décrire les données dont on dispose.

Malheureusement le data analyste se retrouve fréquemment face à des bases de données massives, tant en termes de nombre d'individus qu'en termes de nombre de variables.

Les techniques d'analyse de données (*à la française*) constituent une solution adéquate pour décrire des ensembles de grande dimension.

Le tableau de données

x_i^j désigne la valeur de la j -ème variable (parmi d) observée sur le i -ème individu (parmi n).

Individus	Variables				
	1	...	j	...	d
1	x_1^1	...	x_1^j	...	x_1^d
⋮	⋮		⋮		⋮
i	x_i^1	...	x_i^j	...	x_i^d
⋮	⋮		⋮		⋮
n	x_n^1	...	x_n^j	...	x_n^d

Individus et variables

On confond dans ce qui suit l'**individu i** avec le vecteur :

$$x_i = (x_i^1, \dots, x_i^d)^\top$$

et la **variable j** avec le vecteur :

$$x^j = (x_1^j, \dots, x_n^j)^\top.$$

On note **X** ce tableau de données.

Quelle dimension réduire ?

- ▶ Le nombre de variables : *dimension reduction*.
- ▶ Le nombre d'individus :
 - ▶ choisir un sous-ensemble séquentiel,
 - ▶ choisir un sous-ensemble aléatoire ,
 - ▶ utiliser le *binning* : discréétisation de l'espace pour travailler avec des données moyennées (problème pour contrôler le nombre de points si le design est non uniforme),
 - ▶ regrouper les individus en classes homogènes : *clustering*.

Le clustering

- ▶ En anglais *clustering*, en français *classification non supervisée* (en anglais, *classification* désigne la *classification supervisée*).
- ▶ Une définition : action de **répartir en classes**, en catégories, des choses, des objets, ayant des caractères communs afin notamment d'en faciliter l'étude.
- ▶ Quelques exemples :
 - ▶ Astronomie : classification d'étoiles.
 - ▶ Géographie : délimitation de zones homogènes.
 - ▶ Marketing : détermination de segments de marchés (groupes de consommateurs ayant les mêmes habitudes).
 - ▶ Réseaux sociaux : extraction de communautés.

Un nombre de partitions explosif

Le **nombre de Bell** p_n donne le **nombre de partitions possibles pour n individus** :

n	4	6	10
p_n	15	203	115 975

On constate là qu'il nous faudra disposer d'algorithmes de recherche de partitions **optimales**, il sera impossible de tester toutes les partitions possibles.

Différentes méthodes

- ▶ Les méthodes de partitionnement non hiérarchiques.
→ *K-means*
- ▶ Les méthodes de partitionnement hiérarchique : regroupent (méthode ascendante) ou divisent (méthode descendante) les individus, de manière séquentielle.
→ *CAH & CDH*
- ▶ Les méthodes basées sur la densité (des points).
→ *DBSCAN*
- ▶ Les méthodes probabilistes, basés sur des modèles de mélange de lois.
→ *EM, SEM, etc.*

Des modes communs

Quelle que soit la méthode, il est nécessaire de définir :

- ▶ Une mesure de dissimilarité (ou de similarité) entre individus.
- ▶ Une mesure de l'homogénéité des groupes et la différence entre les différents groupes.

En général, on centre (voire centre et réduit) les individus avant un clustering.

Partition

On dit que $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ est une **partition** de l'espace des individus \mathcal{O} si :

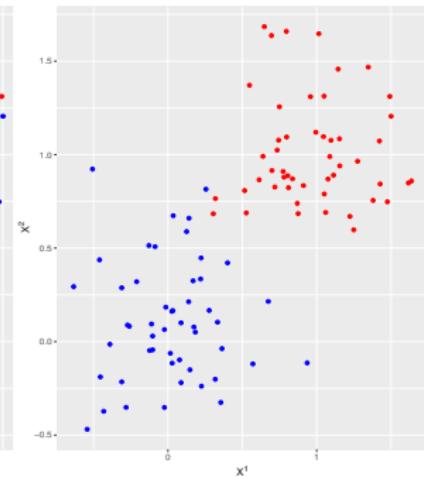
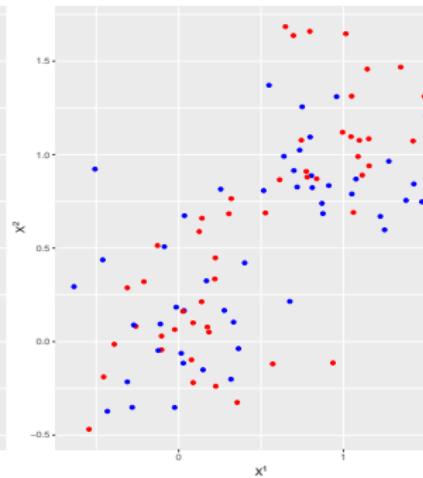
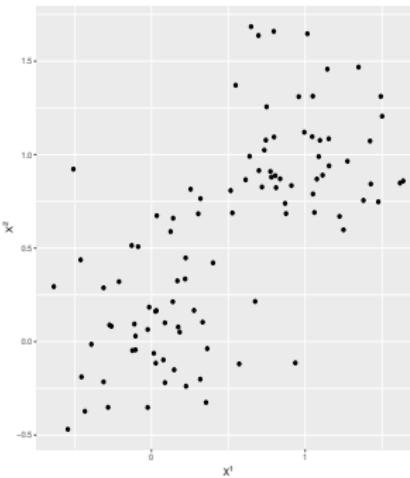
- ▶ $\forall k \in \{1, \dots, K\} : \mathcal{C}_k \neq \emptyset$,
- ▶ $\forall \{k, k'\} \in \{1, \dots, K\}^2 : \mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$,
- ▶ $\bigcup_{k \in \{1, \dots, K\}} \mathcal{C}_k = \mathcal{O}$.

Chaque élément \mathcal{C}_k de la partition est appelé **classe** ou **cluster**.

Caractérisation des classes

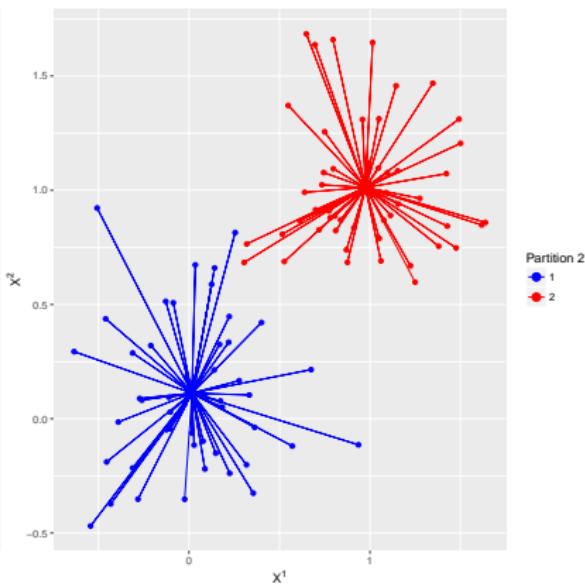
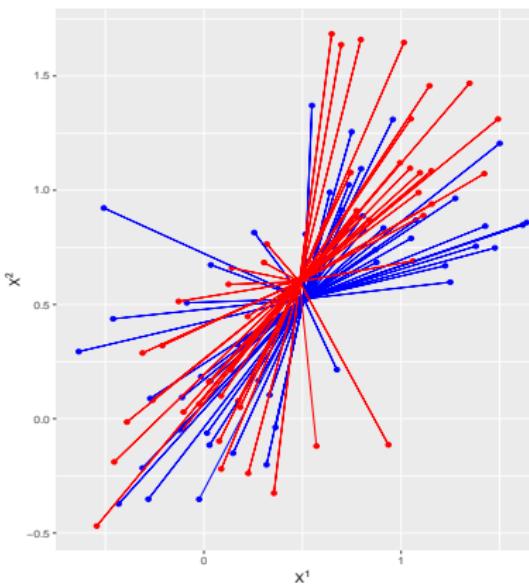
- ▶ On peut décrire chaque classe grâce aux variables actives (celles sur lesquelles on a souhaité différencier les classes), et grâce à toute autre variable supplémentaire.
- ▶ Lorsque les variables sont quantitatives, on peut comparer leurs moyennes sur les différentes classes.
Lorsque les variables sont qualitatives, on compare, pour chaque modalité, sa proportion dans la classe à sa proportion dans la population, afin de déterminer les modalités significativement sur-représentées (ou sous-représentées).
- ▶ On peut aussi rechercher l'individu le plus typique (ou central) de la classe, ou bien encore un noyau d'individus la représentant bien.
- ▶ Il est fréquent de nommer chacune des classes obtenues par un qualificatif résumant la caractérisation.

Qualité du clustering



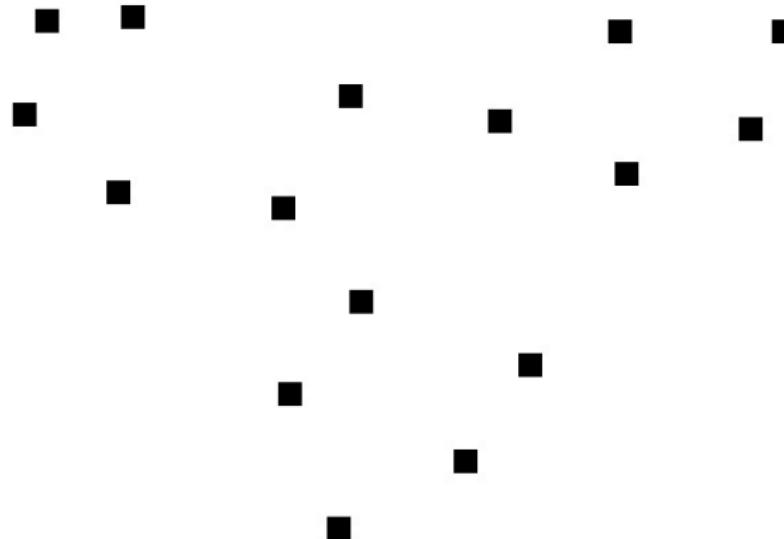
Quelle partition choisissez-vous ?

Qualité du clustering

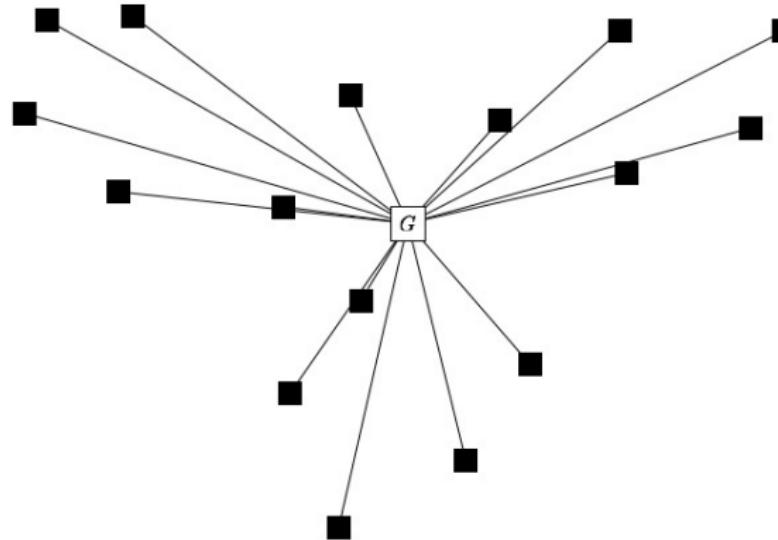


Inertie (notion de variabilité) intra classe et inter classes

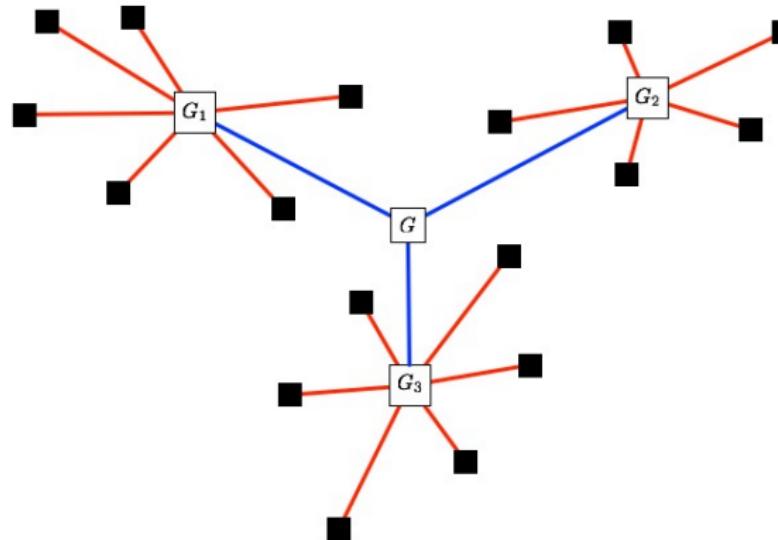
A partir de ce nuage de points :



L'inertie totale de ce nuage de points est symbolisée par les distances en noir entre les points et le centre de gravité G :



L'inertie intra-classes est symbolisée par les distances en rouge et l'inertie inter-classes par les traits en bleu (pour $K = 3$) :



Poids, centre de gravité, inertie

Chaque individu a en général un poids $\omega_i = 1/n$ sinon on note son poids ω_i avec $\sum_{i=1}^n \omega_i = 1$.

Le **barycentre** G d'un nuage de points est :

$$G = \sum_{i=1}^n \omega_i X_i .$$

On appelle **inertie totale** la quantité :

$$\mathcal{I}_{tot} = \sum_{i=1}^n \omega_i d^2 (X_i, G) .$$

où d désigne la distance euclidienne.

Décomposition de l'inertie totale

L'inertie intra-classes mesure la concentration dans les K classes :

$$\mathcal{I}_{intra} = \sum_{k=1}^K \sum_{i \in C_k} \omega_i d^2(X_i, G_k) .$$

L'inertie inter-classes mesure l'éloignement des K classes :

$$\mathcal{I}_{inter} = \sum_{k=1}^K \mu_k d_Q^2(G_k, G) .$$

où μ_k correspond au poids du groupe.

On a toujours (donc quand l'une augmente l'autre diminue)

$$\mathcal{I}_{tot} = \mathcal{I}_{intra} + \mathcal{I}_{inter} .$$

Inerties et clustering

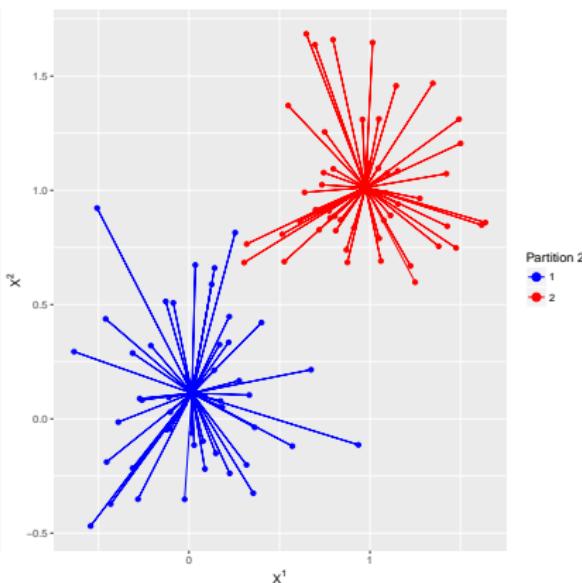
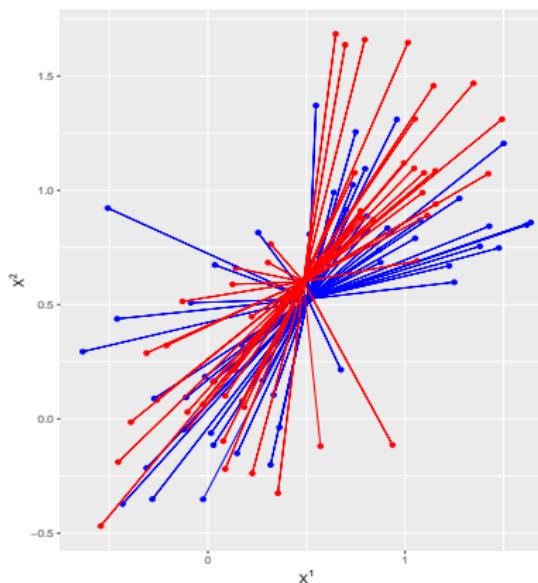
A K fixé, l'idéal est de **minimiser l'inertie intra-classes** (i.e rendre les classes les plus homogènes possible), soit encore **maximiser l'inertie inter-classes** (i.e séparer le plus possible les classes).

La **qualité d'un clustering** peut être évaluée par :

$$\frac{\mathcal{I}_{\text{inter}}}{\mathcal{I}_{\text{tot}}} ,$$

interprétable comme une part d'inertie des n individus expliquée par leur synthèse en K barycentres.

L'objectif de la méthode des k -means est de minimiser l'inertie intra classes



Inertie (notion de variabilité) intra classe et inter classes

Un critère à k fixé

$$g_n(\mathcal{C}) = \sum_{i \in \text{gp1}} \|x_i - \bar{x}_1\|^2 + \sum_{i \in \text{gp2}} \|x_i - \bar{x}_2\|^2$$

Critère des k -means

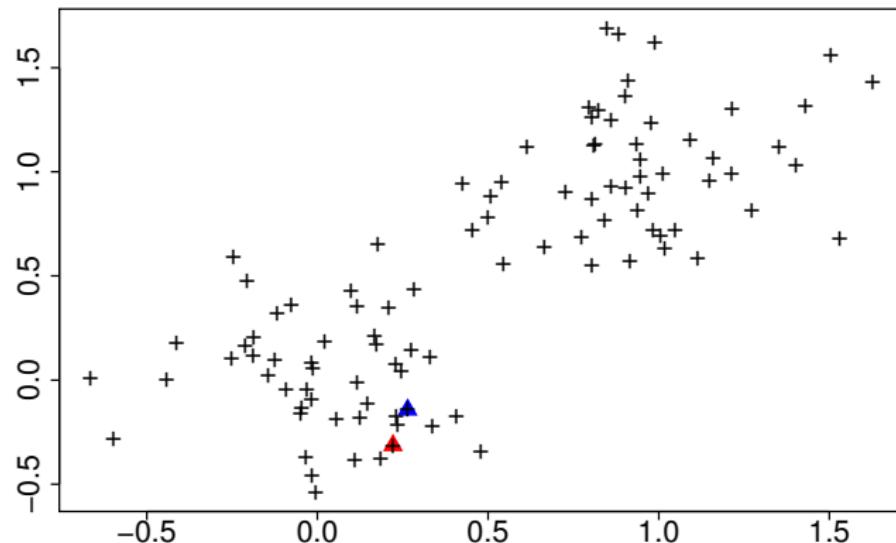
- ▶ Soient des observations $x_1, \dots, x_i, \dots, x_n$ $x_i \in \mathbb{R}^d$
→ p variables **quantitatives continues**
- ▶ Soit $\mathcal{C} = (C_1, C_2, \dots, C_K)$ une partition de $\{1, 2, \dots, n\}$

On recherche la partition qui réalise le minimum du critère

$$g_n(\mathcal{C}) = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \bar{x}_{C_k}\|^2 \quad (1)$$

1. A l'heure actuelle, on ne sait pas trouver le minimum global de ce critère, c'est-à-dire la meilleure partition $\hat{\mathcal{C}}$ qui donne le critère le plus bas; méthodes itératives convergeant vers un minimum local.
2. Ce critère admet d'autres formulations équivalentes.

Illustration graphique

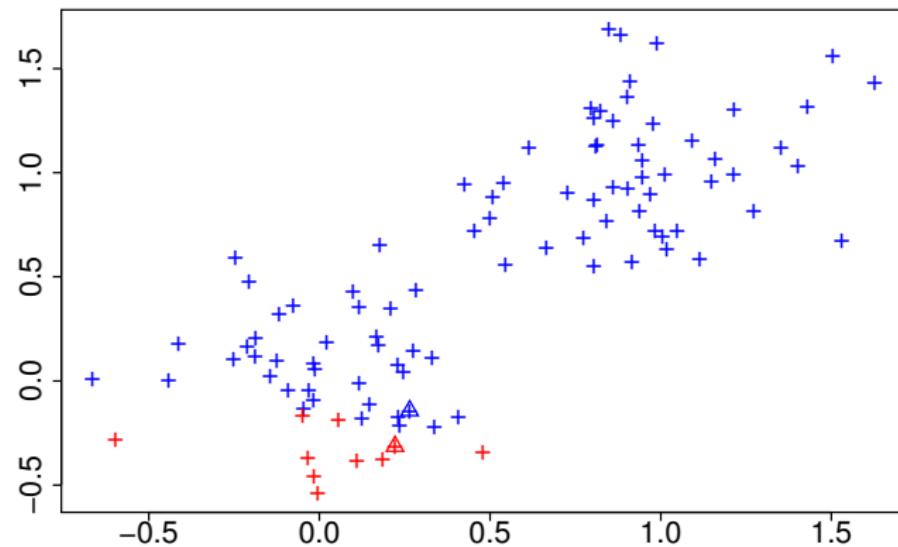


On choisit des centres initiaux

Classification

└ Partitionnement (Classification) avec k -means

└ Un algorithme simple des k -means

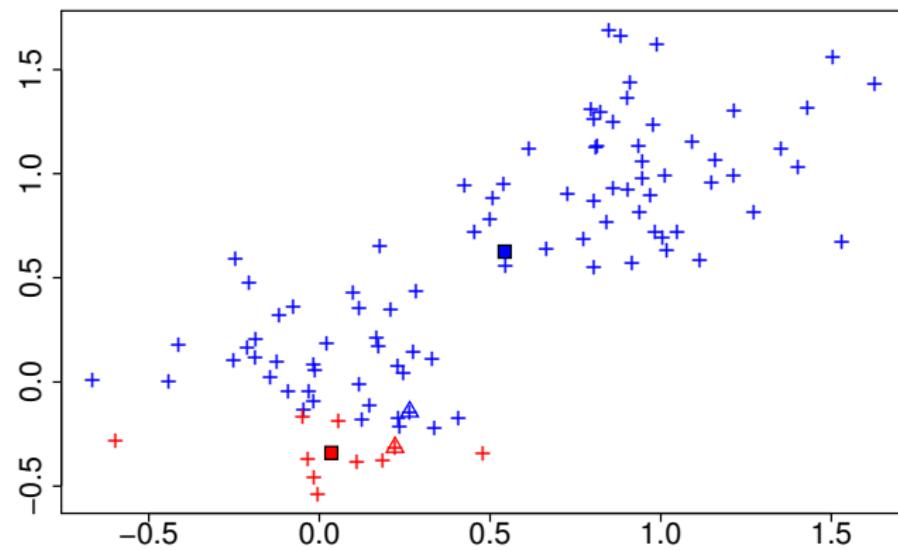


On réaffecte.

Classification

Partitionnement (Classification) avec k -means

Un algorithme simple des k -means

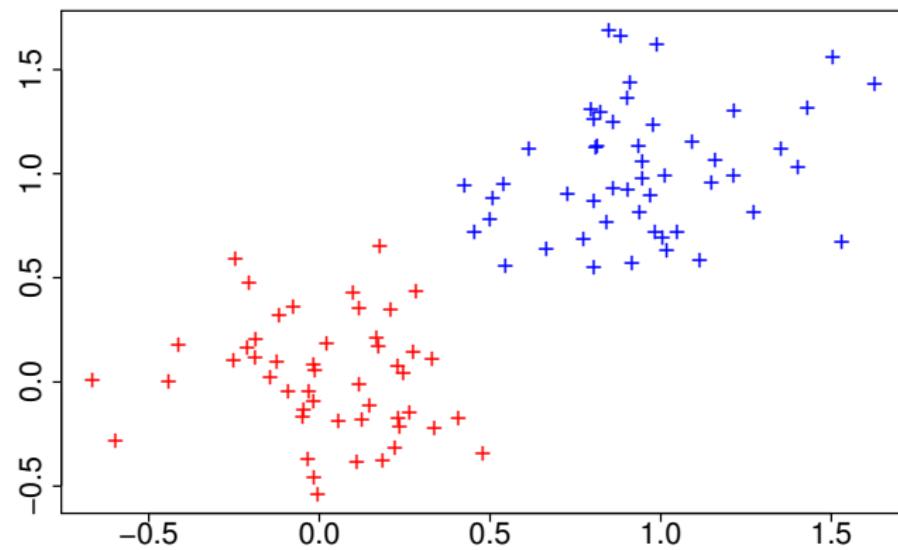


On recalcule les centres...

Classification

└ Partitionnement (Classification) avec k -means

└ Un algorithme simple des k -means



FIN

Formulation générale du critère des *k*-means

Représentant d'une classe

Nous représentons chaque classe k par un point de \mathbb{R}^d (pas forcément la moyenne) : notons $\mathcal{Z} = \{z_1, \dots, z_K\}$, $z_k \in \mathbb{R}^d$ ces représentants.

Méthode des *k*-means

Recherche de la meilleure partition ET des meilleurs représentants avec le critère suivant :

$$g_n(\mathcal{C}, \mathcal{Z}) = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - a_k\|^2 \quad (2)$$

- Quand on fixe une partition \mathcal{C}^* , les meilleurs représentants sont les moyennes $\widehat{\mathcal{Z}} = (\bar{x}_{C_1}, \dots, \bar{x}_{C_K})$

$$g_n(\mathcal{C}^*, \mathcal{Z}) \geq g_n(\mathcal{C}^*, \widehat{\mathcal{Z}}) = g_n(\mathcal{C}^*)$$

- Quand on fixe des représentants \mathcal{Z}^* , la meilleure partition est celle de la distance (carrée) minimale définie par

$$\widehat{\mathcal{C}} = \{\widehat{C}_1, \widehat{C}_2, \dots, \widehat{C}_K\}$$

$$\widehat{C}_k = \{i \in \{1, \dots, n\} \mid \|x_i - z_k\|^2 = \min_j \|x_i - z_j\|^2\}.$$

Elle réalise le minimum à représentants fixés :

$$g_n(\mathcal{C}, \mathcal{Z}^*) \geq g_n(\widehat{\mathcal{C}}, \mathcal{Z}^*)$$

Choix pour appliquer la méthode des k -means

- ▶ Choix du nombre de groupes K
- ▶ La distance entre vecteurs est la distance euclidienne
- ▶ Le représentant de chaque groupe C_k est la moyenne du groupe \bar{x}_{C_k}
- ▶ Choix du point de départ

Nombre de groupes

- ▶ en fonction d'une connaissance à priori
- ▶ à la suite d'une CAH
- ▶ critère ad-hoc : "coude" dans la représentation graphique de l'inertie intra-classes

$$\sum_{k=1}^K \sum_{i \in \hat{C}_k} \|x_i - \bar{x}_{\hat{C}_k}\|^2$$

en fonction du nombre de classes

Extension du Critère des k -means

Recherche de

- ▶ la partition optimale $\hat{\mathcal{C}}$
- ▶ des meilleurs représentants $\hat{\mathcal{Z}}$

qui réalisent le minimum de :

$$h_n(\mathcal{C}, \mathcal{Z}) = \sum_{k=1}^K \sum_{i \in C_k} d(x_i, z_k) \quad (3)$$

Distance

- ▶ distances (carrées ?) classiques (l_2 , l_1 , ...)
- ▶ distances issues de produit scalaire via un noyau
- ▶ distances ad hoc

Représentant z_k

- ▶ moyenne = k -means
- ▶ l'observation du groupe le plus central au sens de la distance choisie : k -medoids (plus robuste)

Point de départ

- ▶ K individus au hasard
- ▶ K individus choisis
- ▶ Choix des individus de départ après CAH (individu moyen par classe, ou un individu au hasard par classe)

Qualité d'une partition

Quand une partition est-elle bonne ?

- ▶ si des individus d'une **même classe** sont **proches** ;
- ▶ si des individus de **2 classes différentes** sont **éloignés**.

Mathématiquement cela se traduit par

- ▶ la variabilité (ou l'inertie) **intra-classe**

$$\mathcal{I}_{\text{intra}} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{C}_k} d^2(x_i, \bar{x}_{\mathcal{C}_k})$$

est **petite**.

- ▶ la variabilité **inter-classe**

$$\mathcal{I}_{\text{inter}} = \frac{1}{n} \sum_{k=1}^K n_k d^2(\bar{x}_{\mathcal{C}_k}, \bar{x}).$$

est **grande**.

Qualité d'une partition

Compromis entre ces 2 variabilités et la qualité d'une classification est mesurée par

$$0 \leq \frac{\mathcal{I}_{\text{inter}}}{\mathcal{I}_{\text{totale}}} \leq 1$$

où $\mathcal{I}_{\text{totale}} = \mathcal{I}_{\text{inter}} + \mathcal{I}_{\text{intra}} = \frac{1}{n} \sum_{i=1}^n d^2(x_i, \bar{x})$.

- ▶ 0 si on a une classe unique;
- ▶ 1 lorsque chaque objet est une classe.

On cherche à ce que ce critère soit proche de 1 sans avoir trop de groupes.

Algorithme de Lloyd ou Forgy

A partir des K centres fournis,

1. Affectation de tous les individus au centre le plus proche

$$\forall i \in \{1, \dots, n\}, \quad : K(i) = \operatorname{argmin}_{c_K \in C} d(x_i, c_K)$$

2. Calcul des nouveaux centres par la moyenne
3. Retour étape 1 tant qu'il y a changement

L'intertie intra-classes diminue à chaque étape.

Variante de Mac Queen

Accélère la convergence mais le résultat dépend de l'ordre des individus.

A partir des K centres fournis

1. Initialisation

- ▶ affectation de tous les individus au centre le plus proche

$$\forall i \in \{1, \dots, n\}, \quad : K(i) = \operatorname{argmin}_{c_K \in C} d(x_i, c_K)$$

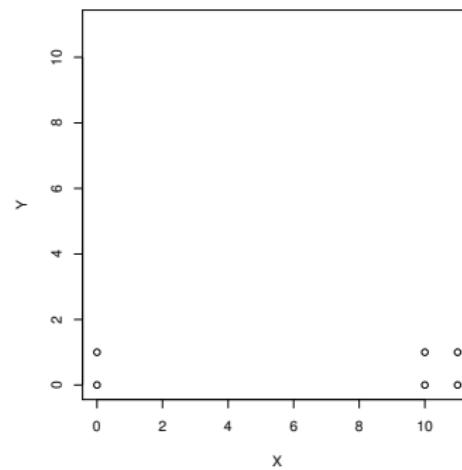
- ▶ calcul des centres par la moyenne

2. Faire pour tout $1 \leq i \leq n$:

- ▶ recherche du centre le plus proche de x_i ;
- ▶ calcul des nouveaux centres par la moyenne si x_i change de groupe (réactualisation d'au plus 2 centres)

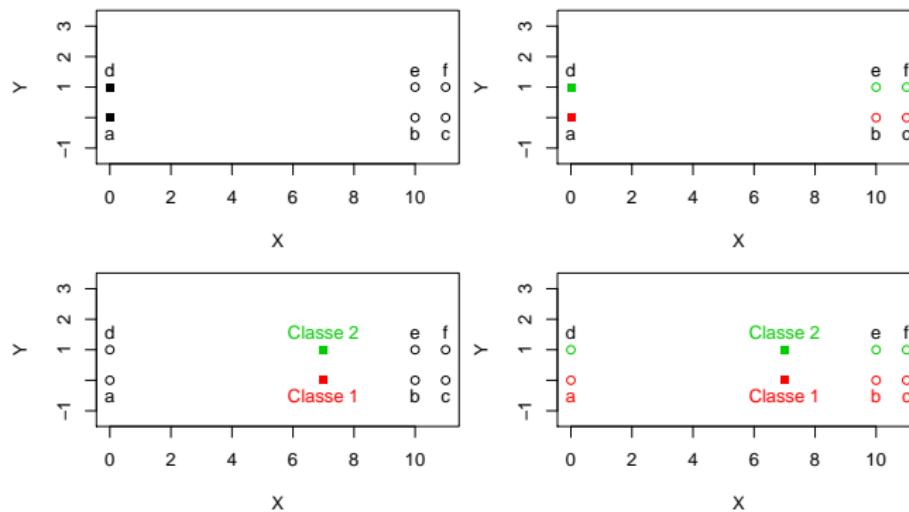
3. Retour étape 2 tant qu'il y a changement

Minima locaux



Imaginons cet exemple avec 2 groupes et les points initiaux sont les 2 points à gauche du graphique on obtient alors

Exemple (contre-exemple)



Les carrés = centre de classe, les ronds = les objets.

Hartigan & Wong

- ▶ Affectation de tous les individus au centre le plus proche

$$\forall i \in \{1, \dots, n\}, \quad : K(i) = \operatorname{argmin}_{c_K \in C} d(x_i, c_K)$$

et mise en mémoire du deuxième centre le plus proche pour chaque individu

- ▶ Mise à jour des centres (calcul de la moyenne)
- ▶ Tous les groupes sont actifs
- ▶ Alternance des étapes de transfert optimal et de transfert rapide pour savoir si un échange de groupe permet de diminuer la valeur du critère.

Transferts

- ▶ Pour chaque point $1 \leq i \leq n$ faire une comparaison de son groupe actuel avec d'autres (chaque changement de groupe d'un individu permettra aux 2 groupes d'être actifs.)
 - ▶ Si le groupe actuel de i est actif, faire la comparaison avec tous les autres groupes
changer l'observation i de groupe et regarder si le critère est diminué par cet échange. Si oui changement de groupe et recalcul des moyennes des 2 groupes. Mise à jour éventuelle des 2 groupes les plus proches.
 - ▶ Si le groupe actuel de i n'est pas actif, alors regarder seulement avec les groupes actifs et mise à jour comme ci-dessus.
- ▶ si plus de groupe actif alors arrêt
- ▶ Pour chaque point $1 \leq i \leq n$ tentative d'échange entre le groupe actuel et le groupe de le plus proche. Chaque changement permet aux 2 groupes d'être actifs. On itère ces changements rapides tant qu'on peut.

Exemple

Configuration	Echange testé	Inertie intra-classe	Echange ?
C1	C2		
{a,b,c}	{d,e,f}		148
{a,b,c}	{d,e,f}	a en C2	112
{b,c}	{a,d,e,f}	b en C2	130
{b,c}	{a,d,e,f}	c en C2	138.4
{b,c}	{a,d,e,f}	d en C1	149.3
{b,c}	{a,d,e,f}	e en C1	82.7
{b,c,e}	{a,d,f}	f en C1	2.5
{b,c,e,f}	{a,d}		2.5

Le coin R

```
> D
  X1 X2
a  0  0
b 10  0
c 11  0
d  0  1
e 10  1
f 11  1
> is.data.frame(D)
[1] TRUE
```

Le coin R

```
> a1 <- kmeans(D, centers=D[c(1,4),])
> a2 <- kmeans(D, centers=D[c(1,4),], algorithm="Lloyd")
> a3 <- kmeans(D, centers=D[c(1,4),], algorithm="MacQueen")
> a1$cluster
[1] 2 1 1 2 1 1
> a1$tot.withinss
[1] 2.5
> a2$cluster
[1] 1 1 1 2 2 2
> a2$tot.withinss
[1] 148
> a3$cluster
[1] 1 1 1 2 2 2
> a3$tot.withinss
[1] 148
> a2bis <- kmeans(D, centers=2, nstart=20, algorithm="Lloyd")
> a2bis$cluster
[1] 2 1 1 2 1 1
```

Minimum local

- ▶ Il faut partir de **plusieurs points de départ** ou utiliser l'algorithme k -means++ qui choisit au départ les points les plus éloignés possibles.
- ▶ Ce premier choix est coûteux mais le **nombre d'itérations** pour la convergence de l'algorithme est **plus petit**.

Données volumineuses

1. choisir un échantillon au hasard
2. appliquer l'algorithme des K-means
 - ▶ considérer les centres de gravité
 - ▶ mesurer la qualité de la partition obtenue avec toutes les données et les centres de gravité obtenus sur l'échantillon.
3. choisir un nombre prédéterminé d'échantillons et répéter
4. comparer les partitions obtenues et conserver la meilleure

Conclusion k -means

Assez rapide $O(I \times kdn)$ avec I, k et $d \ll n$ donc $O(n)$.

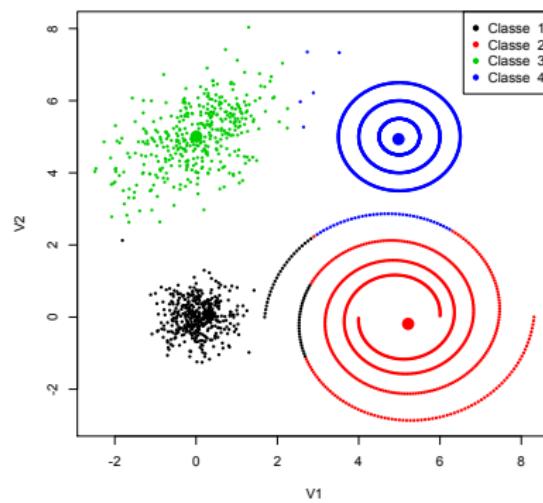
Problèmes :

- ▶ minimum local
- ▶ sensible aux conditions initiales (prendre plusieurs départs nstart)
- ▶ les classes sont constituées par rapport à des centres qui ne sont pas des éléments de la classe (moyenne)
- ▶ sensibles aux valeurs extrêmes

Le coin R

```
> don <- read.table("donclassif.txt",sep=";",  
+ header=TRUE)  
> km <- kmeans(don,centers=4)  
> names(km)  
> plot(don,col=km$cluster)  
> points(km$centers,cex=3,pch=16,col=1:4)  
> legend("topright",legend=paste("Classe ",1:4),  
+ col=1:4,pch=16)
```

Le coin R



Le coin R

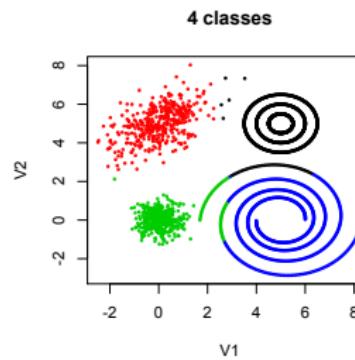
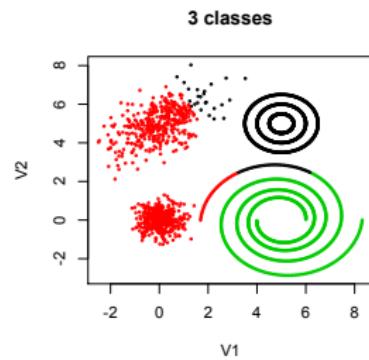
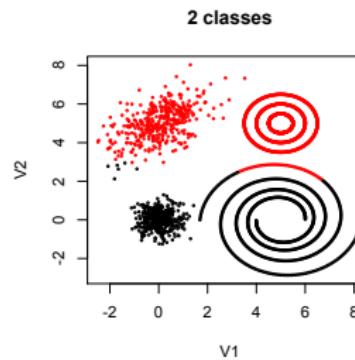
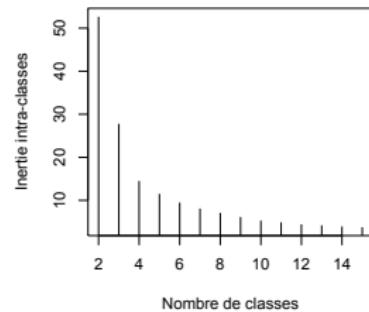
```
> k <- 2:15
> part <- k
> for(i in k){
>   kmk=kmeans(don,centers=i,nstart=20)
>   part[i-1]=sum(kmk$withinss)/kmk$totss*100
> }
> par(mfrow=c(2,2))
> plot(k,part,type="h", xlab="Nombre de classes",
+ ylab="Inertie intra-classes")
```

Classification

Partitionnement (Classification) avec k -means

Algorithmes des k -means

Le coin R



Algorithme PAM (Partitioning Around Medoid)

Le représentant de la classe n'est plus la moyenne mais le medoid.
Moins sensible aux outliers (plus robuste) que k -means.

1. phase d'initialisation :

- ▶ K représentants de classes choisis au hasard
- ▶ tous les éléments sont affectés à la classe dont le représentant est le plus proche
- ▶ la qualité de la partition est évaluée

2. un élément considéré comme représentant de classe est échangé avec un élément qui n'est pas représentant de classe
3. tous les éléments sont affectés à la classe dont le représentant est le plus proche. La qualité de la nouvelle partition est évaluée. L'échange est conservé si la qualité de la partition est améliorée
4. itération jusqu'à ce qu'à ne plus trouver d'échange

Algorithme CLARA (Clustering large applications)

Pour réduire les temps de calcul

1. construire plusieurs échantillons de données
2. pour chaque échantillon
 - ▶ trouver les représentants de classes par l'algorithme standard (ici PAM)
 - ▶ évaluer la qualité de la partition à partir des représentants de l'échantillon et de tous les éléments
3. l'ensemble des représentants qui fournit la solution optimale est considéré pour définir la partition

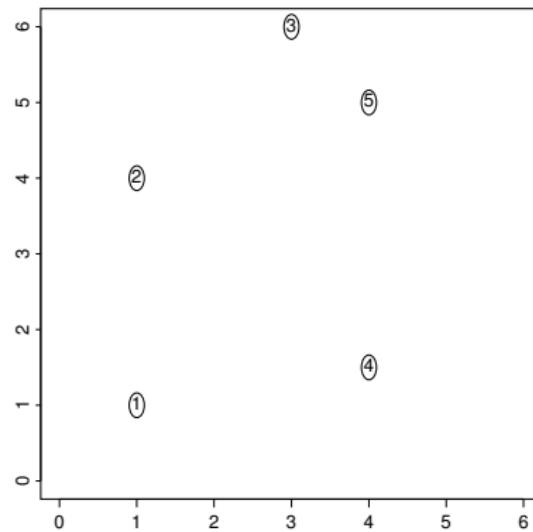
Fonctions **pam** et **clara** du package **cluster**.

Exemple introductif

Considérons l'exemple suivant

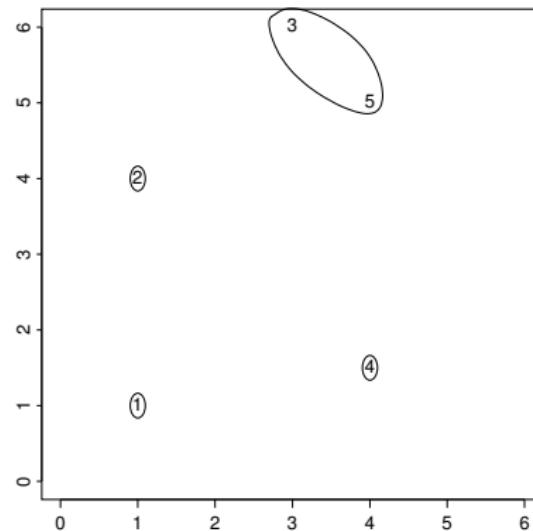
ind	X_1	X_2
1	1	1
2	1	4
3	3	6
4	4	1.5
5	4	5

Exemple suite : distance euclidienne



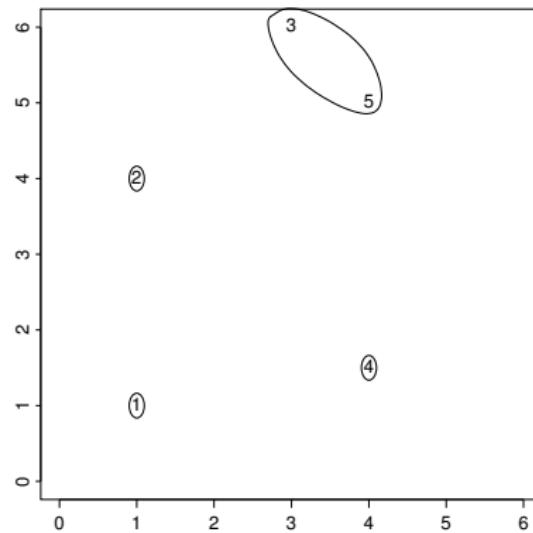
Exemple suite : distance euclidienne

On agrège les deux plus proches



Exemple suite : distance euclidienne

On agrège les deux plus proches

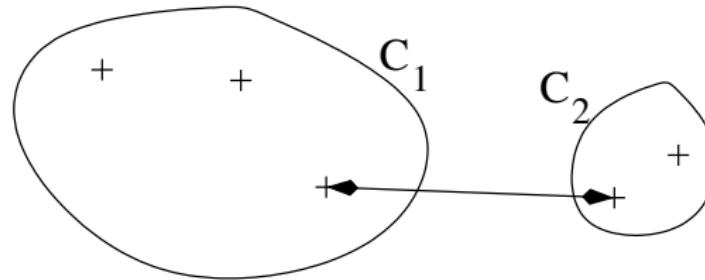


- Définir une dissimilarité ou dissemblance entre ensembles : indice d'agrégation

Saut minimum (minimum linkage ou single linkage)

Le saut minimum associé à une dissimilarité \bar{s} est la dissimilarité minimum que l'on peut trouver entre 2 éléments des deux groupes :

$$\Delta(A, B) = \min_{o_i \in A; o_j \in B} \bar{s}(o_i, o_j).$$

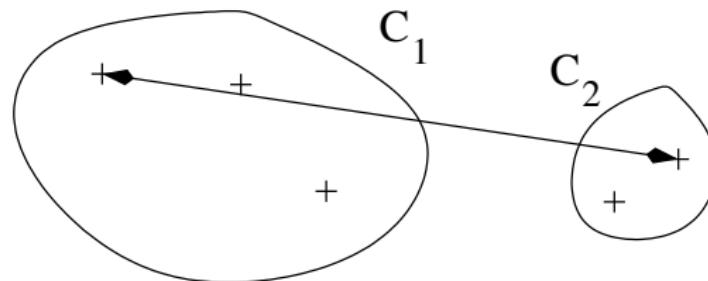


Groupes sont en général allongés, on va ainsi de proche en proche selon la philosophie "le voisin de mon voisin est mon voisin".

Saut maximum (complete linkage)

Le saut maximum est la dissimilarité maximum que l'on peut trouver entre 2 éléments des deux groupes (revient à calculer le diamètre de $A \cup B$ si l'on travaille avec une distance) :

$$\Delta(A, B) = \max_{o_i \in A; o_j \in B} \bar{s}(o_i, o_j).$$



Groupes en général assez compacts. Cependant cela donne souvent de nombreux petits groupes similaires.

Moyenne du groupe (average)

Ici l'indice d'agrégation est la moyenne de toutes les dissimilarités possibles entre 2 objets des 2 groupes :

$$\Delta(A, B) = \text{Moyenne}_{o_i \in A; o_j \in B} \bar{s}(o_i, o_j).$$

Indice de Ward ou d'accroissement de l'inertie

Si l'on travaille avec une distance :

$$\Delta(A, B) = \sqrt{\frac{|A||B|}{|A| + |B|} d(\bar{x}_A, \bar{x}_B)}$$

On regroupe les classes de poids/effectif « faible » et dont les centres de gravités sont proches. A chaque étape on augmente l'inertie intra-classes de façon minimale (on diminue l'inertie inter-classes de façon minimale).

Inertie

l'inertie d'une classe A (par rapport à son centre de gravité $\bar{x}(A)$) est

$$I(A) = \frac{1}{n} \sum_{i \in A} d^2(x_i, \bar{x}_A).$$

L'indice d'agrégation est donc

$$\Delta(A, B)^2 = I(A \cup B) - I(A) - I(B)$$

Inertie

Pour une partition $\mathcal{C}_1, \dots, \mathcal{C}_K$, si tous les points possèdent le même poids $1/n$, et \bar{x}_k désigne le centre de gravité de C_k ($k \in \{1, \dots, K\}$),

- ▶ l'inertie (ou variabilité) **intra-classe** est

$$\mathcal{I}_{\text{intra}} = \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{C}_k} d^2(x_i, \bar{x}_k)$$

- ▶ l'inertie **inter-classe** est

$$\mathcal{I}_{\text{inter}} = \frac{1}{n} \sum_{k=1}^K n_k d^2(\bar{x}_k, \bar{x}).$$

- ▶ l'inertie totale ne dépend pas de la partition et est :

$$\mathcal{I}_{\text{totale}} = \mathcal{I}_{\text{inter}} + \mathcal{I}_{\text{intra}} = \frac{1}{n} \sum_{i=1}^n d^2(x_i, \bar{x}).$$

Classification Ascendante Hiérarchique

1. n objets à classer,
2. Choix d'une dissimilarité \bar{s} entre objets (une distance ou une norme ou un produit scalaire),
3. Choix d'un indice d'agrégation Δ : mesure de dissemblance entre groupes d'objets.

Algorithme

Constitution : Groupes étape 1 : n singltons (chaque objet constitue un groupe) :
 $\mathcal{C}_1^{(1)}, \dots, \mathcal{C}_n^{(1)}$.

Comparaison : Agréger les deux groupes qui se ressemblent le plus (les moins dissemblables).

Itérations jusqu'à n'avoir plus qu'un groupe.

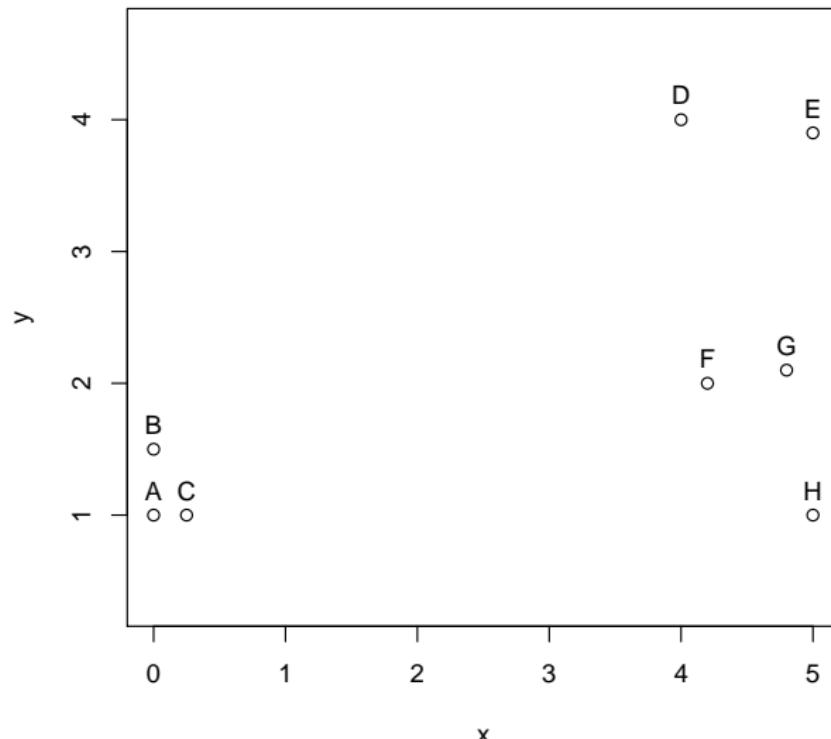
Exemple

- ▶ 8 objets dans \mathbb{R}^2
- ▶ Distance euclidienne
- ▶ indice d'aggrégation entre G_1 et G_2

$$\Delta(G_1, G_2) = \min_{\omega_i \in G_1; \omega_j \in G_2} d(\omega_i, \omega_j).$$

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

Exemple graphiquement

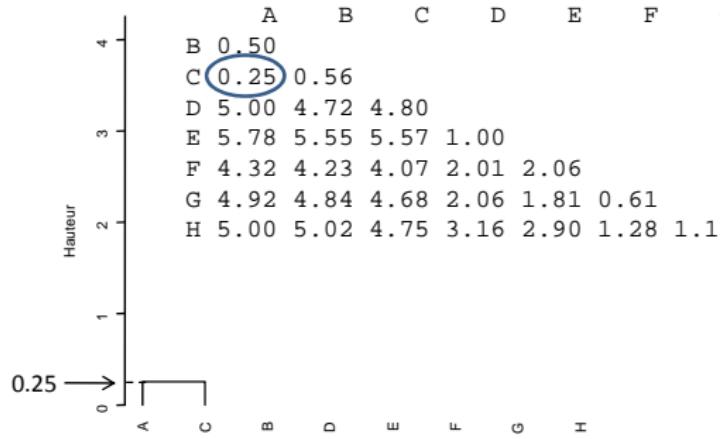
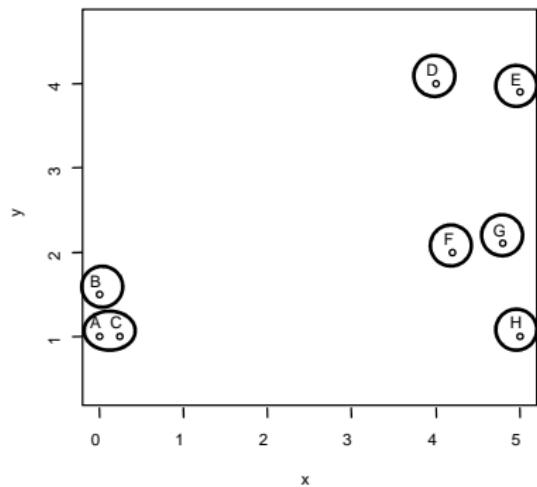


Classification

└ Classification Ascendante Hiérarchique

└ Construction de l'arbre sur un exemple

Etape 2

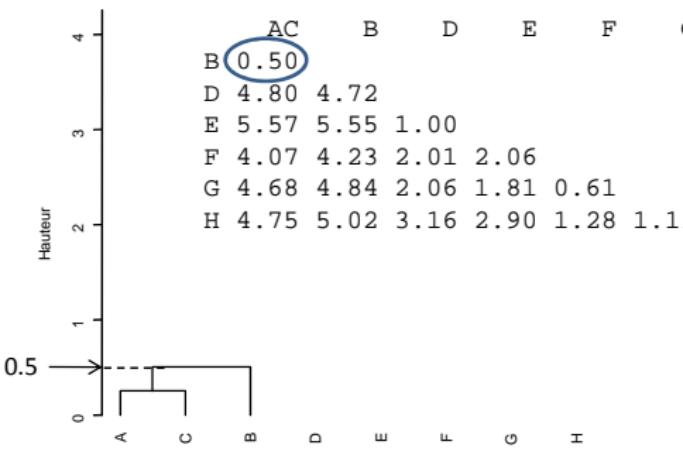
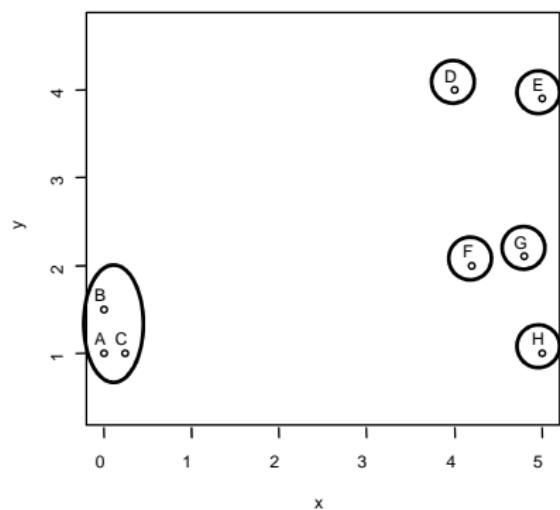


Classification

└ Classification Ascendante Hiérarchique

└ Construction de l'arbre sur un exemple

Etape 3

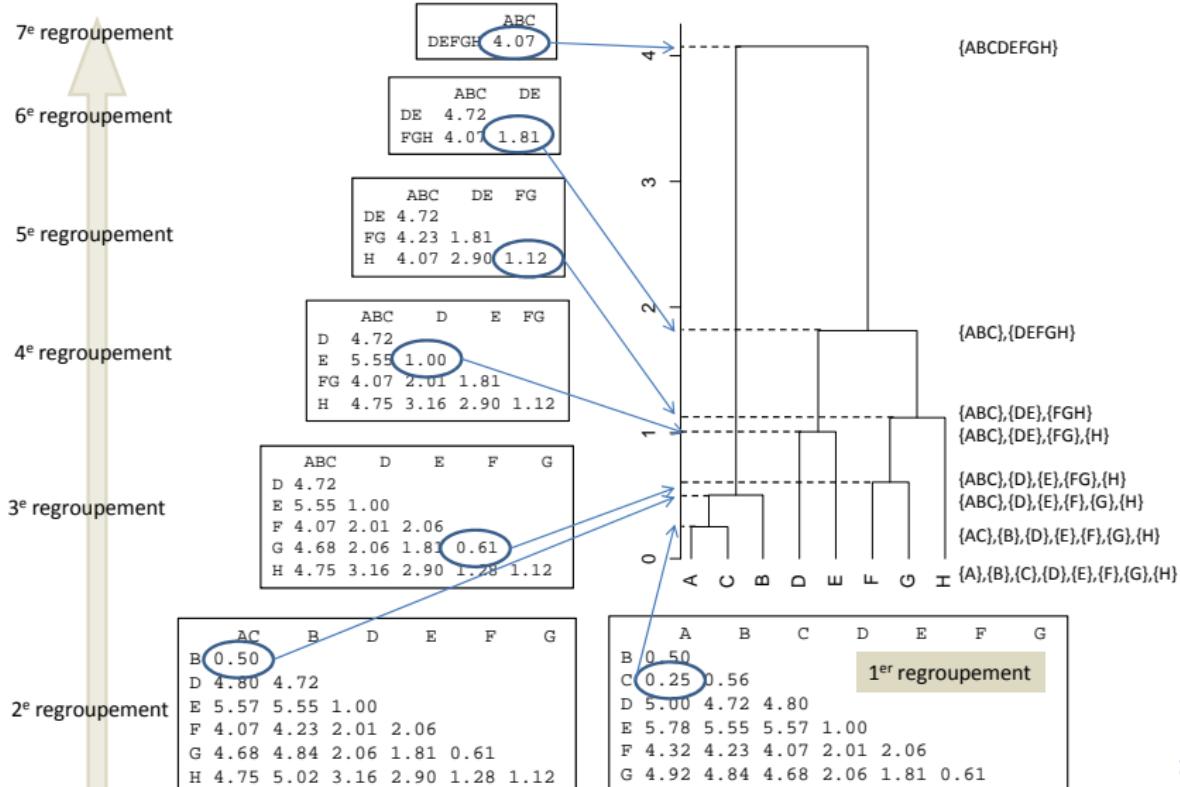


Classification

Classification Ascendante Hiérarchique

Construction de l'arbre sur un exemple

Récapitulatif



Propriétés des indices

- ▶ **Non-inversion** : la réunion de deux classes (non incluses l'une dans l'autre) présente toujours un indice d'agrégation plus grand que le maximum d'indice d'agrégation de chacune
- ▶ **Convexité** : si les objets à classer sont dans un espace euclidien, les enveloppes convexes des partitions générées par la CAH sont d'intersection vide (« convex admissibility »)
- ▶ **Invariance par réPLICATION** : si certains objets sont répliqués, les frontières des partitions générées par la CAH ne changent pas (« point proportionnal admissibility »);
- ▶ **Monotonie** : une transformation monotone des dissimilarités entre objets ne change pas la CAH (« monotone admissibility »).

Propriétés des indices

Indice	Non-inversion	Convexité	RéPLICATION	Monotonie
Saut minimum	Non-inversion	Non	Oui	Oui
Saut maximum	Non-inversion	Non	Oui	Oui
Moyenne	Non-inversion	Non	Non	Non
Ward	Non-inversion	Oui	Non	Non

Idéalement, il faudrait tenir compte des caractéristiques des données pour choisir un indice d'agrégation... mais c'est compliqué !

Dendrogramme

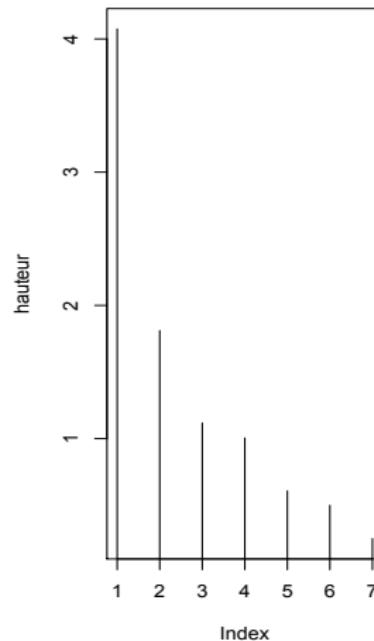
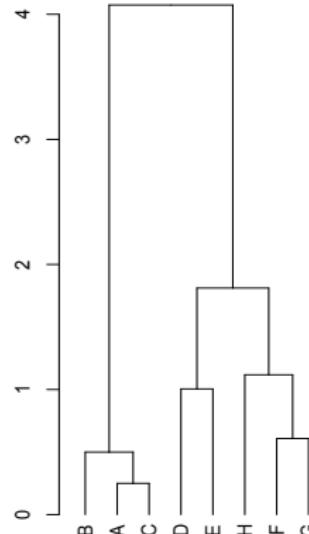
- ▶ Le dendrogramme représente, sous forme d'arbre binaire, les agrégations successives jusqu'à la réunion en une seule classe de tous les individus. On parle de **racine** (1 seule classe), de **feuilles** (n classes), de **branches** et de **noeuds**.
- ▶ La hauteur d'une branche est égale à l'indice de la hiérarchie, soit usuellement la distance (ultramétrique) entre les deux sous-groupes regroupés. La hauteur donne la difficulté pour deux groupes d'individus à être réunis dans le même groupe.
- ▶ Lorsqu'on coupe l'arbre, on peut comptabiliser le nombre de classes retenues.
- ▶ En coupant le dendrogramme au niveau d'un saut important, on espère obtenir une partition de bonne qualité : les individus regroupés auparavant étaient proches, tandis que ceux regroupés après la coupure deviennent trop éloignés.

Classification

└ Classification Ascendante Hiérarchique

└ Construction de l'arbre sur un exemple

Découpage (exemple 8 objets dans \mathbb{R}^2)



Conclusion pour la CAH

Avantages de la CAH :

- ▶ Il n'est pas nécessaire de fixer un nombre de classes a priori.
- ▶ La CAH ne dépend pas de conditions initiales (contrairement à la méthode des *K*-means).

Inconvénients de la CAH :

- ▶ La complexité algorithmique est en $O(n^2 \ln n)$: la CAH devient chronophage si le nombre d'individus est important.

Le coin R

Vous pouvez retrouver les résultats présentés en utilisant les fonctions **hclust** et **cutree** de la base, ou **hclust** du package **fastcluster**.

Essayer les différents indices d'agrégation avec l'argument method

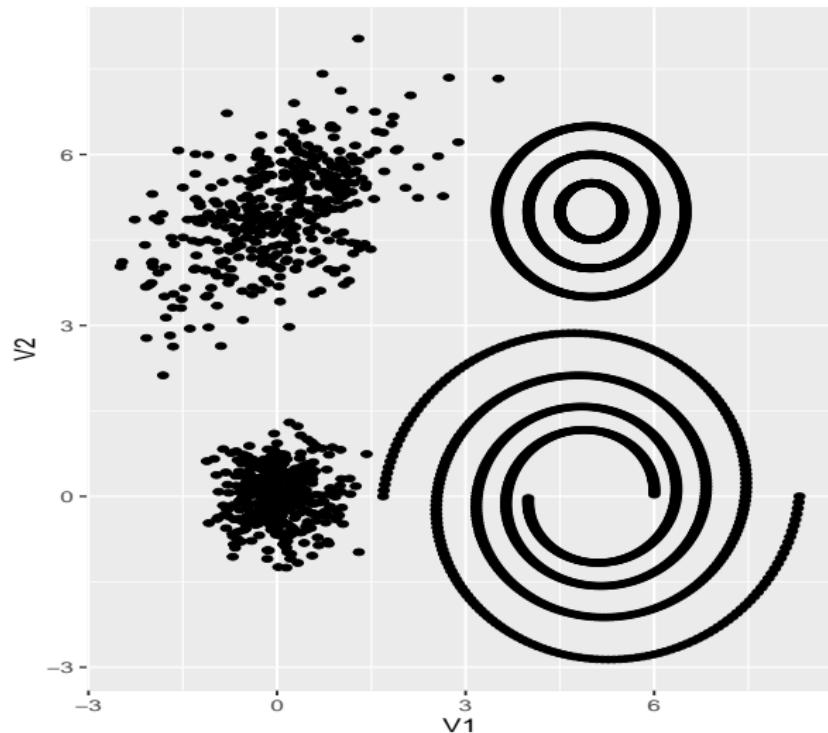
```
> donnees <- data.frame(x=c(0,0,0.25,4,5,4.2,4.8,5),  
+ y=c(1,1.5,1,4,3.9,2,2.1,1))  
> rownames(donnees) <- LETTERS[1:8]  
> round(dist(donnees),2)  
> cah <- hclust(dist(donnees), method="single")  
> plot(as.dendrogram(cah))  
> plot(sort(cah$height,dec=T),type="h")  
> gpcah <- cutree(cah,h=3)  
> plot(donnees,col=gpcah)
```

Classification

└ Classification Ascendante Hiérarchique

└ Construction de l'arbre sur un exemple

Un exemple jouet

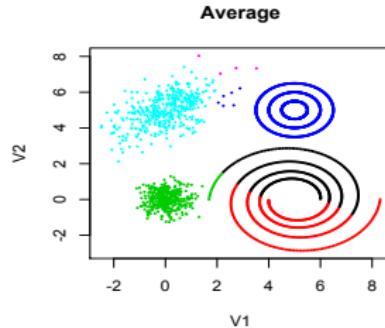
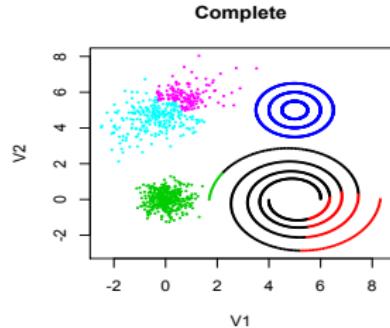
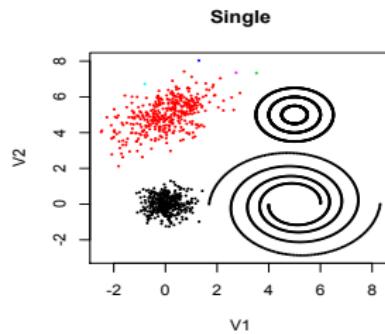
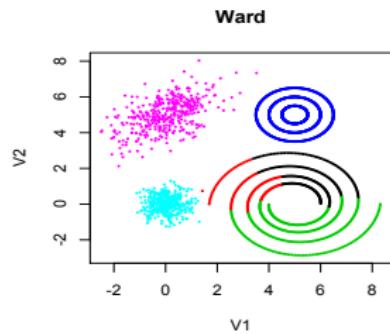


Classification

└ Classification Ascendante Hiérarchique

└ Construction de l'arbre sur un exemple

Exemple : CAH 6 groupes

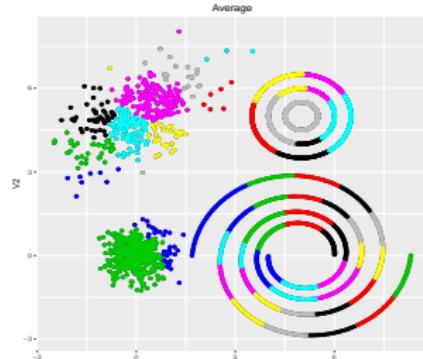
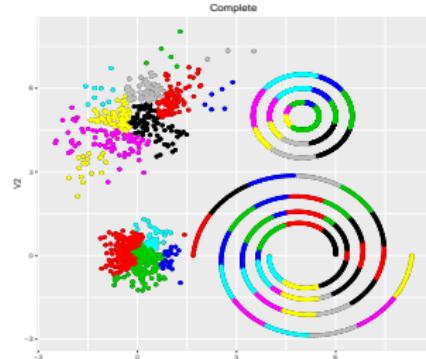
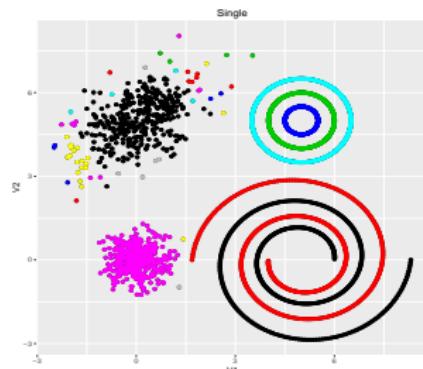
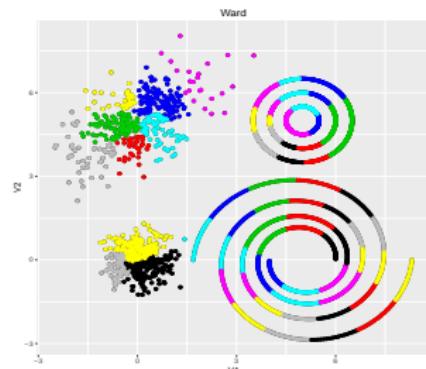


Classification

└ Classification Ascendante Hiérarchique

└ Construction de l'arbre sur un exemple

Exemple : CAH 40 groupes



Grands jeux de données

- ▶ En présence d'un **grand nombre d'individus** la CAH se révèle très **couteuse en temps de calcul** (contrairement au *k-means*).
- ▶ On peut alors procéder en 2 étapes :
 1. faire un *k-means* avec beaucoup de classes (par exemple 1000) ;
 2. faire la CAH sur les centres des classes calculés à l'étape précédente.

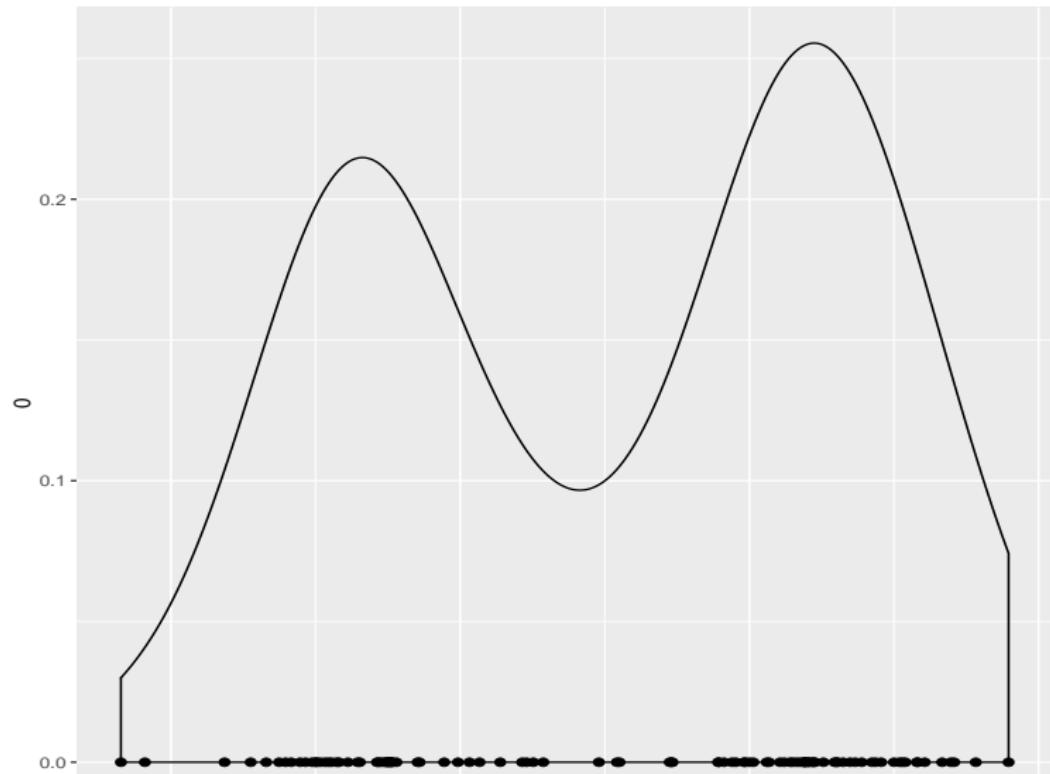
Consolidation de la CAH on n'en parle pas ?

- ▶ Le principe hiérarchique de la procédure fait que les partitions faites étapes après étapes ne sont **jamais remises en cause**.
- ▶ A l'issue de la CAH, des observations de certains groupes peuvent se trouver **proches de barycentres d'autres groupes**.
- ▶ La **consolidation** consiste alors à faire un **kmeans** en utilisant le **nombre de groupes** de la CAH et en initialisant avec les **centres des groupes de la CAH**.

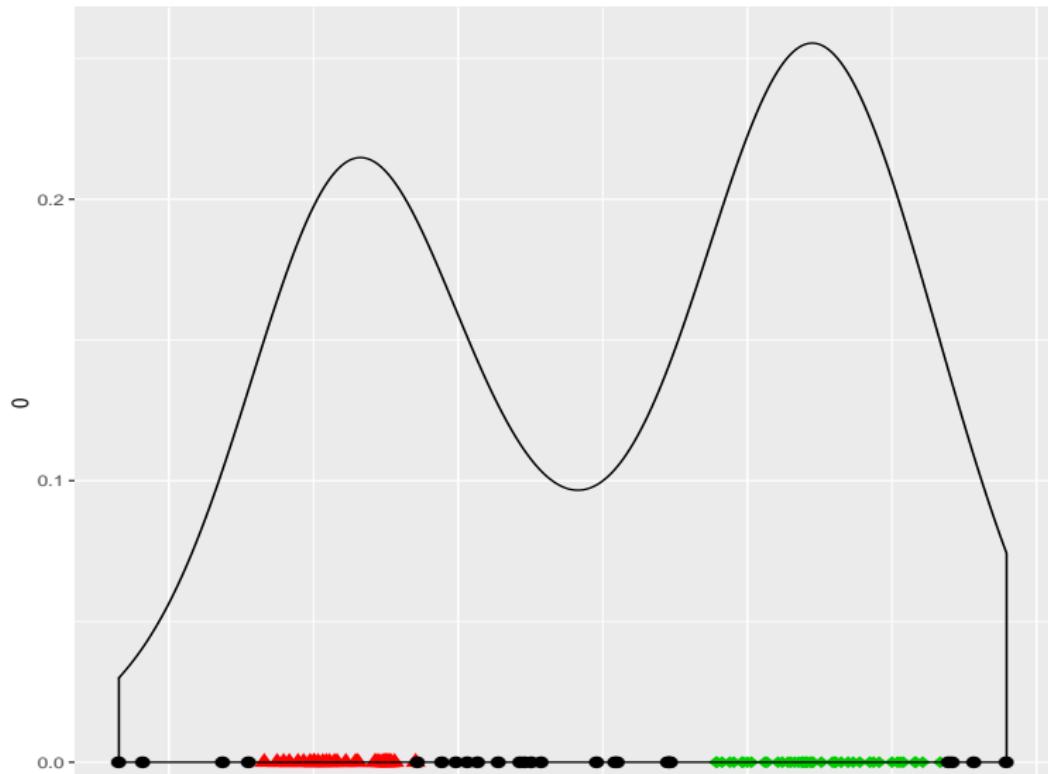
Introduction

- ▶ Le principe est de déterminer les **classes** d'une partition à partir des **zones de forte densité**.
- ▶ Les zones de **faible densité** sont utilisées pour **délimiter les classes**.
- ▶ Les éléments sont **regroupés de proche en proche** et les éléments éloignés des zones de forte densité sont ignorés et considérés comme des outliers.
- ▶ ? DBSCAN (Density-based spatial clustering of applications with noise)

L'idée



L'idée



DBSCAN : noyau et point de bordure

- ▶ Soit $\varepsilon > 0$ et $MinPts \leq n$ fixés.
- ▶ On note $B_\varepsilon(y)$ le voisinage centré sur y et de rayon ε et $|B_\varepsilon(y)|$ le nombre de points dans $B_\varepsilon(y)$.

Definition

- ▶ Si $|B_\varepsilon(y)| > MinPts$ alors y est un **noyau** et est dans une zone de forte densité.
- ▶ Si $|B_\varepsilon(y)| < MinPts$ alors y est un **point bordure** et n'est pas dans une zone de forte densité.
- ▶ Les **clusters** vont être constitués par des **noyaux proches**.
- ▶ Les points de bordure seront des **outliers** ou des **points de séparation des clusters**.

Accessibilité

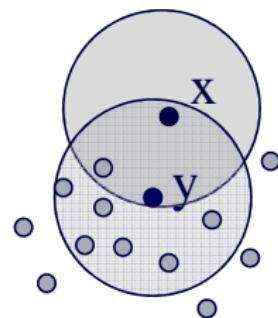
Definition

- ▶ x est **directement accessible** depuis y si $x \in B_\epsilon(y)$ et y est un noyau.
- ▶ x est **accessible depuis** y si il existe une chaîne de points $p_1 = y, p_2, \dots, p_k = x$ telle que $\forall i, p_{i+1}$ est directement accessible depuis p_i .

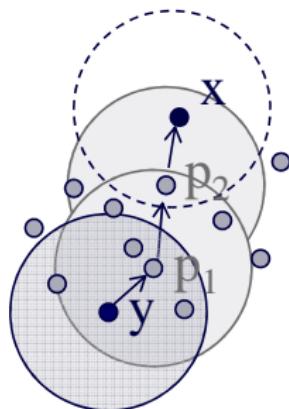
Definition

- ▶ Deux éléments x et y sont **connectés** s'ils sont tous les deux accessibles depuis un même élément z (l'élément z peut éventuellement être x ou y).
- ▶ Un cluster est **constitué** par un ensemble d'éléments connectés.

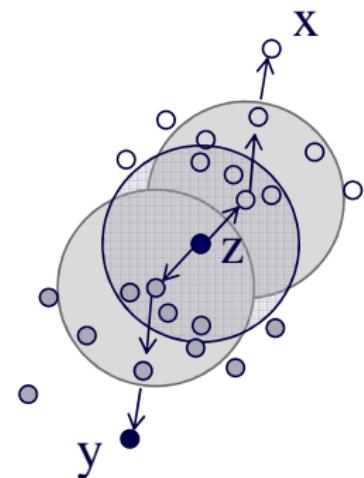
Exemple : $MinPts = 4$



x bordure, y noyau



x accessible depuis y
 y non accessible depuis x

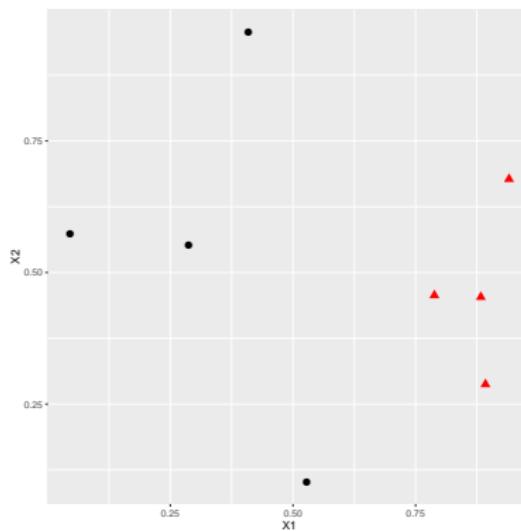


x et y connectés

Un exemple

```
> X[ind,]  
      X1          X2  
2  0.7883051  0.4566147  
4  0.8830174  0.4533342  
5  0.9404673  0.6775706  
8  0.8924190  0.2875775  
> dist(X[ind,])  
      2          4          5  
4  0.09476907  
5  0.26828125  0.23147891  
8  0.19852780  0.16602305  0.39294181  
> b <- dbSCAN(X, eps=0.25, minPts=3)
```

Résultat



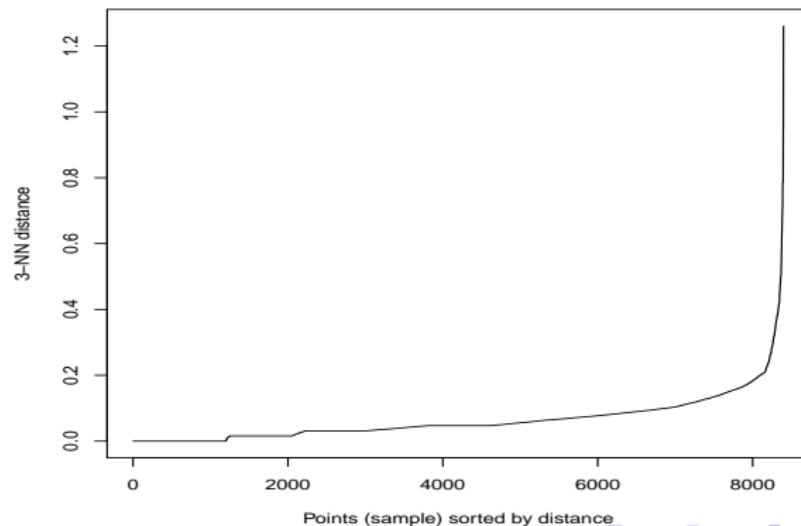
2,5,8 sont accessibles depuis 4 qui est un noyau. Tous ces points sont donc connectés.

Choix des paramètres

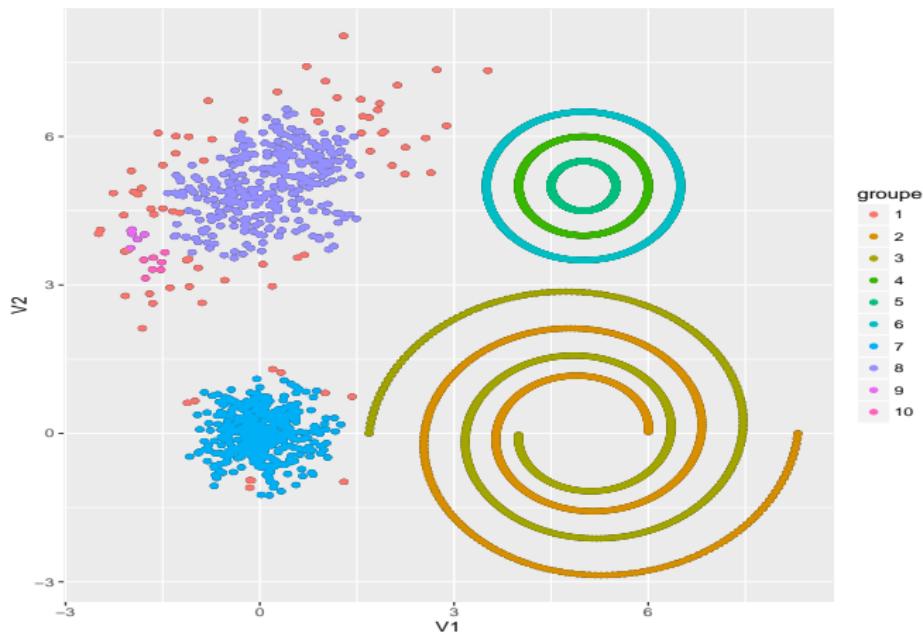
- ▶ 2 paramètres sont à calibrer ε et $minPts$; Leur choix est crucial...
- ▶ ε grand : **peu de groupes** et **réciproquement**.
- ▶ De façon empirique, il a été recommandé :
 - ▶ $minPts \approx 4$;
 - ▶ ε : calculer la distance entre chaque élément et son 3ème voisin. Tracer ces distances par ordre décroissant. Prendre pour ε la distance au coude.

Retour à l'exemple benchmark

```
> kNNdistplot(don[,1:2],k=3)
```

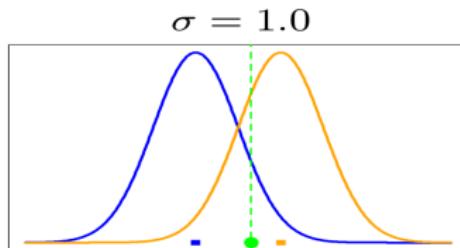


```
> db <- dbscan(don[,1:2], eps=0.25)
```

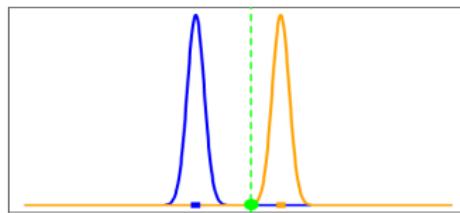


Introduction

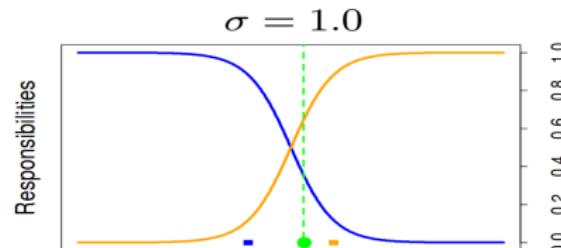
Considérons un mélange de loi $f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$



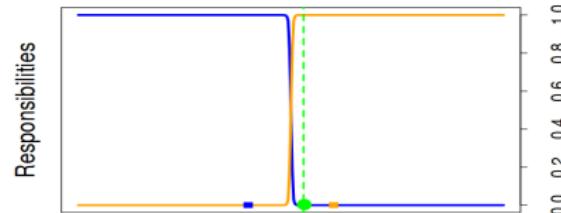
$\sigma = 0.2$



$\sigma = 0.2$



$\sigma = 0.2$



$\sigma = 0.2$

et les densités relatives $\pi_i f_i(x) / (\pi_1 f_1(x) + \pi_2 f_2(x))$. Si elles étaient connues, on affecterait un point à la classe de densité relative maximale.

Le principe est de déterminer des **sous-populations** dont la forme de la distribution est connue.

L'algorithme est présenté dans le cas classique d'un mélange de gaussiennes (unidimensionnelles ici), ce résultat peut être étendu à tout distribution qui admet une solution pour le maximum de vraisemblance.

Hypothèses

Soient n réalisations :

$$\{x_1, \dots, x_n\}$$

des variables aléatoires :

$$\{X_1, \dots, X_n\}$$

suivant k ($k < n$) lois normales :

$$\mathcal{N}(\mu_1, \sigma_1), \dots, \mathcal{N}(\mu_k, \sigma_k).$$

L'objectif est d'attribuer à ces variables aléatoires les lois qu'elles suivent. La problématique consiste à estimer les paramètres du modèle, c'est-à-dire la moyenne et la variance pour chaque composante, les pondérations ainsi que le nombre de composantes.

Exemple simple

Considérons un mélange de 2 gaussiennes à valeur dans \mathbb{R}

$$f(x) = \pi_1 \Phi(x, \mu_1, \sigma_1) + \pi_2 \Phi(x, \mu_2, \sigma_2).$$

Ayant un échantillon x_1, \dots, x_n nous écrivons la logvraisemblance

$$\mathcal{L}(x, \theta) = \sum_{i=1}^n \log (\pi_1 \Phi(x_i, \mu_1, \sigma_1) + \pi_2 \Phi(x_i, \mu_2, \sigma_2))$$

Trouver le maximum est compliqué !

Mais cela serait simple si on connaissait les étiquettes des x_i !

Exemple simple

Imaginons que l'on nous donne z_i qui indique la composante utilisée dans le mélange pour obtenir x_i . z_{i1} vaut 1 quand x_i vient de la composante 1 (i.e. $z_i = 1$) et 0 sinon, z_{i2} vaut 1 quand x_i vient de la composante 2 et 0 sinon. On peut écrire la log vraisemblance du modèle complété

$$\begin{aligned}\mathcal{L}(x, z, \theta) &= \sum_{i=1}^n z_{i1} \log \pi_1 + \sum_{i=1}^n z_{i1} \log \Phi(x_i, \mu_1, \sigma_1) \\ &\quad + \sum_{i=1}^n z_{i2} \log \pi_2 + \sum_{i=1}^n z_{i2} \log \Phi(x_i, \mu_2, \sigma_2)\end{aligned}$$

Pb : les z_i sont inconnus !!!

Exemple simple

Remplaçons les alors par leur valeur espérée

$$\gamma_k(x_i) = P(z = k|x_i)$$

et c'est ce que nous avons présenté comme étant la valeur des densités relatives

$$\gamma_k(x_i) = \frac{\pi_k \Phi(x_i, \mu_k, \sigma_k)}{\sum_{\ell} \pi_{\ell} \Phi(x_i, \mu_{\ell}, \sigma_{\ell})}$$

Pour estimer ces paramètres, nous allons utiliser l'algorithme itératif EM (expectaction/maximisation).

Algorithme

- ▶ On choisit au hasard $\mu_1^{iter}, \sigma_1^{iter}, \dots, \mu_K^{iter}, \sigma_K^{iter}, \pi_K^{iter}, \dots, \pi_K^{iter}$
 $iter = 0$.
- ▶ Expectation

Calculs des densités relatives

$$\gamma_k^{iter}(x_i) = \frac{\pi_k^{iter} \Phi(x_i, \mu_k^{iter}, \sigma_k^{iter})}{\sum_{\ell} \pi_{\ell}^{iter} \Phi(x_i, \mu_{\ell}^{iter}, \sigma_{\ell}^{iter})}$$

- ▶ Maximisation

$$\mu_k^{iter+1} = \frac{\sum_{i=1}^n x_i \gamma_k^{iter}(x_i)}{\sum_{i=1}^n \gamma_k^{iter}(x_i)}$$

Moyenne pondérée, idem pour le calcul des variances. Et

$$\pi_k^{iter+1} = \sum_{i=1}^n \gamma_k^{iter}(x_i) / n.$$

$iter = iter + 1$ et on repart à la partie expectation.

Choix de K

Pour un certain nombre de valeurs de K , on optimise les paramètres par la démarche précédente (EM). On obtient donc des $\hat{\theta}(K)$ et on peut aussi calculer la log-vraisemblance obtenue avec ces valeurs des paramètres, pénalisée par votre critère favori

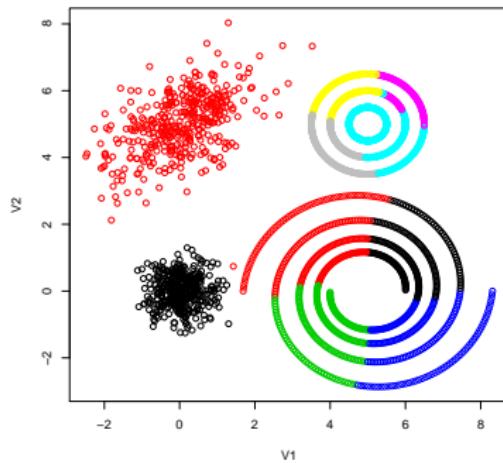
$$\mathcal{L}(x, \hat{\theta}(K)) - Pen(\#\hat{\theta}(K)).$$

Ici, $\#\hat{\theta}(K)$ désigne le nombre de paramètres estimés dans le modèle K .

On choisit K qui maximise la log-vraisemblance pénalisée.

L'exemple benchmark

10 classes retenues si l'on fait varier $K \in \{1, \dots, 12\}$. Attention les 2 groupes rouges sont différents (idem pour les noirs) mais il n'y a que 8 couleurs dans R.



Le coin R

Vous pouvez retrouver les résultats présentés en utilisant la fonction **Mclust** du package `mclust`. Essayer différents nombres de composantes.

```
> don <- read.table("data_benchmark1.txt",
+ sep=";",header=TRUE)
> melange <- Mclust(don,1:12)
> summary(melange)
> names(summary(melange))
> melange$bic
> plot(don, col=summary(melange)$classification,
+ cex=.3,main="K=10")
```

Les données quantitatives

X	
1	d

Individus dans \mathbb{R}^d

Ligne = Individu

Colonne=Variable Quantitative

1	1	d
	n	

Variables dans \mathbb{R}^n

Exemple de données

- ▶ $d = 10$ variables **quantitatives**, mesurées pour $n = 41$ athlètes
- ▶ Variables quantitatives continues : 100m, longueur, poids, hauteur, 400m, 110m haies, disque, perche, javelot, 1500m

Jeu de données présenté dans ? et accessible :

[http://math.agrocampus-ouest.fr/infoglueDeliverLive/
enseignement/support2cours/livres/statistiques.avec.R](http://math.agrocampus-ouest.fr/infoglueDeliverLive/enseignement/support2cours/livres/statistiques.avec.R)

Exemple (suite)

Premier individu

	100m	Longueur	Poids	Hauteur	400m	110m.haies
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05

	Disque	Perche	Javelot	1500m
Sebrle	48.72	5	70.52	280.01

Variable 100m (première colonne)

10.85	10.44	10.50	10.89	10.62	10.91	10.97	10.80	10.69	10.98
10.90	11.14	10.85	10.55	10.68	10.89	11.06	10.87	11.14	10.92
11.08	11.10	11.33	10.86	11.23	11.36	11.04	10.76	11.02	11.02
11.11	11.13	10.83	11.64	11.37	11.33	11.33	11.36		

Objectif(s)

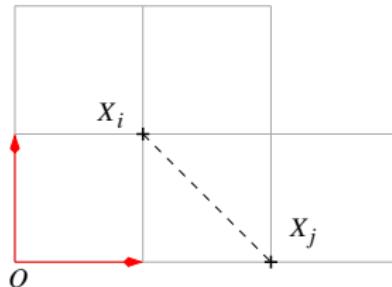
Décrire le jeu de données

1. Quels sont les individus qui se ressemblent (ou non) ? Distance entre individus mais il faudrait alors calculer et analyser $n(n - 1)/2$ distances
2. Quelles sont les relations entre les variables **quantitatives** (celles qui se ressemblent ou non) ? Analyse des corrélations ? Mais alors analyse des corrélations 2 à 2, comment faire pour obtenir une information plus globale ?

L'ACP permet d'étudier les ressemblances entre individus et de dégager des profils. Elle permet également de réaliser un bilan des liaisons linéaires entre variables. En reliant les deux études : caractériser des individus par les variables.

Distance entre individus

$$d^2(X_i, X_j) = \sum_{k=1}^d (X_{ik} - X_{jk})^2 = \langle X_i - X_j, X_i - X_j \rangle$$



Problème : n individus $\rightarrow n(n - 1)/2$ distances à analyser !

Proximités entre variables

Classiquement on utilise la covariance (ou la corrélation)

$$\text{cov}(Z, Y) = \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y}).$$

Ecart à la moyenne.

Dans certaines applications, il est utile de travailler avec des poids p_i éventuellement différents d'un individu à l'autre (e.g. données regroupées) auquel cas

$$\text{cov}(Z, Y) = \sum_{i=1}^n p_i (Z_i - \bar{Z})(Y_i - \bar{Y}).$$

Proximités entre variables (suite)

Si les variables sont **centrées** ($\bar{Y} = \bar{Z} = 0$) et les individus de même poids $1/n$ alors

$$\text{cov}(Z, Y) = \frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y}) = \frac{1}{n} \langle Z, Y \rangle = \langle Z, Y \rangle_*$$

Si de plus les variables sont **réduites** alors

$$\rho(Z, Y) = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2} \sqrt{\sum_{i=1}^n (Z_i - \bar{Z})^2}} = \langle Z, Y \rangle_*$$

Si les individus ont des poids différents, on travaille avec le produit scalaire associé à $D = \text{diag}(p_1, \dots, p_n)$.

Echelle de mesure

Pour réaliser une ACP, il est nécessaire de centrer les variables.

Quand les variables sont mesurées sur différentes échelles, il faut en outre les réduire (l'ACP donne à chaque variable un poids correspondant à son écart-type). Si les variables sont mesurées à la même échelle (même unité), réduire ou non donne lieu à deux analyses différentes.

centrage origine repère = point moyen (ou centre de gravité)

réduction toutes les variables ont même unité de variation :

s'affranchir des hétérogénéités des unités de mesures.

Dans ce cas **toutes les variables sont de longueur 1** car

$$\|Y\|_D^2 = \langle Y, Y \rangle_D = \rho(Y, Y) = 1$$

Fait **automatiquement** par le logiciel (argument scale).

Illustration graphique 1



Illustration graphique 2



→ Séparer le plus possible les points !

ACP naïve et illustration graphique

La méthode

1. Centrage-réduction des données
2. Recherche d'axes qui permettent de séparer au mieux les individus
Ou recherche d'axes qui permettent de conserver le plus de variation (de variabilité)
Vecteurs directeurs orthonormés (repère)
3. Représentation graphique des individus projetés sur ces axes, plan par plan

Inertie

On appelle **Inertie totale** du nuage de points

$$I(\mathcal{N}) = \frac{1}{n} \sum_{i=1}^n d^2(X_i, \bar{X}).$$

L'inertie portée par un sous espace \mathcal{F} est l'inertie du nuage projeté

Remarque : si les individus ont des poids différents alors l'inertie est donnée par

$$I(\mathcal{N}) = \sum_{i=1}^n p_i d^2(X_i, \bar{X}) \quad \text{avec} \quad \sum_{i=1}^n p_i = 1$$

Projection sur un axe

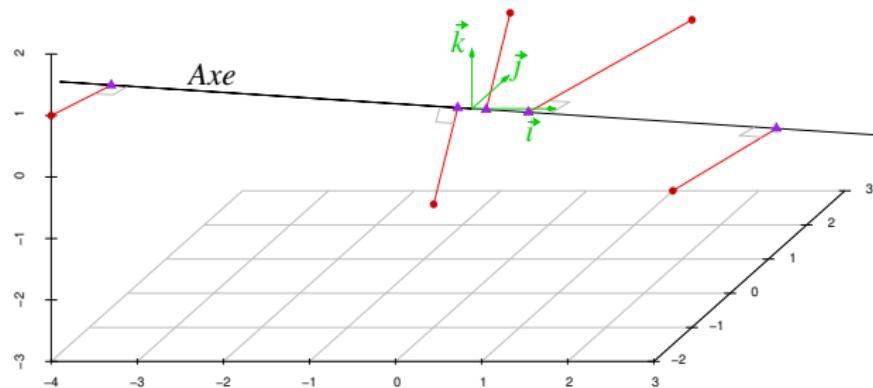


Figure – Projection sur un axe

Objectif : trouver une droite de vecteur unitaire $a_1 \in \mathbb{R}^d$ tq l'inertie du nuage projeté soit maximum.

Choix du sous-espace de dimension 1

Chercher a_1 unitaire qui maximise l'inertie portée par a_1 revient à

$$\text{maximiser} \quad \frac{1}{n} a_1' X' X a_1 \quad \text{sc} \quad \|a_1\|^2 = 1.$$

La solution est obtenue par le vecteur propre normé associé à la plus grande valeur propre λ_1 de la matrice $\Sigma = \frac{1}{n} X' X$.

Remarques

- ▶ Σ est symétrique semi-définie positive elle est diagonalisable et toutes ses valeurs propres sont positives ou nulles
- ▶ a_1 est appelé premier axe principal
- ▶ les coordonnées sur l'axe 1 sont regroupées dans un vecteur \tilde{c}_1 obtenu par $\tilde{c}_1 = Xa_1$ où $\|\tilde{c}_1\|_*^2 = \lambda_1$:

$$\frac{1}{n} \tilde{c}'_1 \tilde{c}_1 = \frac{1}{n} a'_1 X' X a_1 = a'_1 \Sigma a_1 = a'_1 \lambda_1 a_1 = \lambda_1.$$

Ce vecteur, obtenu comme combinaison linéaire “optimale” des variables initiales, est appelé composante principale.

Et ainsi de suite

Pour trouver le second axe, il faut maximiser l'inertie et être orthogonal à a_1 .

La solution est le vecteur propre normé a_2 associé à la seconde plus grande valeur propre λ_2 de Σ . La seconde composante principale est Xa_2 .

Et ainsi de suite...

Chercher les axes factoriels revient à diagonaliser Σ et donc à faire un changement de base.

Choix des axes

L'inertie portée par le j ème axe est λ_j et l'inertie totale est

$$I(\mathcal{N}) = \text{tr}(\Sigma) = \sum_{j=1}^d \lambda_j$$

Le pourcentage d'inertie portée par l'axe k est

$$\frac{\lambda_k}{\sum_{j=1}^d \lambda_j} \times 100.$$

Pas de méthode pour choisir le nombre d'axes, critères empiriques.

Choix des axes (suite)

Analyse de la décroissance des valeurs propres : on retient en général le nombre d'axes donné par le "coude".

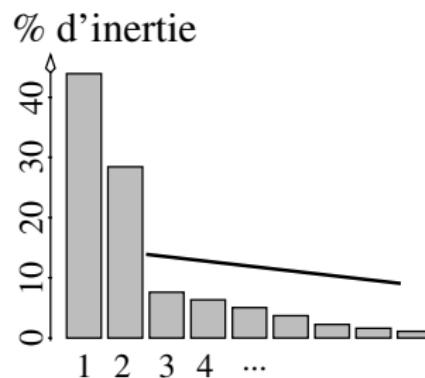
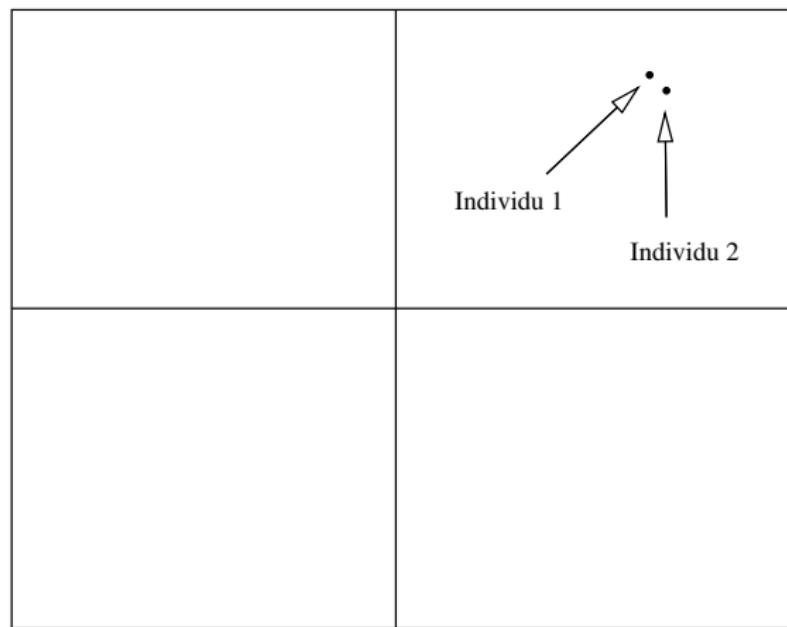


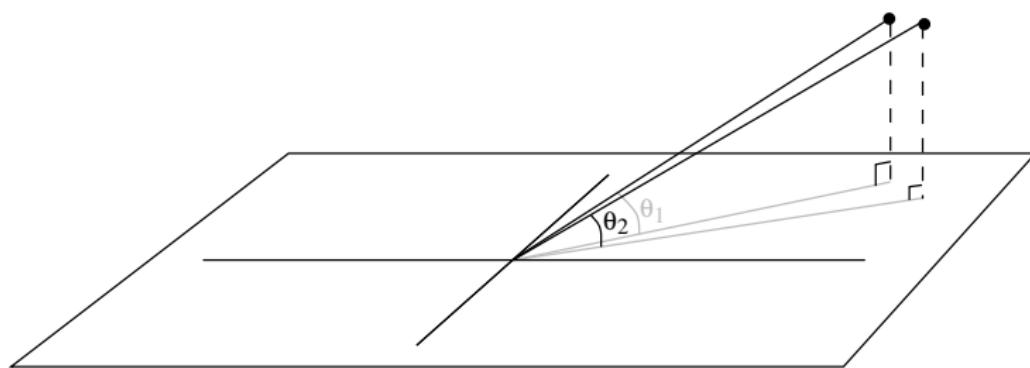
Figure – Valeurs propres et inertie de chaque axe

Comment mesurer la qualité de représentation d'un individu sur le sous-espace choisi ?

Individus projetés

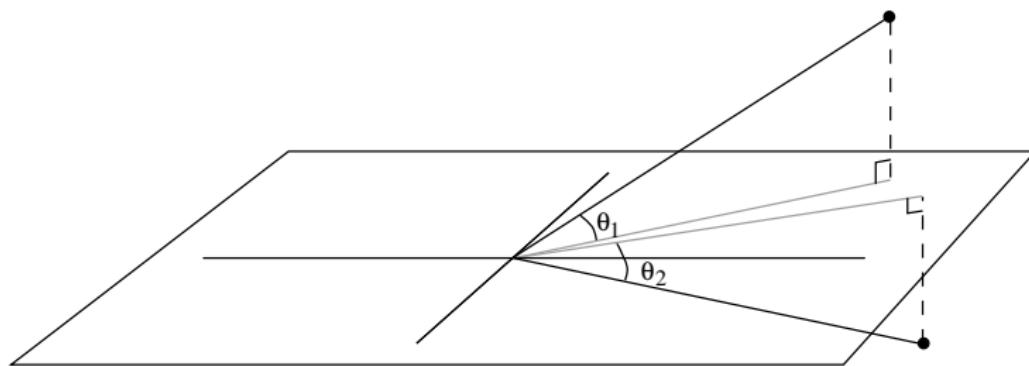


Individus proches



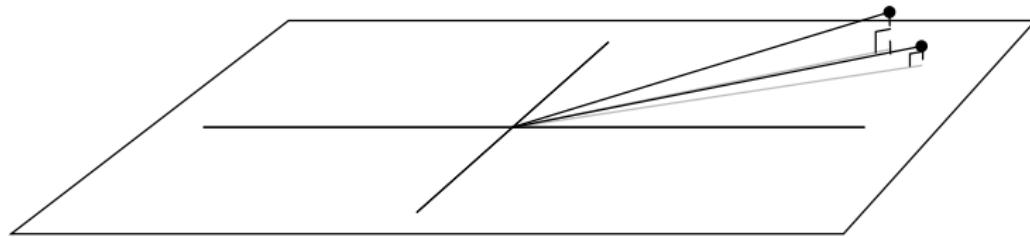
Mais éloignés du plan on pourrait avoir

Individus éloignés



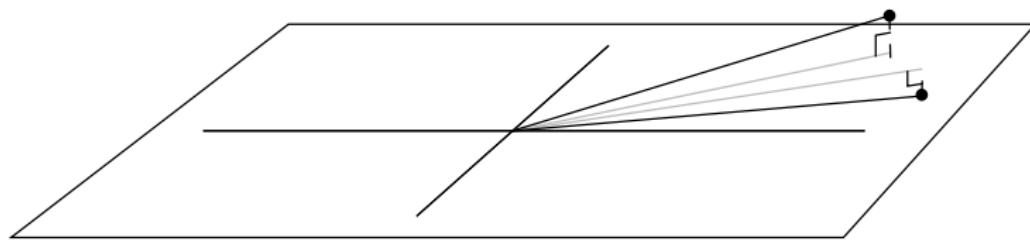
Alors que si l'angle est petit

Individus (très) proches



les individus restent proches même si les angles sont opposés.

Individus proches



Calcul de l'angle

En fait on calcule le \cos^2 grâce à Pythagore. En effet

$$\cos^2(\theta_i) = \frac{\tilde{c}_{1i}^2 + \tilde{c}_{2i}^2}{\|x_i\|^2}$$

En regardant la projection des points sur le sous-espace et en analysant le \cos^2 des angles on peut savoir si des individus sont proches (mais rien ne vous empêche de retourner voir le tableau de données).

- ▶ On s'intéresse au nuage des variables $\{X^1, \dots, X^d\}$
- ▶ Pour prendre en compte les poids des individus on munit \mathbb{R}^n de la métrique $D = \text{diag}(p_1, \dots, p_n)$ où p_i est le poids de l'individu i (dans le cas usuel, $p_i = 1/n$).

On a

- ▶ $\|X^j\|_D$ est l'écart-type de X^j ($=1$ dans le cas réduit)
- ▶ $\frac{\langle X^k, X^l \rangle_D}{\|X^j\|_D \|X^k\|_D} = \rho(X^k, X^l) = \cos(X^k, X^l)$

Une méthode naturelle pour donner une signification à une composante principale c est de la lier aux variables initiales en calculant les $\rho(c, X^j)$. Pour un couple de composantes principales c_1, c_2 on synthétise usuellement ces corrélations sur une figure appelée **Cercle des corrélations**.

Objectif en termes de variables

Rechercher un vecteur c_1 D -normé (\star) à l'unité de \mathbb{R}^n qui soit le plus corrélé avec les variables initiales. Il faut donc maximiser

$$\sum_{j=1}^d \rho^2(X^j, c_1) = \sum_{j=1}^d (\langle X^j, c_1 \rangle_\star)^2.$$

Cela s'écrit

$$\text{maximiser } \frac{1}{n} c_1' X X' c_1 \quad \text{sc} \quad \|c_1\|_D^2 = 1,$$

ou encore

$$\text{maximiser } c_1' D X X' D c_1 \quad \text{sc} \quad c_1' D c_1 = 1.$$

La solution est obtenue par le vecteur propre normé associé à la plus grande valeur propre λ_1 de la matrice $WD = XX'D$.

Remarque : formule de transition

$$\sum a_1 = \lambda_1 a_1 \quad \text{or} \quad \Sigma = \frac{1}{n} X' X$$

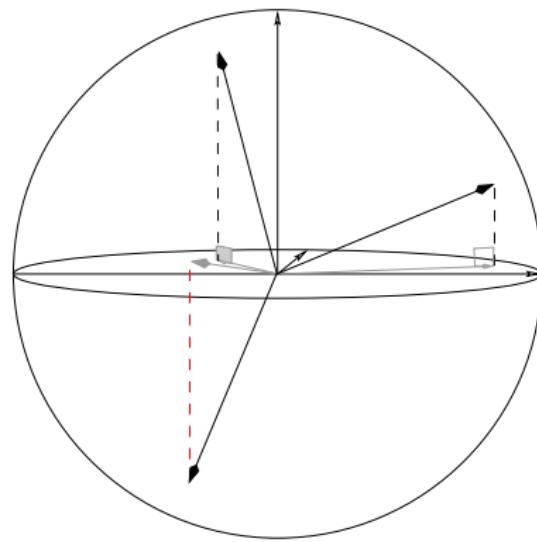
$$X' D X a_1 = \lambda_1 a_1$$

$$X X' D X a_1 = \lambda_1 X a_1$$

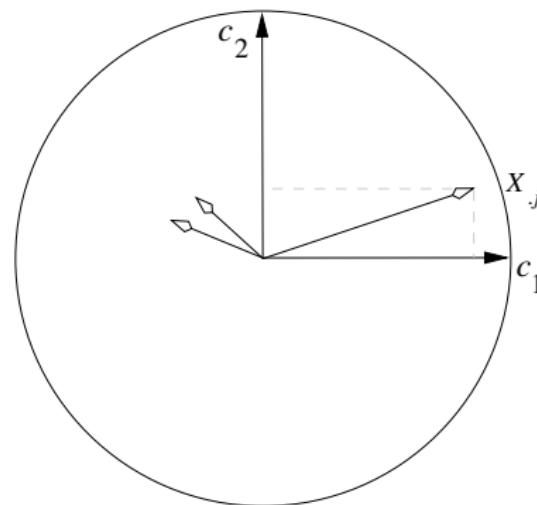
$$X X' D \tilde{c}_1 = \lambda_1 \tilde{c}_1.$$

Dans le cas de données centrées-réduites, le cercle des corrélations fournit la projection des variables initiales (centrées-réduites) sur le sous-espace engendré par les composantes principales correspondantes.

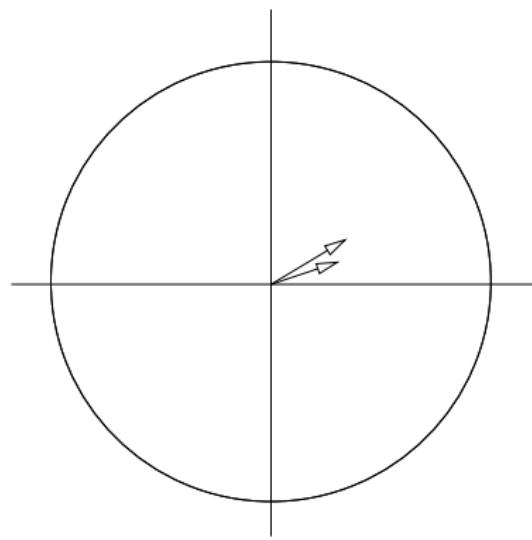
Cercle de corrélation (3D)



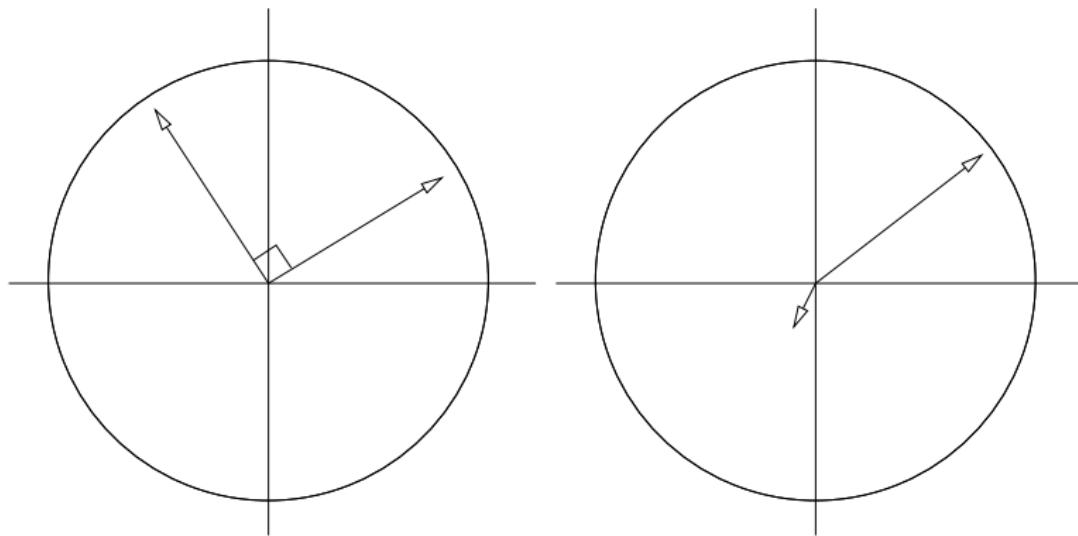
Cercle de corrélation (2D) : projection



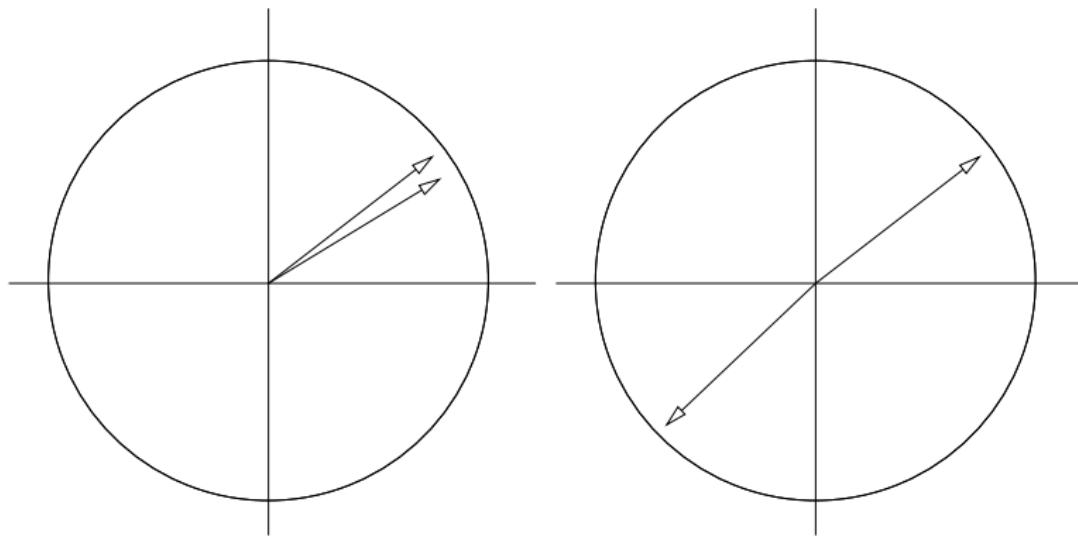
Aucune interprétation



Non corrélation



Corrélation



Relation entre les analyses individus/variables

- ▶ Analyse de $X'X$ pour l'analyse des individus
- ▶ Analyse de XX' pour l'analyse des variables

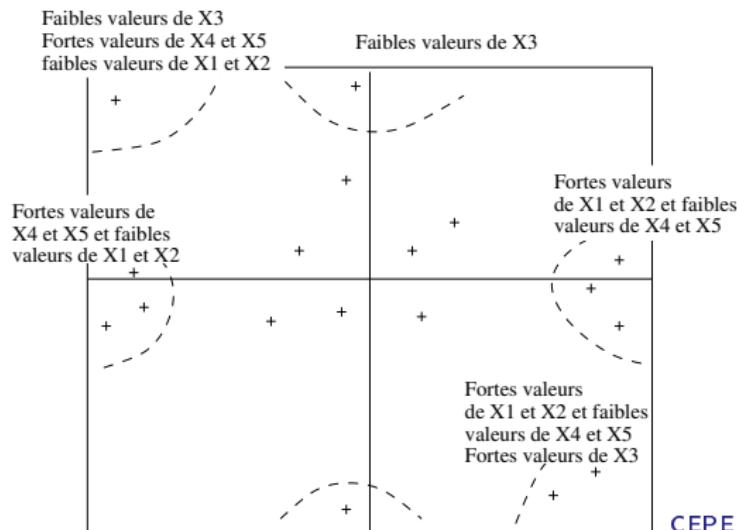
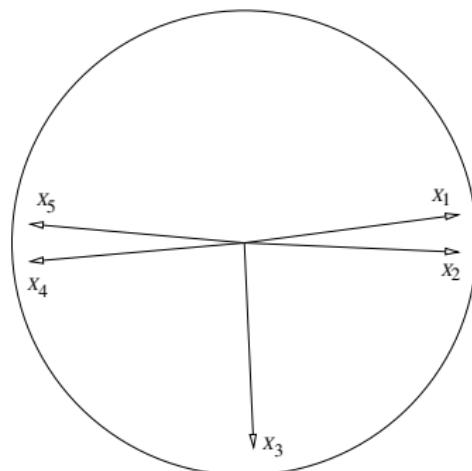
Même valeurs propres

- ▶ Les axes factoriels de \mathbb{R}^n (ceux des variables) se déduisent de ceux de \mathbb{R}^d (ceux des individus)

$$c_i = Xa_i$$

Variables ↔ individus

Représentation des variables sur un cercle de corrélations ;
 représentation des individus dans le plan engendré par deux axes principaux.



Pour différentes raisons on peut être amené à

- ▶ analyser des individus supplémentaires (nouvelles observations qu'on a décidé de ne pas utiliser dans l'analyse des individus aberrants)
- ▶ analyser des variables supplémentaires (nouvelles variables ou variables calculées à partir des variables initiales cf décathlon)

On peut être intéressé par visualiser ces individus ou variables. Pour cela

- ▶ mêmes transformations (centrage réduction)
- ▶ projection

A faire en TP avec les données de décathlon.

Le coin R

- ▶ La fonction **princomp** permet de réaliser une ACP de façon simpliste. La fonction **PCA** du package **FactoMineR** permet l'ajout d'éléments supplémentaires et la construction simple de graphiques.
- ▶ Les variables qualitatives supplémentaires sont représentées sur le graphique des individus (chaque modalité est représentée au barycentre des individus prenant cette modalité).
- ▶ Les variables quantitatives supplémentaires sont représentées sur le cercle des corrélations.

Le coin R

```
> decath<-read.table("decathlon.csv",sep="; ",  
+ dec=". ",header=T,row.names=1)  
> res.pca=PCA(decath,quanti.sup=11:12,quali.sup=13)  
> names(res.pca)  
> barplot(res.pca$eig[,1],  
+ names=paste("Dim",1:nrow(res.pca$eig)),  
+ main="inertie expliquée")  
> names(res.pca$ind)  
> plot(res.pca,choix="ind",habillage=13,cex=.7)
```

Contexte

- ▶ AFC : Analyse Factorielle des Correspondances.
- ▶ On considère ici de 2 variables qualitatives X^1 et X^2 , avec respectivement m_1 et m_2 modalités, observées sur n individus.
- ▶ Pour $i \in \{1, \dots, m_1\}$ et $j \in \{1, \dots, m_2\}$, on note n_{ij} le nombre d'individus ayant comme modalité i pour la variable X^1 et j pour la variable X^2 .
- ▶ On peut représenter cet ensemble de données dans le tableau de contingence \mathbb{N} à m_1 lignes et m_2 colonnes, de terme n_{ij} .

Notations

On peut en déduire la fréquence conjointe associée :

$$f_{ij} = \frac{n_{ij}}{n} .$$

On utilise les notation suivantes :

$$\sum_{i=1}^{m_1} n_{ij} = n_{\cdot j} .$$

$$\sum_{j=1}^{m_2} n_{ij} = n_{i\cdot} .$$

$$\sum_{i=1}^{m_1} f_{ij} = f_{\cdot j} .$$

$$\sum_{j=1}^{m_2} f_{ij} = f_{i\cdot} .$$

Profils lignes

Afin de comparer les lignes entre elles, on divise chaque ligne par la somme de la ligne.

On travaille donc sur les quantités :

$$\frac{n_{ij}}{n_{i\cdot}} .$$

Chaque point i a pour coordonnées $\frac{n_{ij}}{n_{i\cdot}}$ pour $j \in \{1, \dots, m_2\}$ avec $n_{i\cdot}$ comme poids.

Le barycentre des points est donc $(f_{1\cdot}, \dots, f_{m_2\cdot})^\top$.

Profils colonnes

Afin de comparer les colonnes entre elles, on divise chaque colonne par la somme de la colonne.

On travaille donc sur les quantités :

$$\frac{n_{ij}}{n_{\cdot j}}.$$

Chaque point j a pour coordonnées $\frac{n_{ij}}{n_{\cdot j}}$ pour $i \in \{1, \dots, m_1\}$ avec $n_{\cdot j}$ comme poids.

Le barycentre des points est donc $(f_{\cdot 1}, \dots, f_{\cdot m_1})^\top$.

Métrique du khi-deux

Pour 2 individus i et i' :

$$d^2(i, i') = n \sum_{j=1}^{m_2} \frac{1}{n_{\cdot j}} \left(\frac{n_{ij}}{n_{i \cdot}} - \frac{n_{i'j}}{n_{i' \cdot}} \right)^2.$$

Pour 2 variables j et j' :

$$d^2(j, j') = n \sum_{i=1}^{m_1} \frac{1}{n_{i \cdot}} \left(\frac{n_{ij}}{n_{\cdot j}} - \frac{n_{ij'}}{n_{\cdot j'}} \right)^2.$$

AFC et ACP

Soient :

$$D_1 = \text{diag}(n_{1.}, \dots, n_{m_1.}) ,$$

$$D_2 = \text{diag}(n_{.1}, \dots, n_{.m_2}) .$$

On peut alors, de manière duale, effectuer une ACP :

- ▶ Sur les profils lignes

On considère le tableau de données $\mathbb{X} = D_1^{-1}\mathbb{N}$ avec les poids $W = \frac{D_1}{n}$.

La métrique considérée est $Q = n D_2^{-1}$.

- ▶ Sur les profils colonnes

On considère le tableau de données $\mathbb{X} = D_2^{-1}\mathbb{N}^\top$ avec les poids $W = \frac{D_2}{n}$.

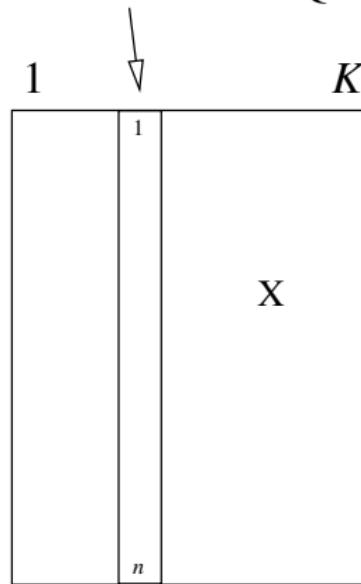
La métrique considérée est $Q = n D_1^{-1}$.

Résultats

- ▶ On peut superposer les 2 nuages projetés, centrés sur le barycentre.
- ▶ On s'intéresse essentiellement aux points ayant une forte contribution relative, étant donné que la contribution mesure la proximité entre les points et les axes.
- ▶ On pourra observer que :
 - ▶ Deux modalités de la même variable sont proches si leurs profils sont similaires.
 - ▶ Deux modalités de 2 variables différentes sont proches si leurs individus respectifs ont des barycentres proches.

Tableau de Données

Colonne=Variable Qualitative



K variables qualitatives mesurées sur n individus

Objectif

- ▶ Objectif : décrire le jeu de données
 1. Quelles sont les relations entre les K variables **qualitatives** (celles qui se ressemblent ou non) ?
 2. Quels sont les individus qui se ressemblent (ou non) ?
- ▶ Problèmes
 1. Comment généraliser à plusieurs variables (plus facile si $K = 2$)
 2. Comment connaître les modalités liées (qui contribuent à l'écart à l'indépendance) ?
 3. Comment représenter les individus ?
- ▶ Idée : se servir du cadre de l'analyse factorielle (ACP)
 - ▶ Distance entre individus
 - ▶ Distance entre variables (ou modalités?)

Remarque : des variables quantitatives peuvent être intégrées à l'étude à condition d'être découpées en classes et considérées comme facteurs.

Exemple d'un tableau de variables qualitatives

Le tableau de départ

$$Y = \begin{bmatrix} A1 & B2 & C3 \\ A2 & B1 & C1 \\ A2 & B2 & C2 \\ A3 & B2 & C1 \\ A3 & B1 & C2 \end{bmatrix}$$

Remplacer A1 par 1, A2 par 2, A3 par 3 : idiot.

Codage disjonctif complet du tableau

Le tableau d'arrivée : codage disjonctif complet

$$U = \left[\begin{array}{ccc|cc|ccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{array} \right]$$

somme des colonnes → $n_1^{(1)} n_2^{(1)} n_3^{(1)} | n_1^{(2)} n_2^{(2)} | n_1^{(3)} n_2^{(3)} n_3^{(3)}$

Nous pouvons faire des calculs sur U

- ▶ Produits scalaires entre individus (lignes de U)
- ▶ Produits scalaires entre modalités (colonnes de U)

Distance modifiée

Cependant distance classique compte le nombre de fois que les modalités apparaissent en même temps (ne dépend pas de l'effectif).

Division de chaque colonne $k(j)$ (modalité j de la variable k) de U par la fréquence d'apparition de la modalité k (notée $n_j^{(k)}/n$).

Résumé

1. Tableau de variables qualitatives X
2. Transformation en tableau disjonctif complet U
3. Division des colonnes par la fréquence de la modalité (colonne) en question
4. ACP du tableau obtenu sans réduction.

La fonction **MCA** du package FactoMineR fait cela automatiquement.