Classification

CEPE

Fev 2025

Plan

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).

		Variables				
		1	...	j	...	d
Individus	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 :

$$X_i = \left(x_i^1, \ldots, x_i^d\right)^\top$$

et la variable j avec le vecteur :

$$X^j = \left(x_1^j, \ldots, x_n^j\right)^\top .$$

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

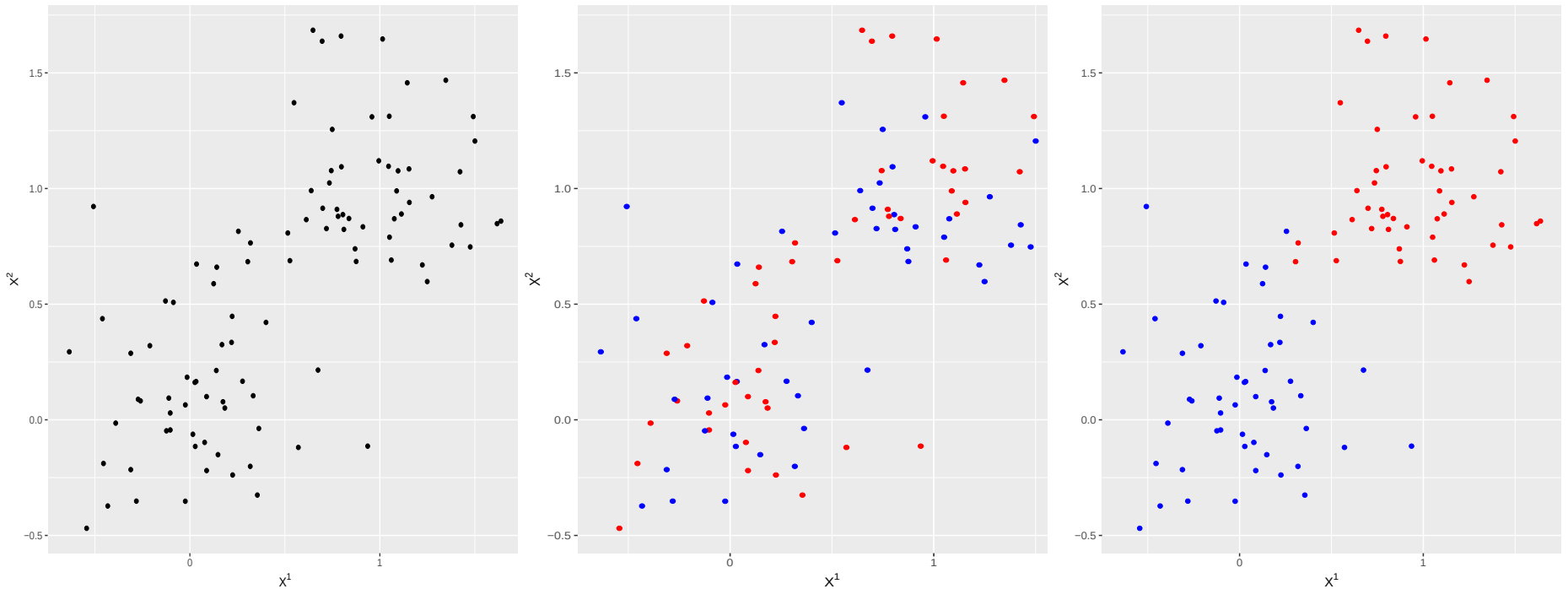
- $\forall k \in \{1, \ldots, K\} : \mathcal{C}_k \neq \emptyset$,
- $\forall \{k, k'\} \in \{1, \ldots, K\}^2 : \mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$,
- $\bigcup_{k \in \{1,\ldots,K\}} \mathcal{C}_k = \mathcal{O}$.

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

Caractérisation des classes

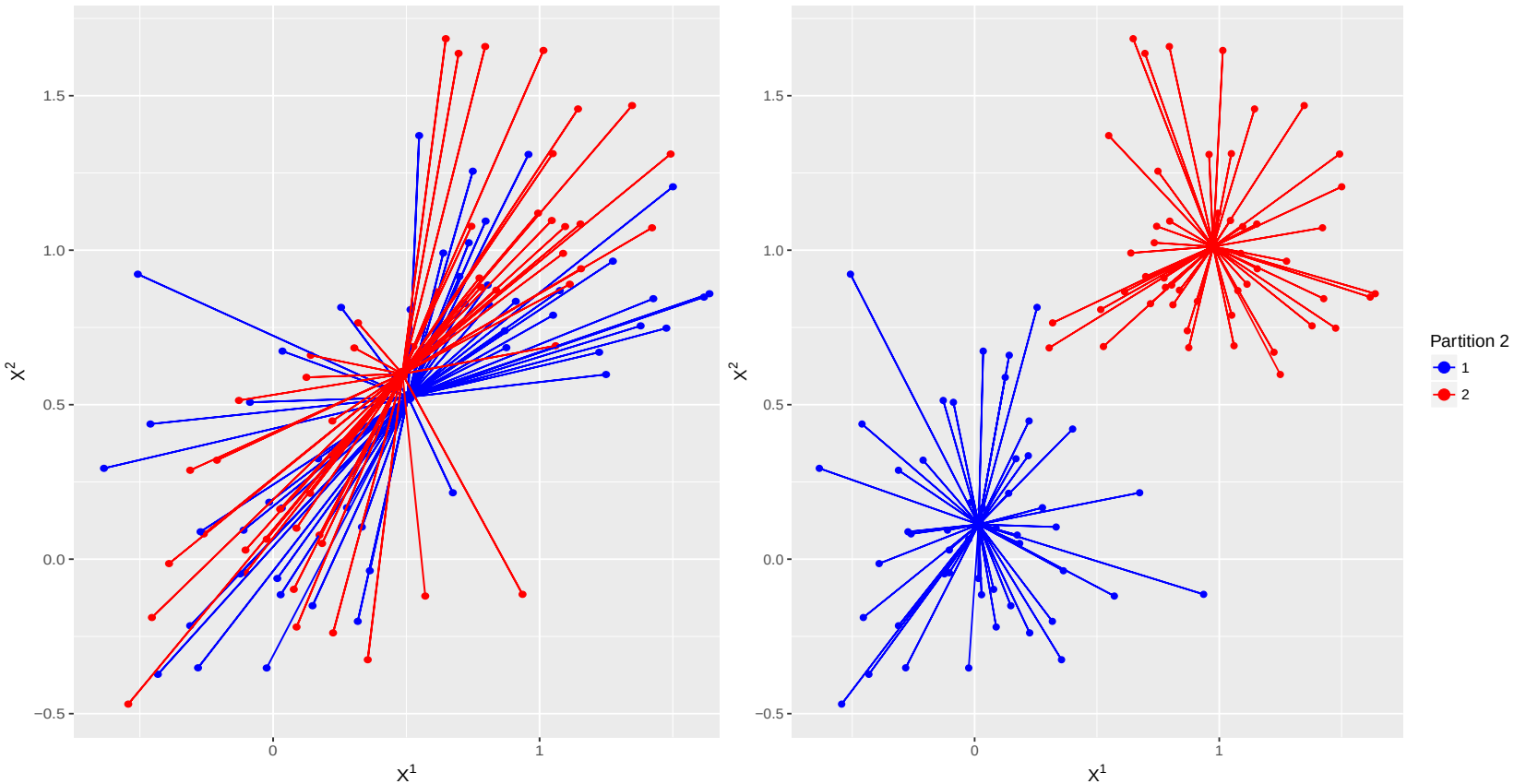
- On peut décrire chaque classe grâce aux variables actives (celles sur lesquelles on a souhaité différencier les classes), et grâce à toute autre variable supplémentaire.
- Lorsque les variables sont quantitatives, on peut comparer leurs moyennes sur les différentes classes.
  Lorsque les variables 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



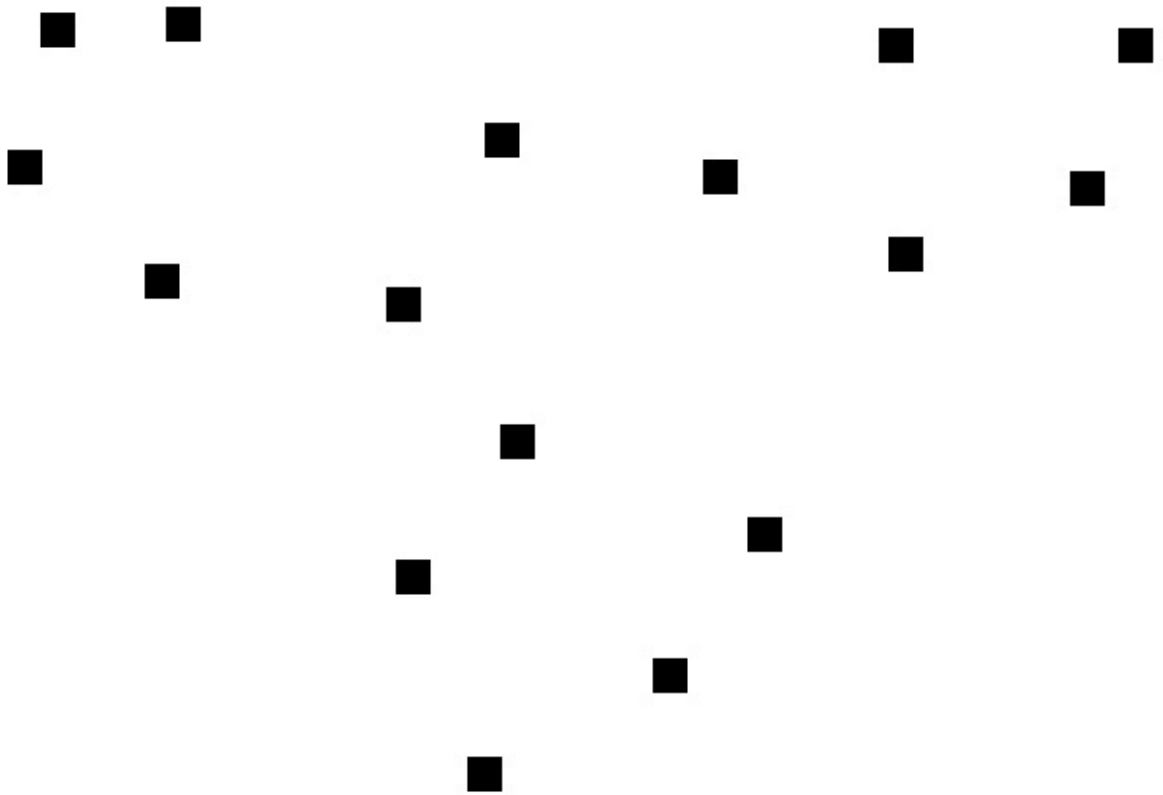
Quelle partition choisissez-vous ?

Qualité du clustering

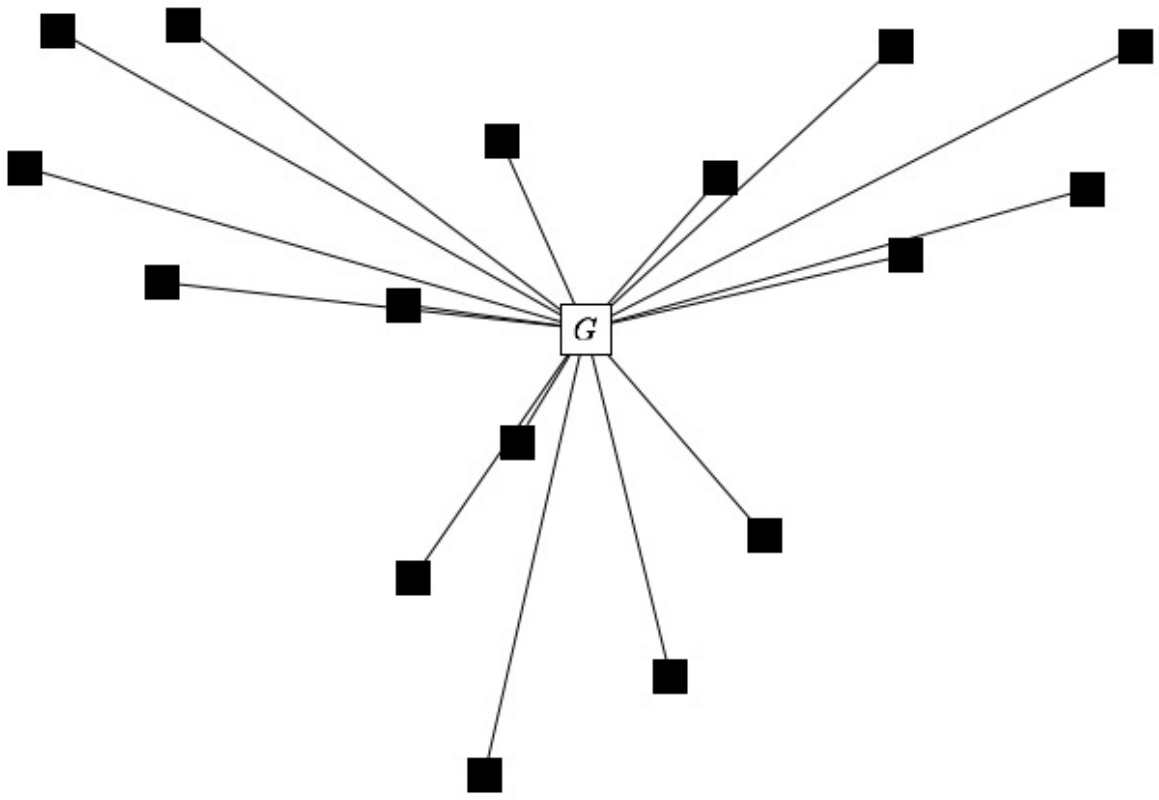


Inertie (notion de variablité) intra classe et inter classes

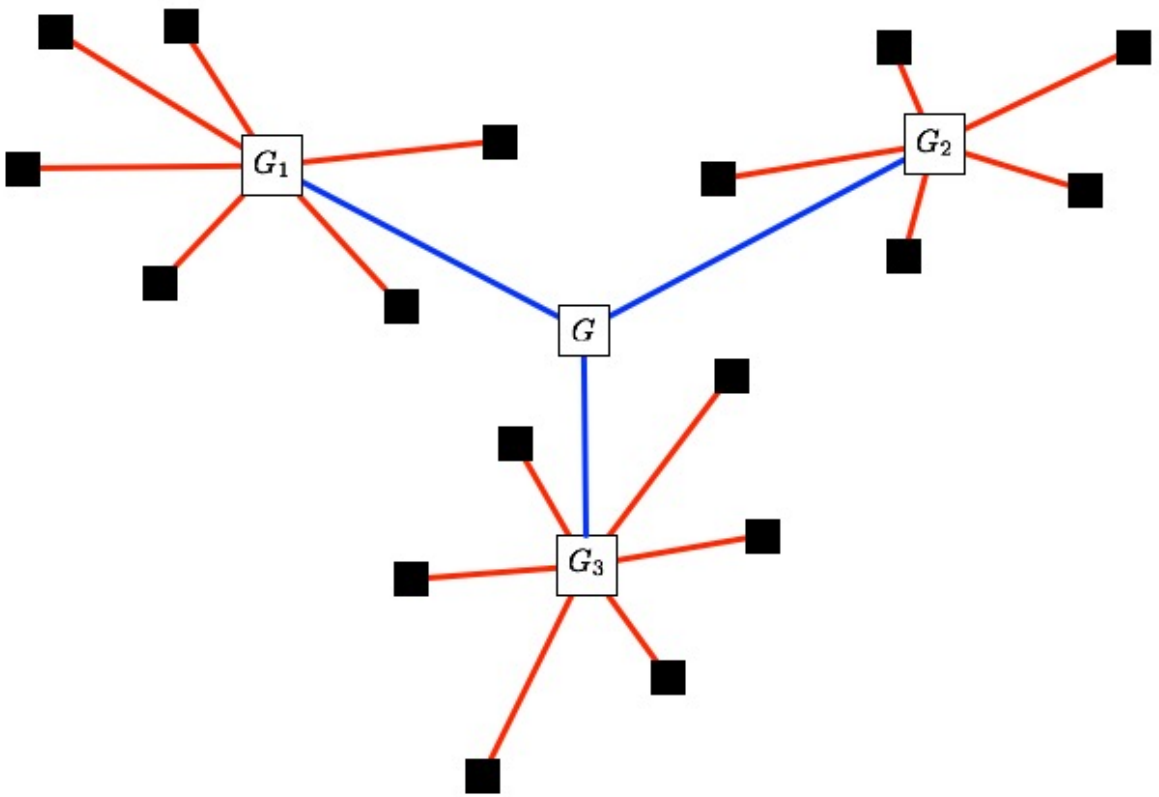
A partir de ce nuage de points :



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 :

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

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

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

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

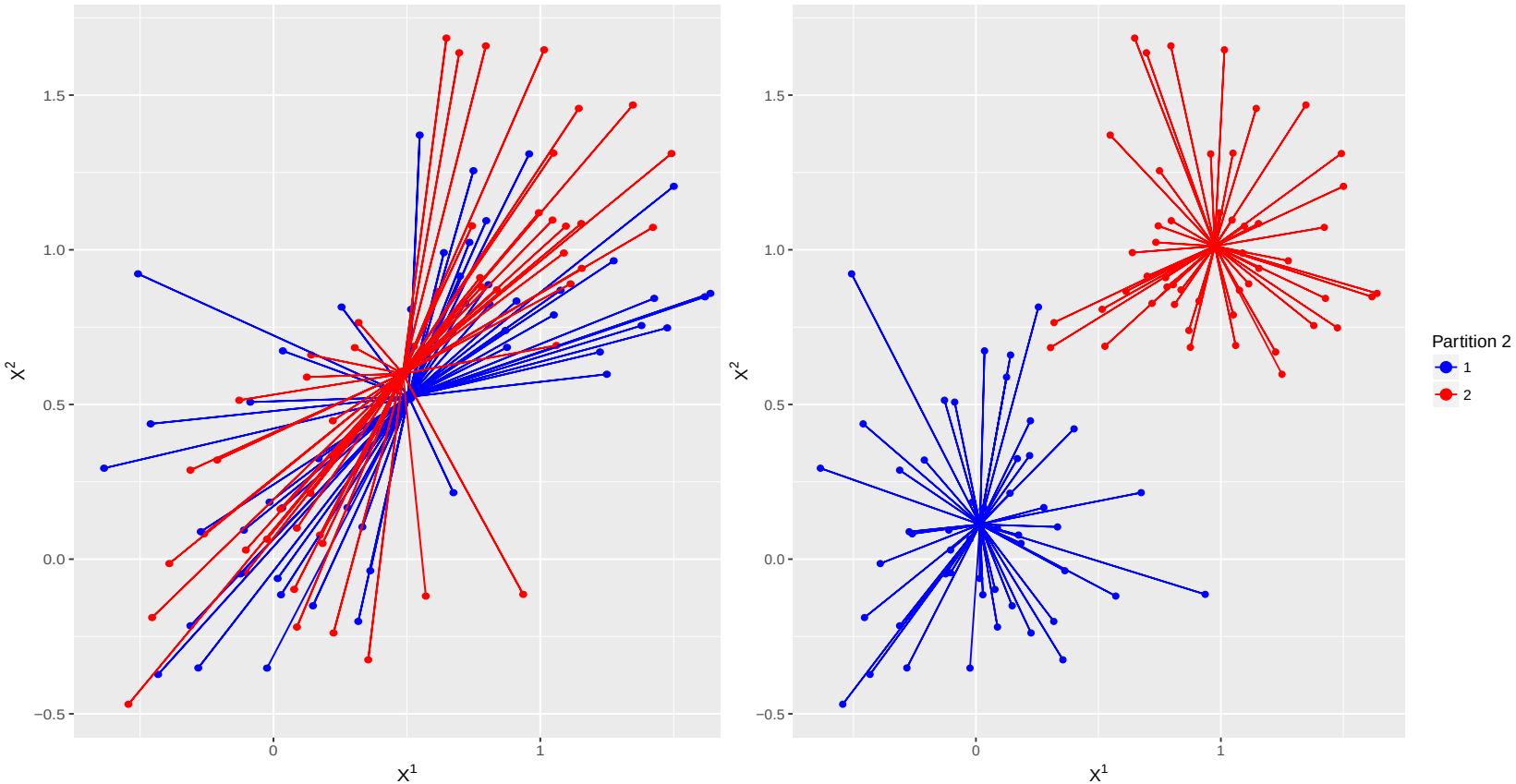
A K fixé, l'idéal est de minimiser l'inertie intra-classes (i.e rendre les classes les plus homogènes possible), soit encore maximiser l'inertie inter-classes (i.e séparer le plus possible les classes).

La qualité d'un clustering peut être évaluée par :

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

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 variablité) 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

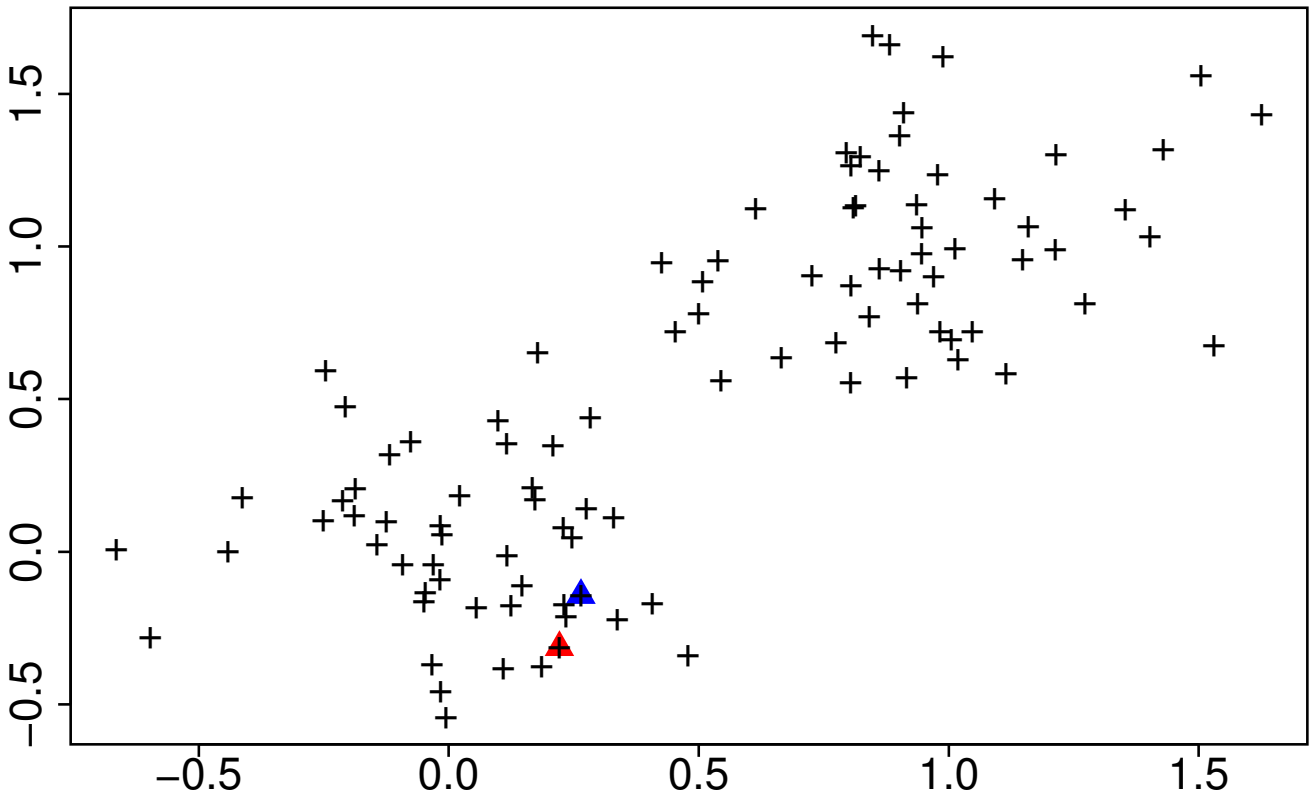
- Soient des observations $x_1, \ldots, x_i, \ldots, x_n \quad x_i \in \mathbb{R}^d$
  $\rightarrow$ p variables quantitatives continues
- Soit $\mathcal{C} = (C_1, C_2, \ldots, C_K)$ une partition de $\{1, 2, \ldots, n\}$

On recherche la partition qui réalise le minimum du critère

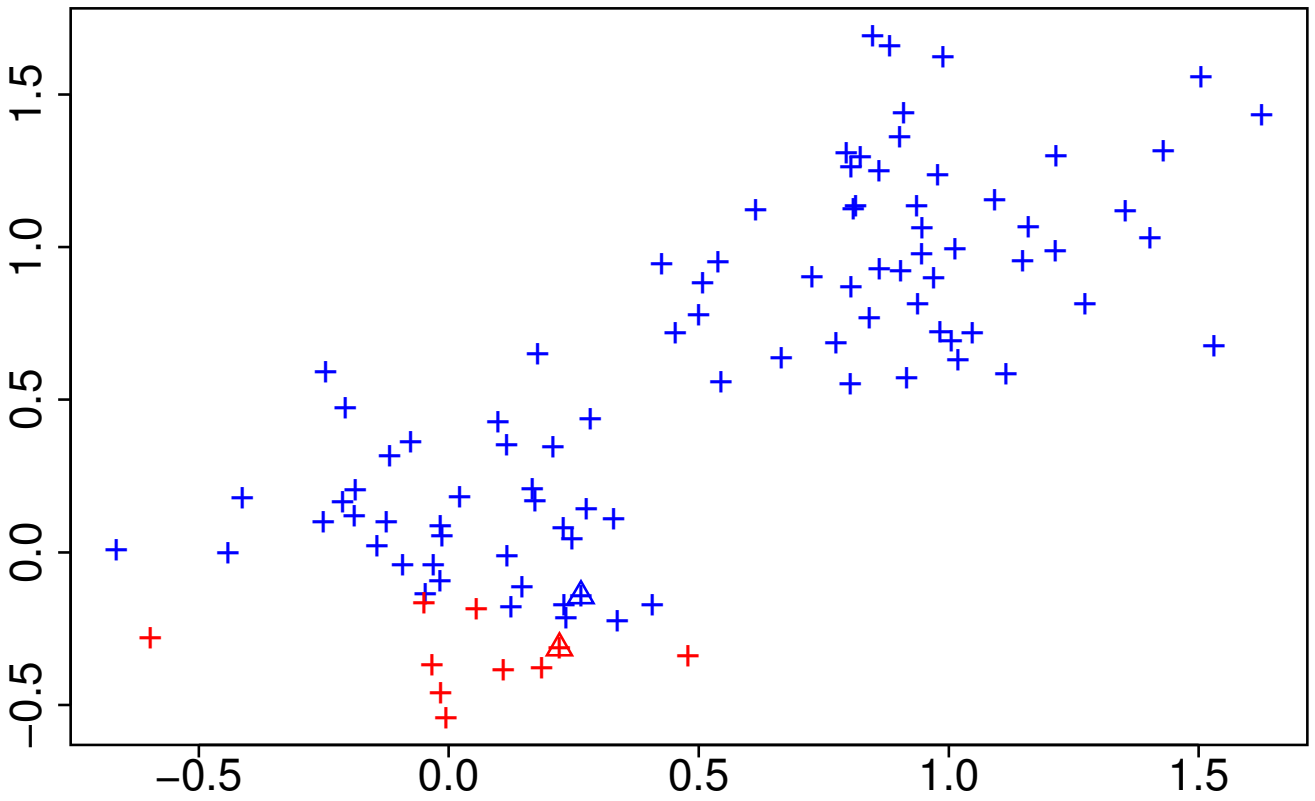
$$g_n(\mathcal{C}) \quad = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}_{C_k}\|^2 \tag{1}$$

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

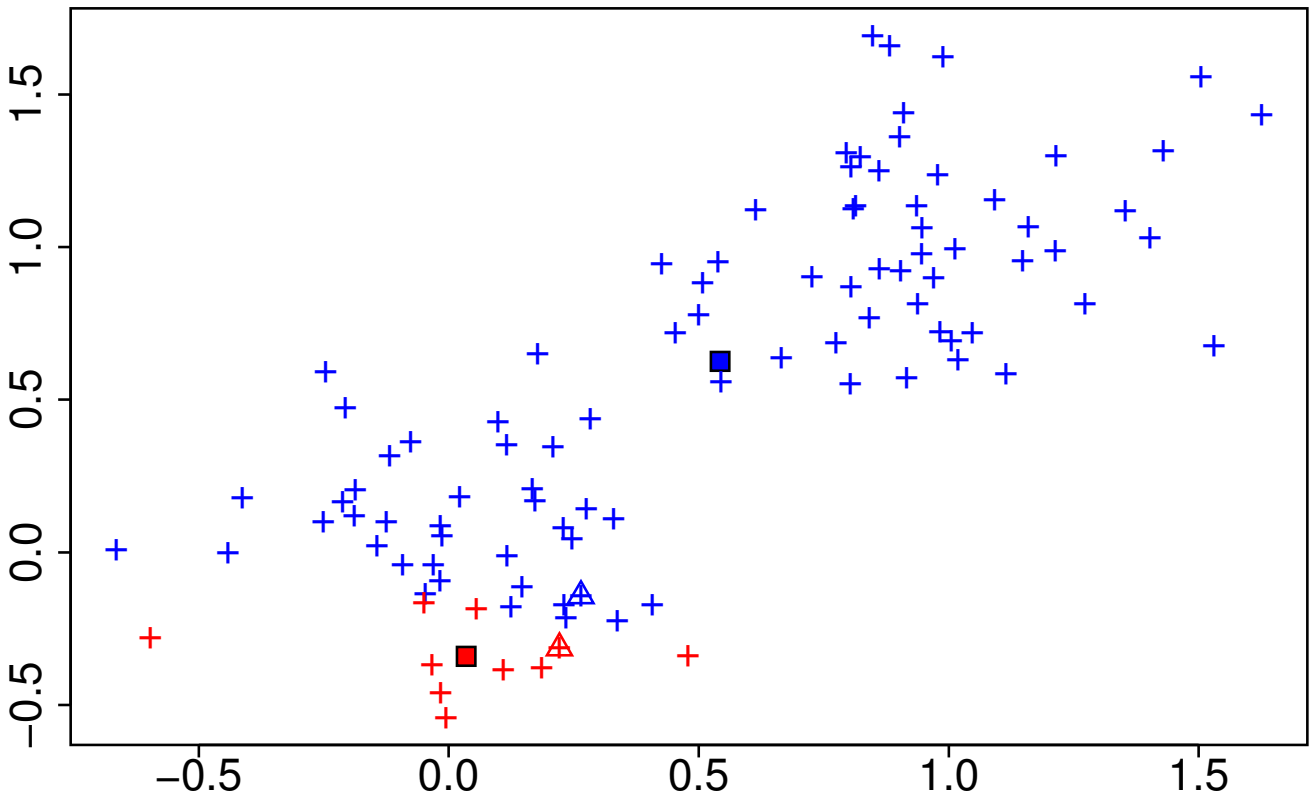
Illustration graphique



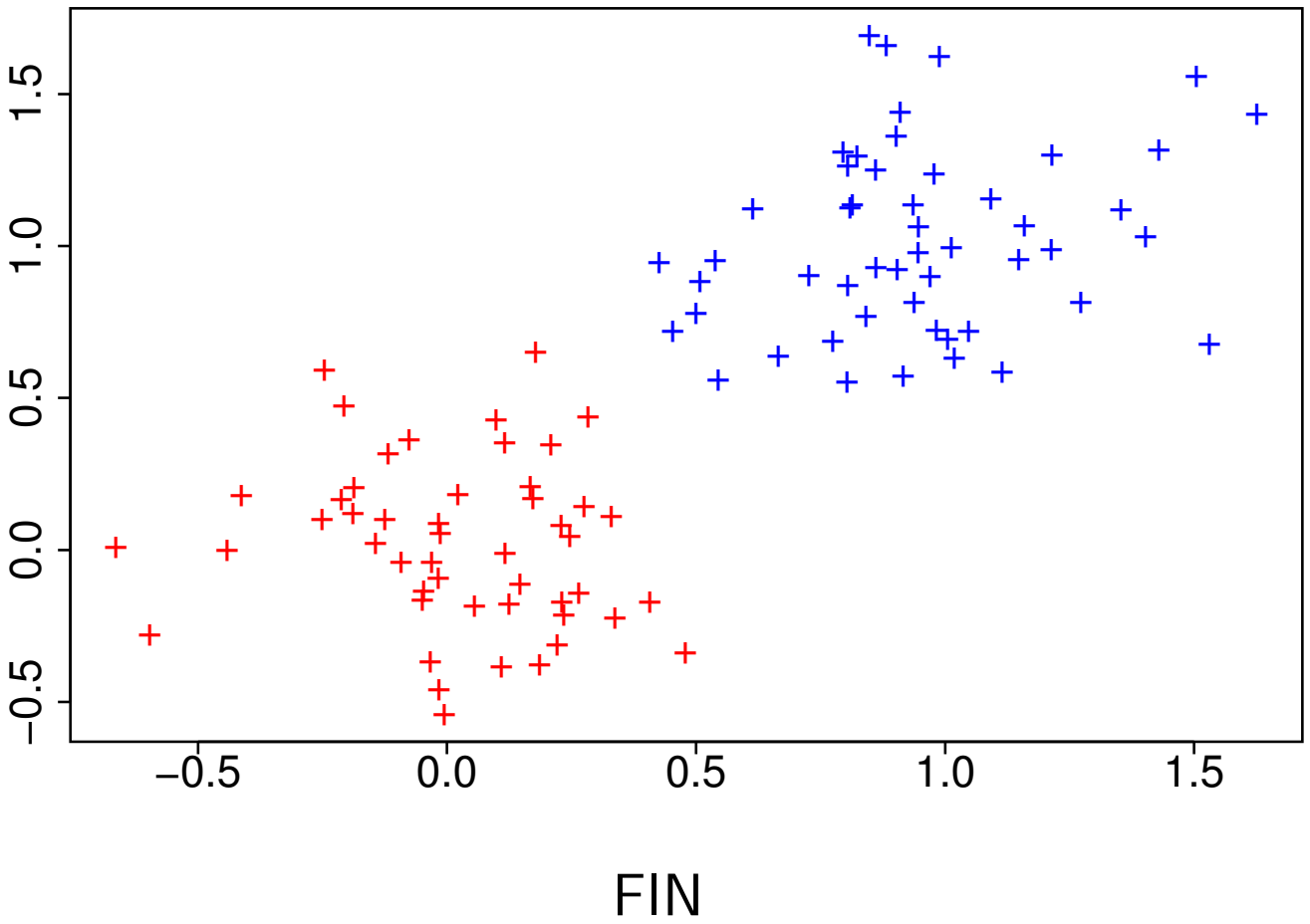
On choisit des centres initiaux



On réaffecte.



On recalcule les centres...


1.5
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
1.5
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 $\mathcal{Z} = \{z_1, \ldots, z_K\}$, $z_k \in \mathbb{R}^d$ ces représentants.

Méthode des k-means

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

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

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

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

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

$$\widehat{\mathcal{C}} = \{\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_K\}$$
$$\hat{C}_k = \{i \in \{1, \ldots, n\} | \|x_i - z_k\|^2 = \min_j \|x_i - z_j\|^2\}.$$

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 $\widehat{\mathcal{C}}$
- des meilleurs représentants $\widehat{\mathcal{Z}}$

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

$$\forall i \in \{1, \cdots, n\}, \quad : K(i) = argmin_{c_K \in C} d(x_i, c_K)$$

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.

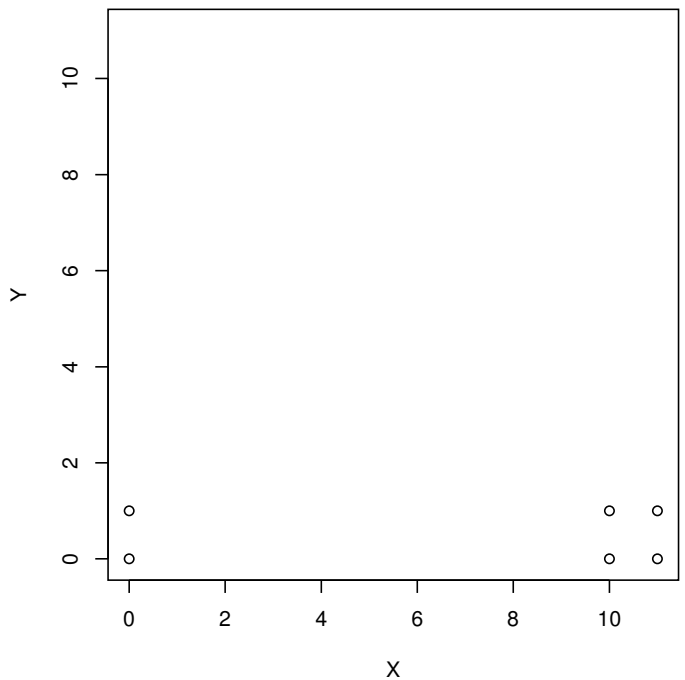
A partir des K centres fournis

1. Initialisation
   - affectation de tous les individus au centre le plus proche

$$\forall i \in \{1, \cdots, n\}, \quad : K(i) = argmin_{c_K \in C} d(x_i, c_K)$$

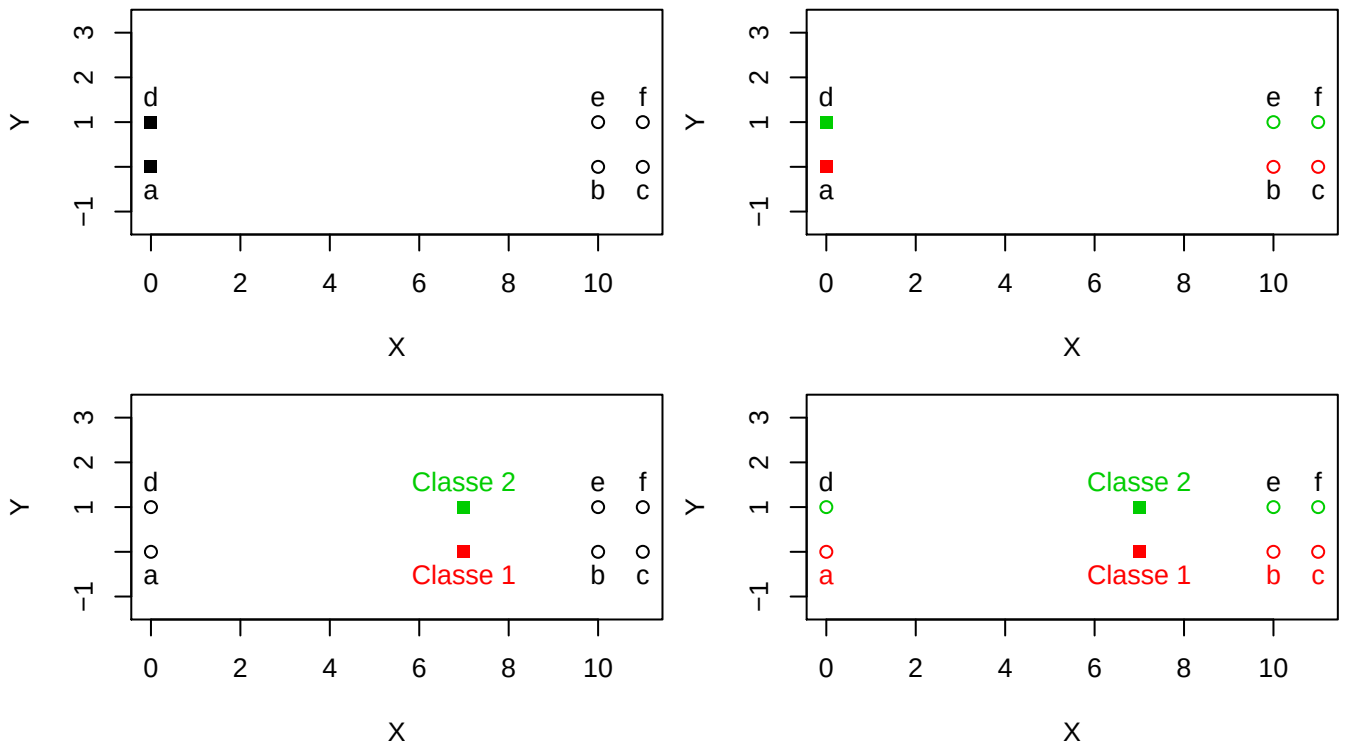
   - calcul des centres par la moyenne
2. Faire pour tout $1 \leq i \leq n$:
   - recherche du centre le plus proche de x_i
   - calcul des nouveaux centres par la moyenne si x_i change de groupe (réactualisation d'au plus 2 centres)
3. Retour étape 2 tant qu'il y a changement

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

$$\forall i \in \{1, \cdots, n\}, \quad : K(i) = argmin_{c_K \in C} d(x_i, c_K)$$

  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 C1	Configuration C2	Echange testé	Inertie intra-classe	Echange ?
{a,b,c}	{d,e,f}		148	
{a,b,c}	{d,e,f}	a en C2	112	oui
{b,c}	{a,d,e,f}	b en C2	130	non
{b,c}	{a,d,e,f}	c en C2	138.4	non
{b,c}	{a,d,e,f}	d en C1	149.3	non
{b,c}	{a,d,e,f}	e en C1	82.7	oui
{b,c,e}	{a,d,f}	f en C1	2.5	oui
{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 $O(n)$.

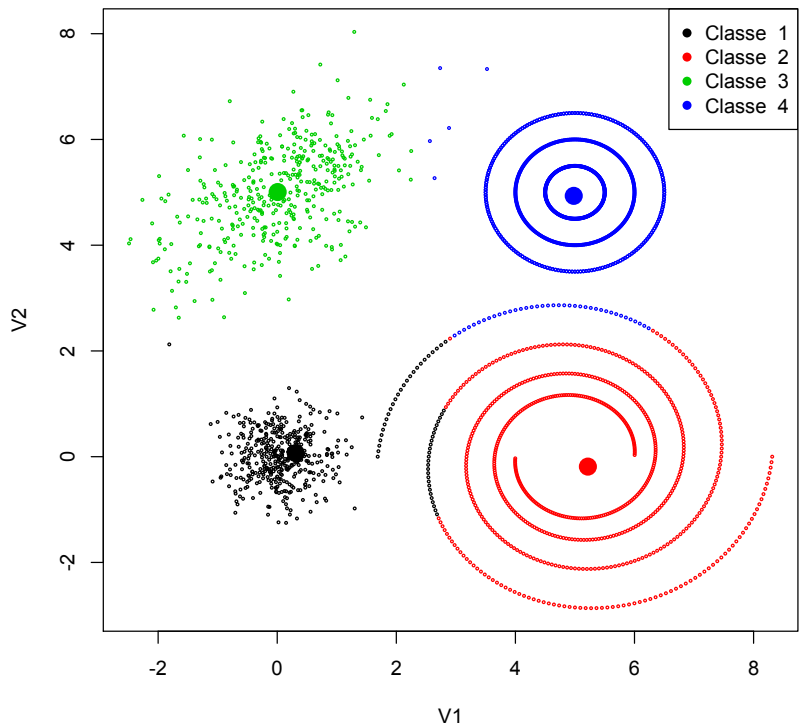
Problèmes :

- 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)
```

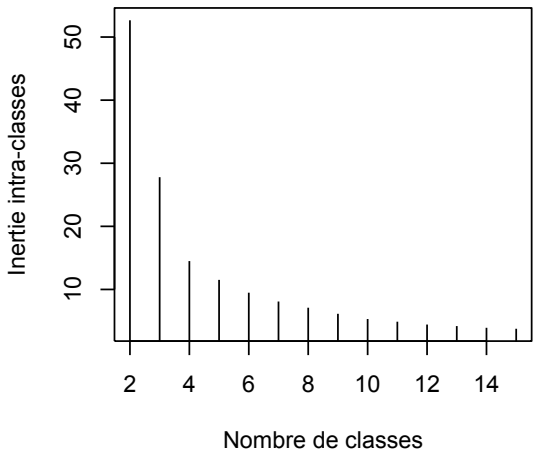
Le coin R

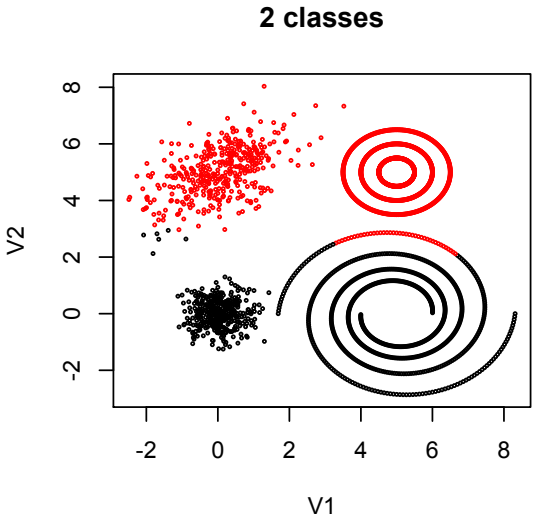


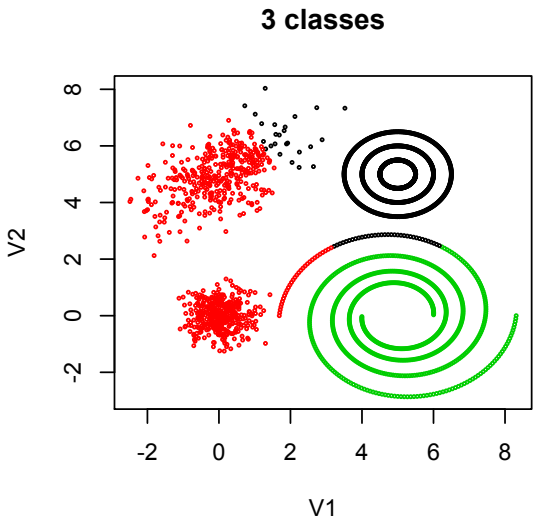
Le coin R

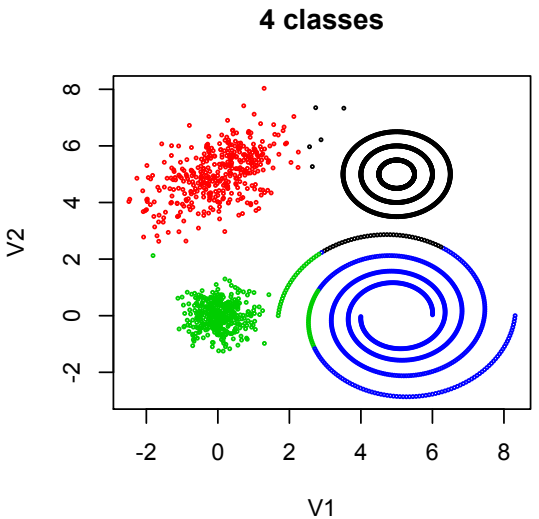
```
> k <- 2:15
> part <- k
> for(i in k){
>  kmk=kmeans(don,centers=i,nstart=20)
>  part[i-1]=sum(kmk$withinss)/kmk$totss*100
> }
> par(mfrow=c(2,2))
> plot(k,part,type="h", xlab="Nombre de classes",
+ ylab="Inertie intra-classes")
```

Le coin R


Inertie intra-classes
10
20
30
40
50
2
4
6
8
10
12
14
Nombre de classes


2 classes
V2
-2
0
2
4
6
8
V1


3 classes
V2
-2
0
2
4
6
8
V1


4 classes
V2
-2
0
2
4
6
8
V1

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

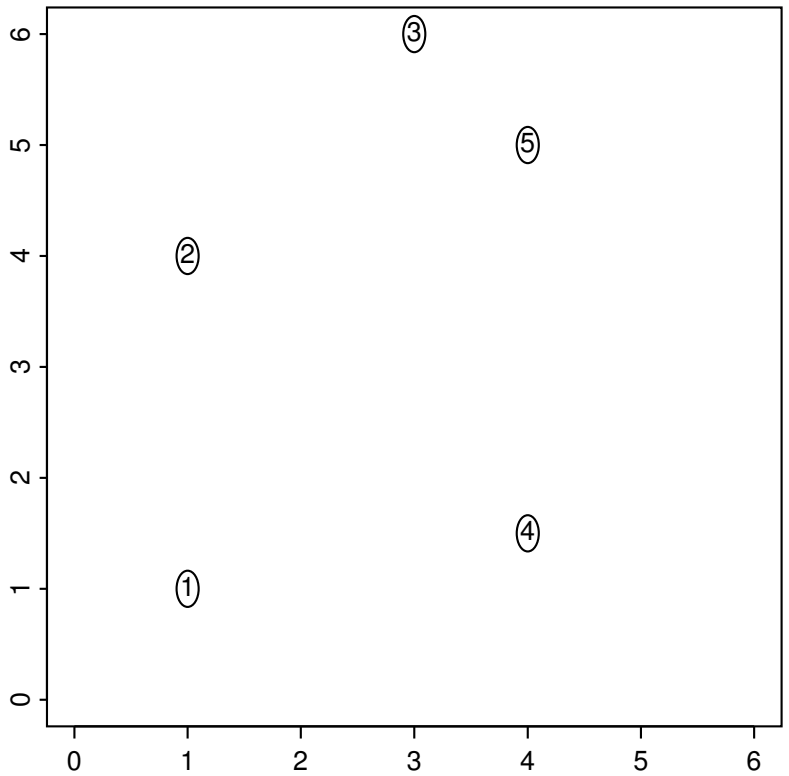
Fonctions **pam** et **clara** du package **cluster**.

Exemple introductif

Considérons l'exemple suivant

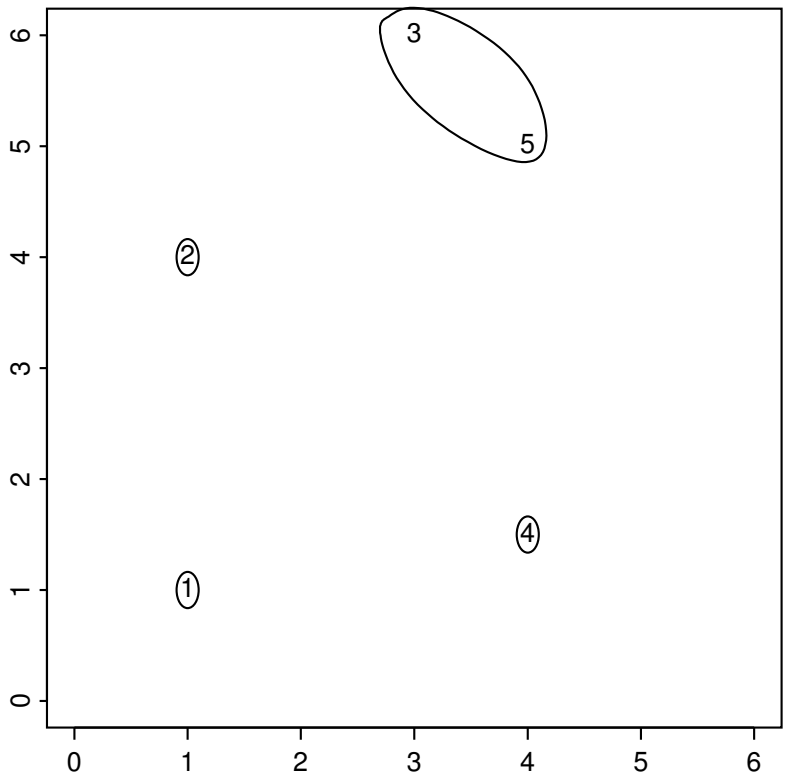
ind	X_1	X_2
1	1	1
2	1	4
3	3	6
4	4	1.5
5	4	5

Exemple suite : distance euclidienne



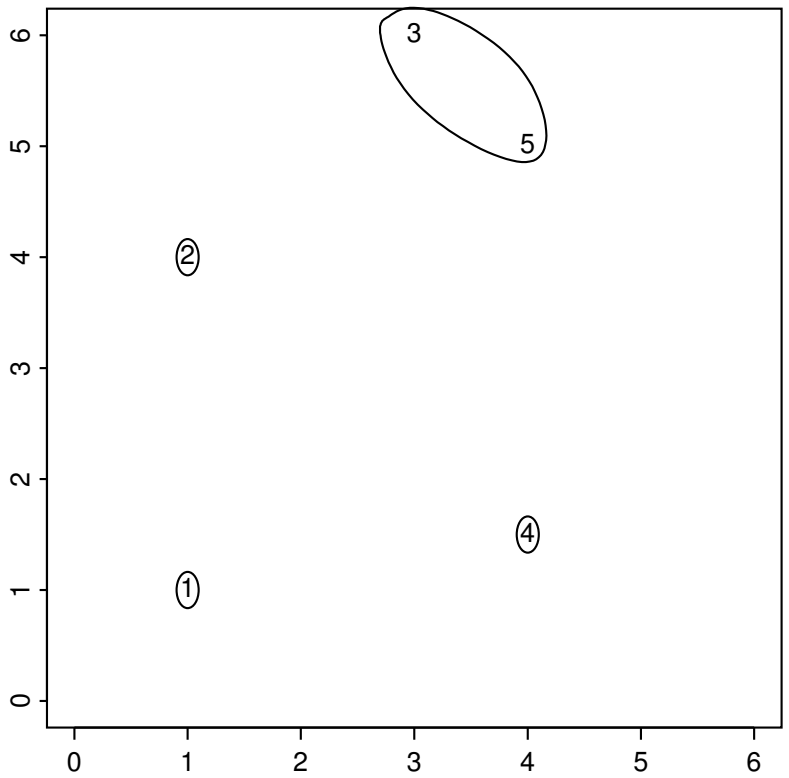
Exemple suite : distance euclidienne

On aggrège les deux plus proches



Exemple suite : distance euclidienne

On aggrège les deux plus proches

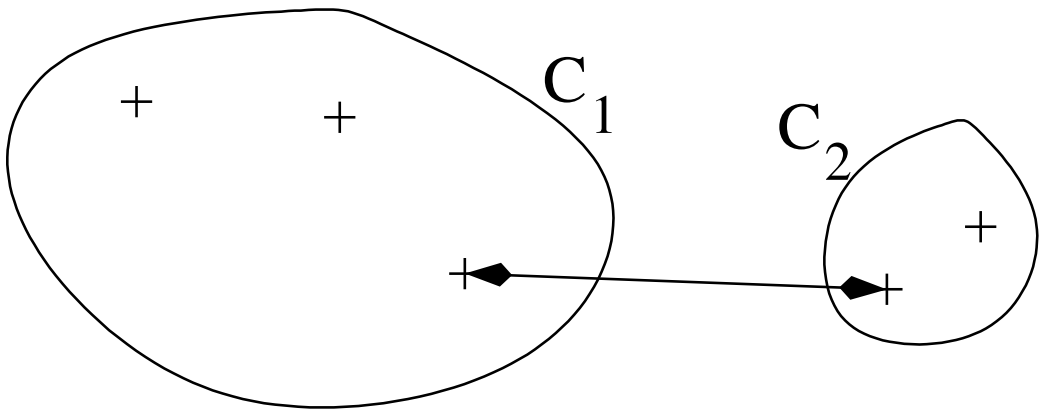


- Définir une dissimilarité ou dissemblance entre ensembles : indice d'agrégation

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 :

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

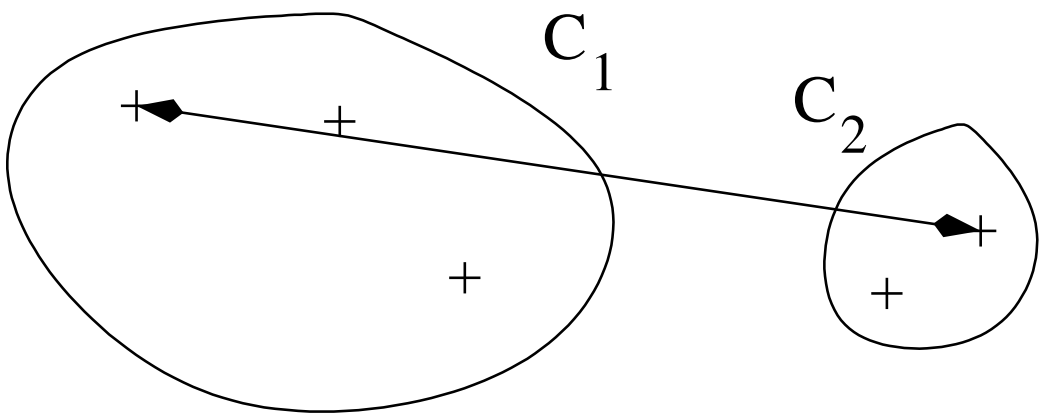


Groupes sont en général allongés, on va ainsi de proche en proche selon la philosophie "le voisin de mon voisin est mon voisin".

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 :

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

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'aggrégation est donc

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

Inertie

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

- 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 singletons (chaque objet constitue un groupe) : $\mathcal{C}_1^{(1)}, \ldots, \mathcal{C}_n^{(1)}$.

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

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

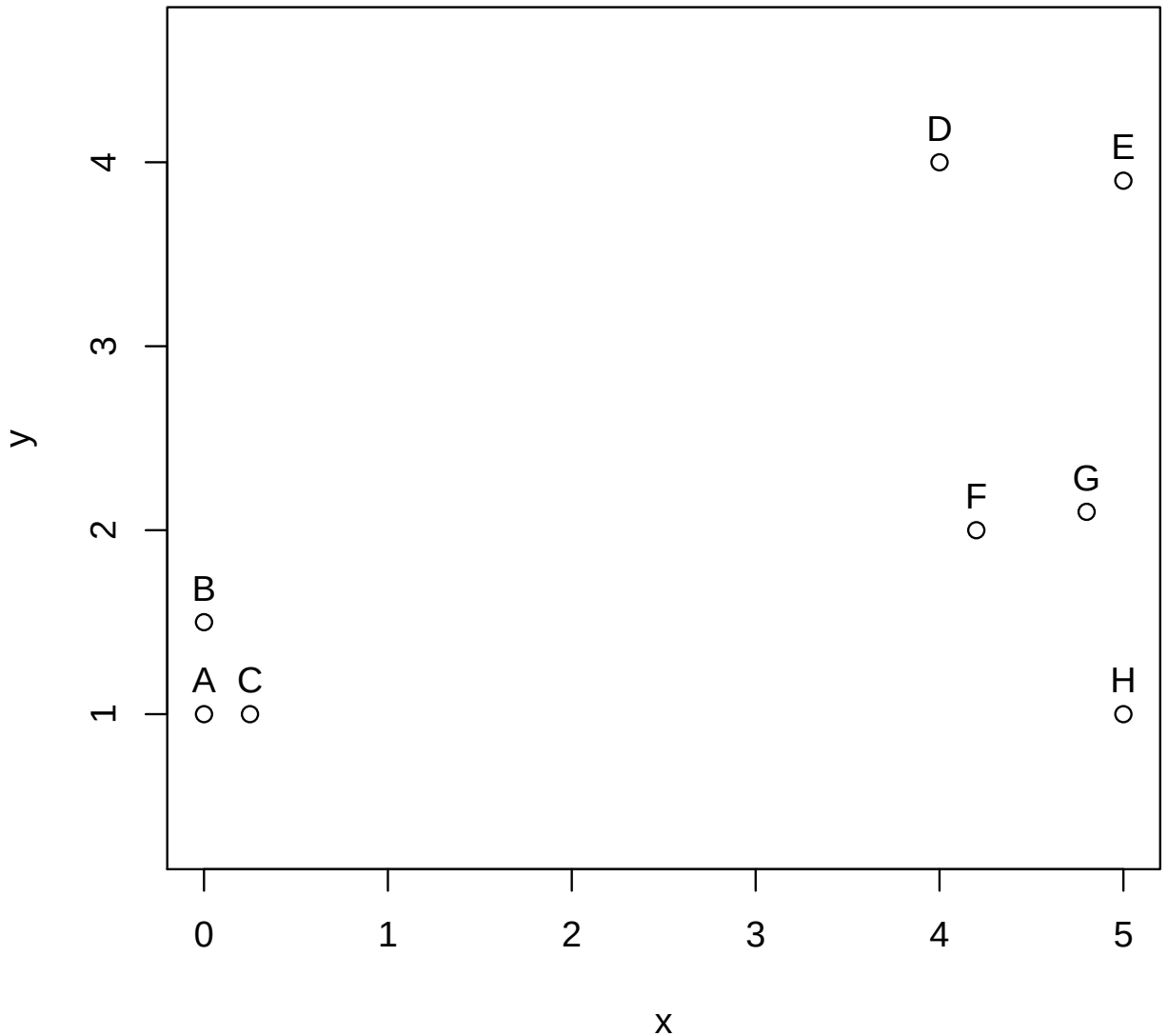
Exemple

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

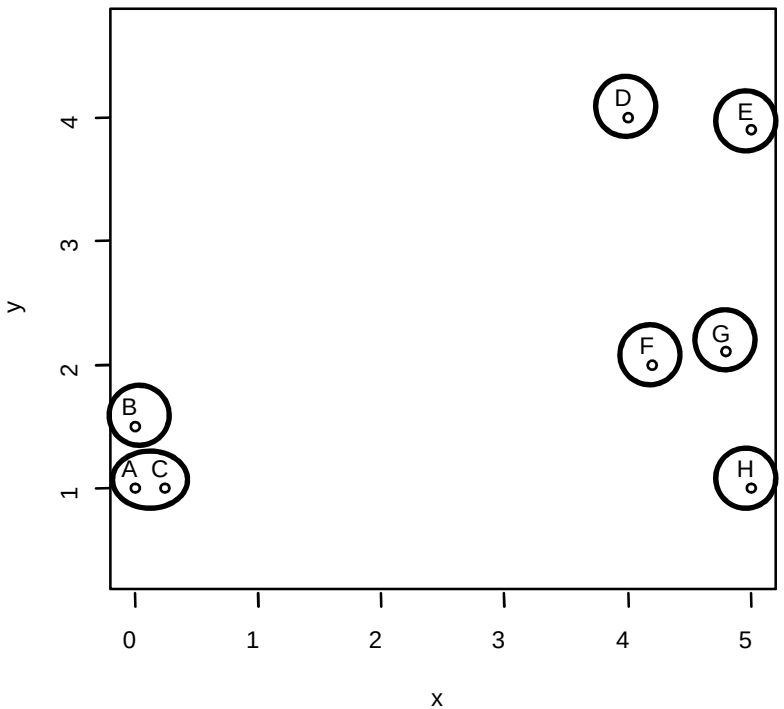
$$\Delta(G_1, G_2) = \min_{\omega_i \in G_1; \omega_j \in G_2} d(\omega_i, \omega_j).$$

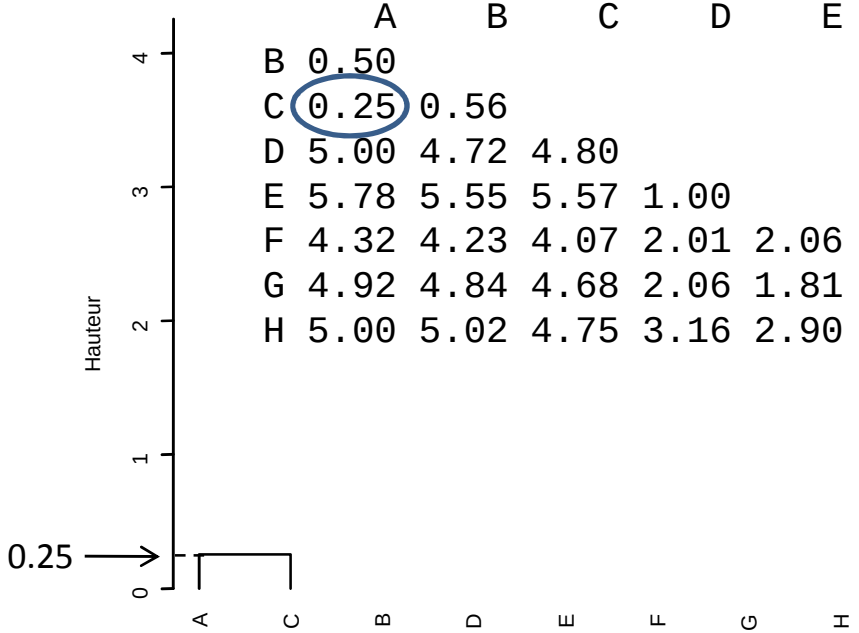
	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

Exemple graphiquement



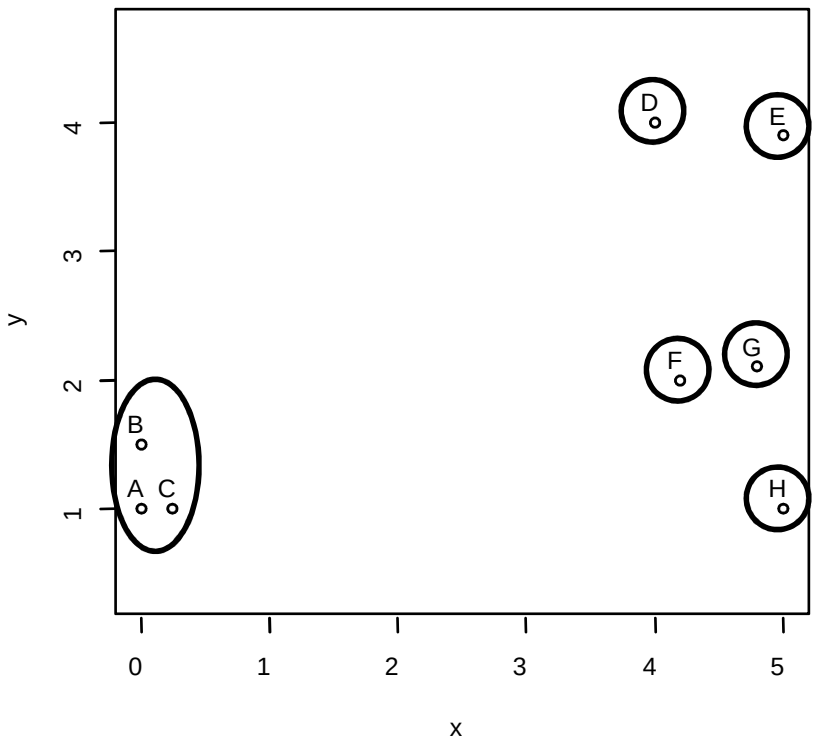
Etape 2

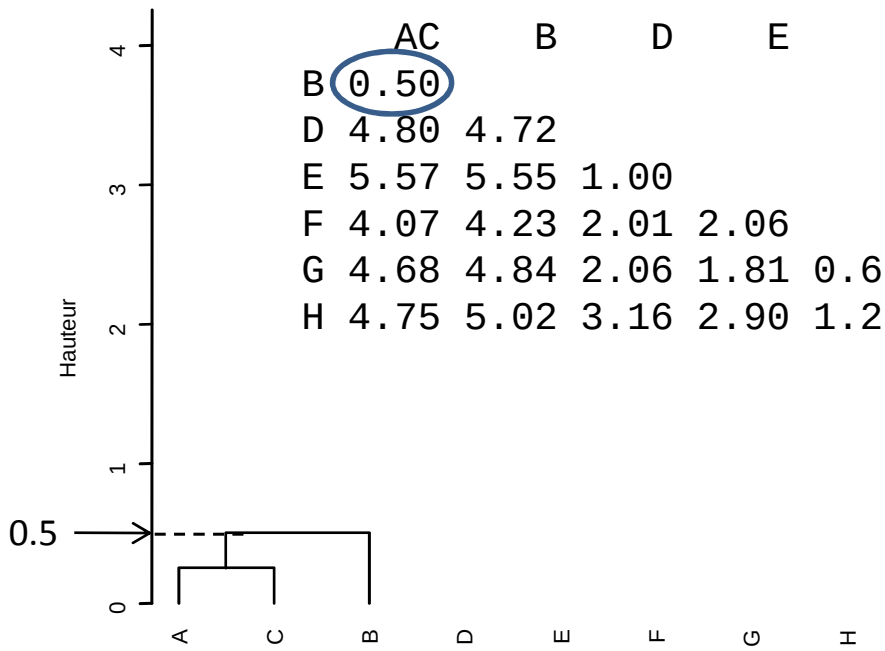




	A	B	C	D	E	F
B	0.50					
C	0.25	0.56				
D	5.00	4.72	4.80			
E	5.78	5.55	5.57	1.00		
F	4.32	4.23	4.07	2.01	2.06	
G	4.92	4.84	4.68	2.06	1.81	0.61
H	5.00	5.02	4.75	3.16	2.90	1.28 1.1

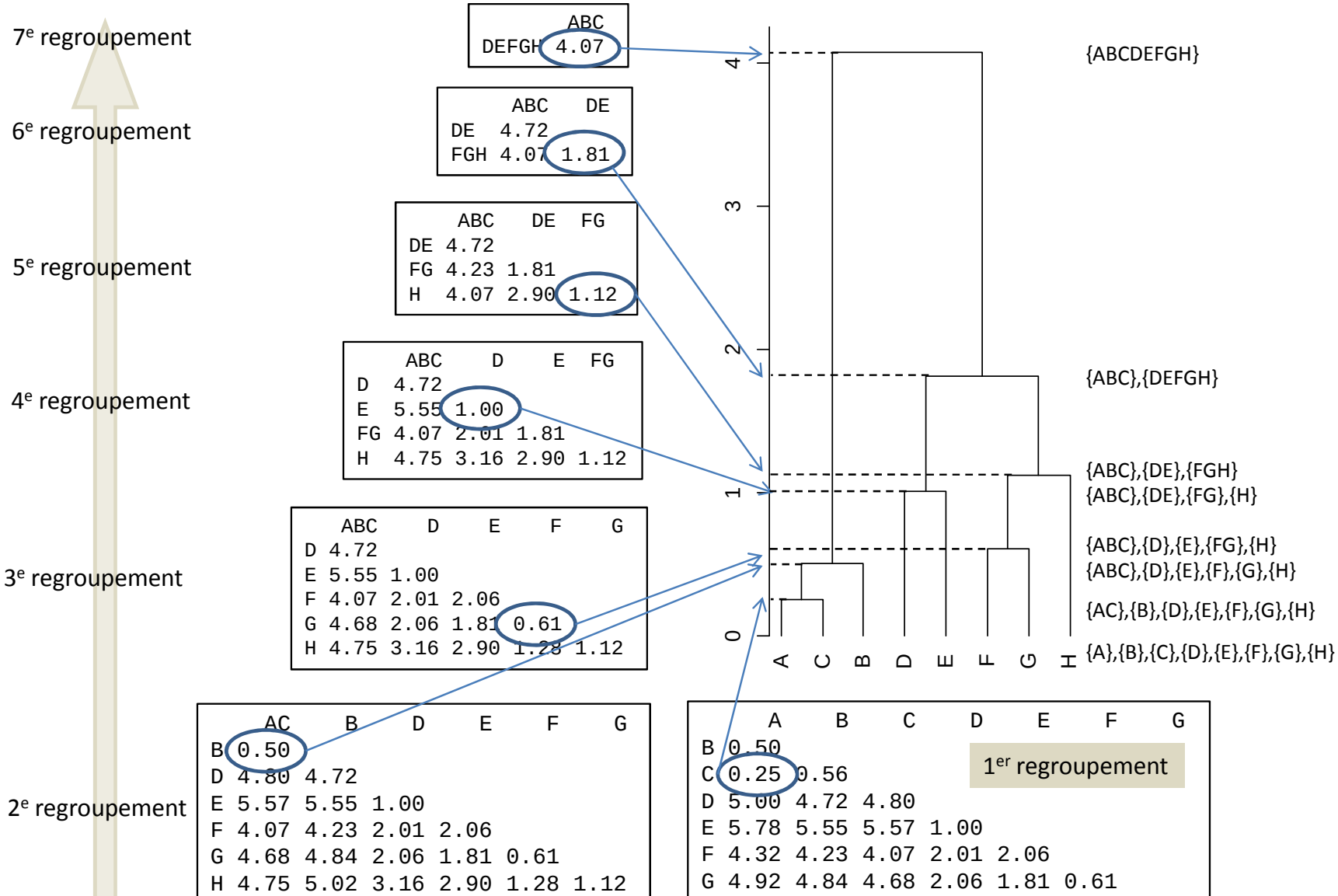
Etape 3





	AC	B	D	E	F
B	0.50				
D	4.80	4.72			
E	5.57	5.55	1.00		
F	4.07	4.23	2.01	2.06	
G	4.68	4.84	2.06	1.81	0.61
H	4.75	5.02	3.16	2.90	1.28

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

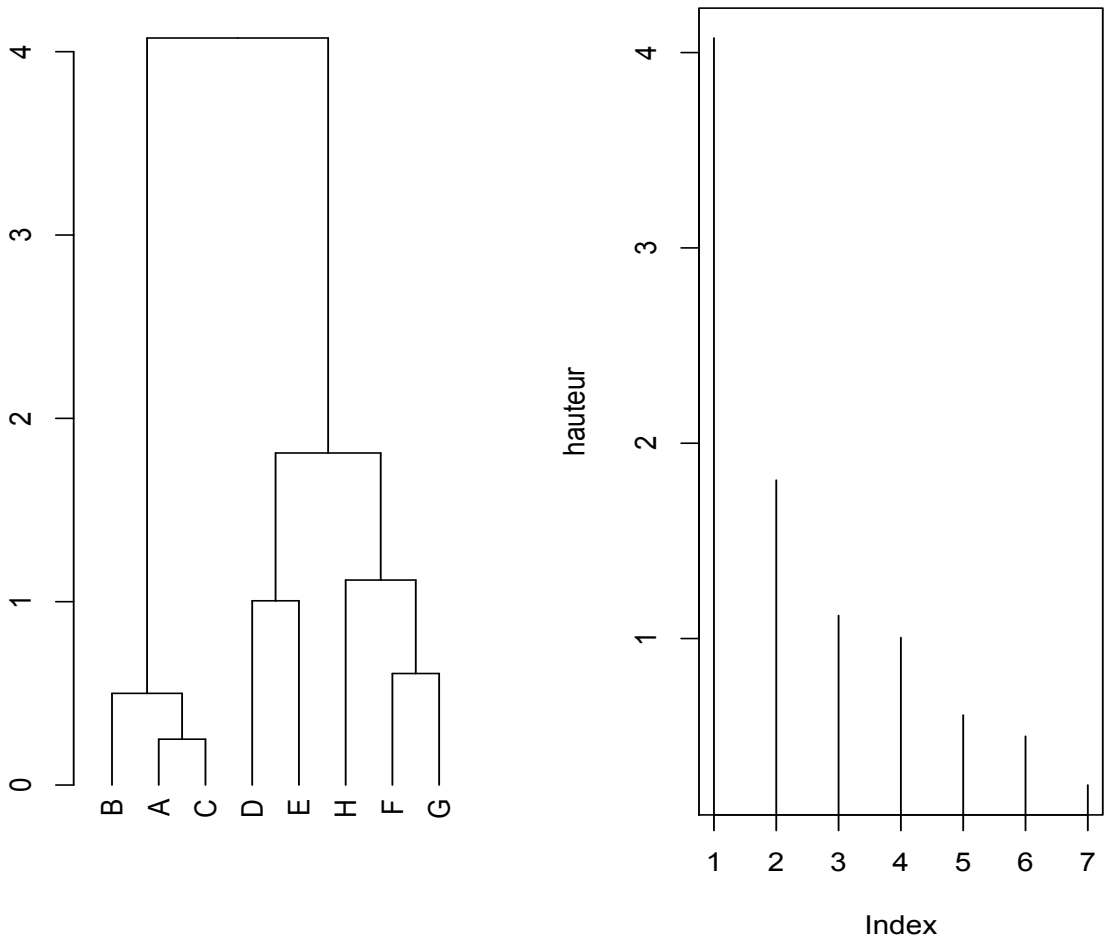
Indice	Non-inversion	Convexité	Réplication	Monotonie
Saut minimum	Non-inversion	Non	Oui	Oui
Saut maximum	Non-inversion	Non	Oui	Oui
Moyenne	Non-inversion	Non	Non	Non
Ward	Non-inversion	Oui	Non	Non

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.

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 :

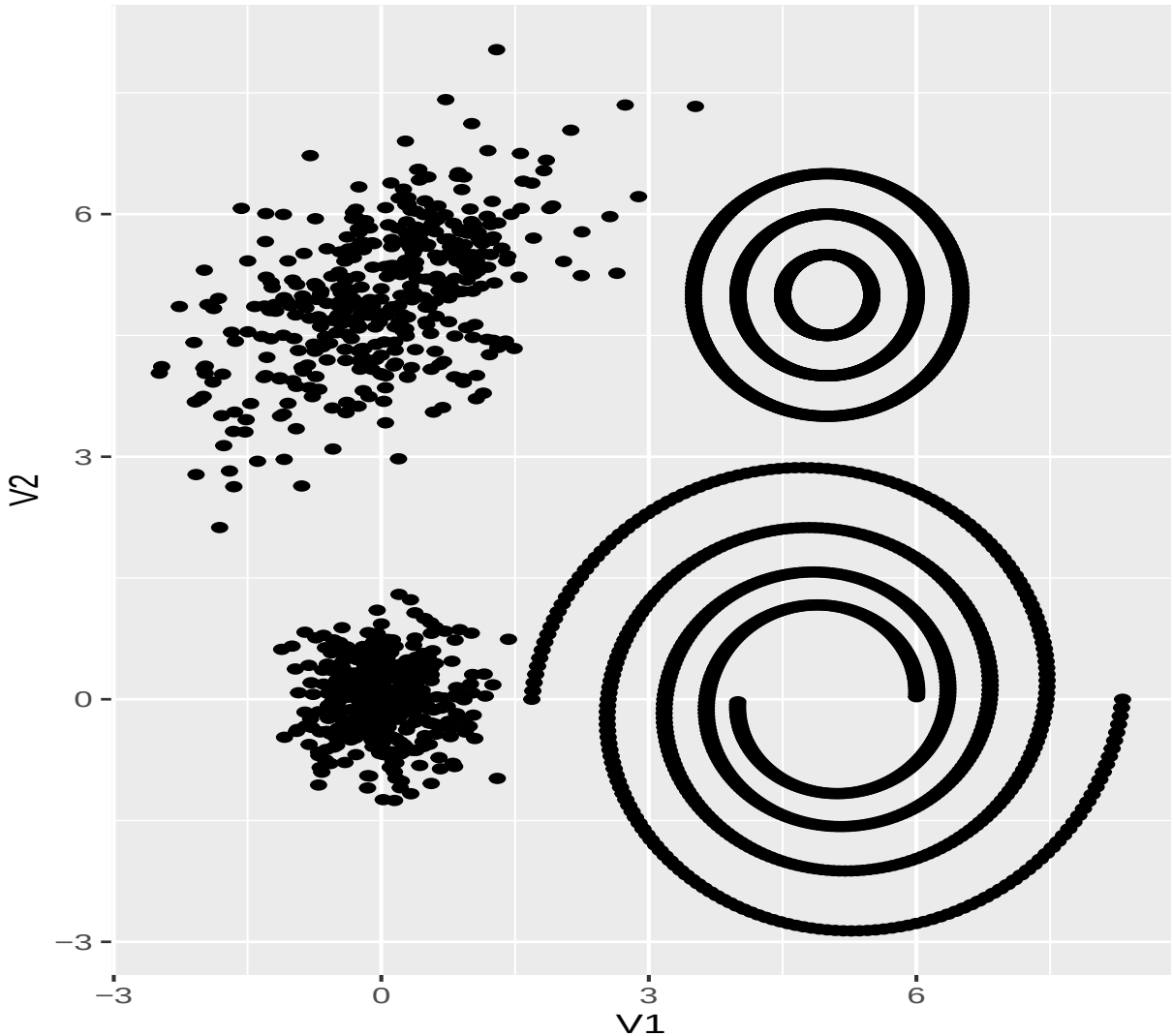
- La complexité algorithmique est en $O(n^2 \ln n)$: la CAH devient chronophage si le nombre d'individus est important.

Le coin R

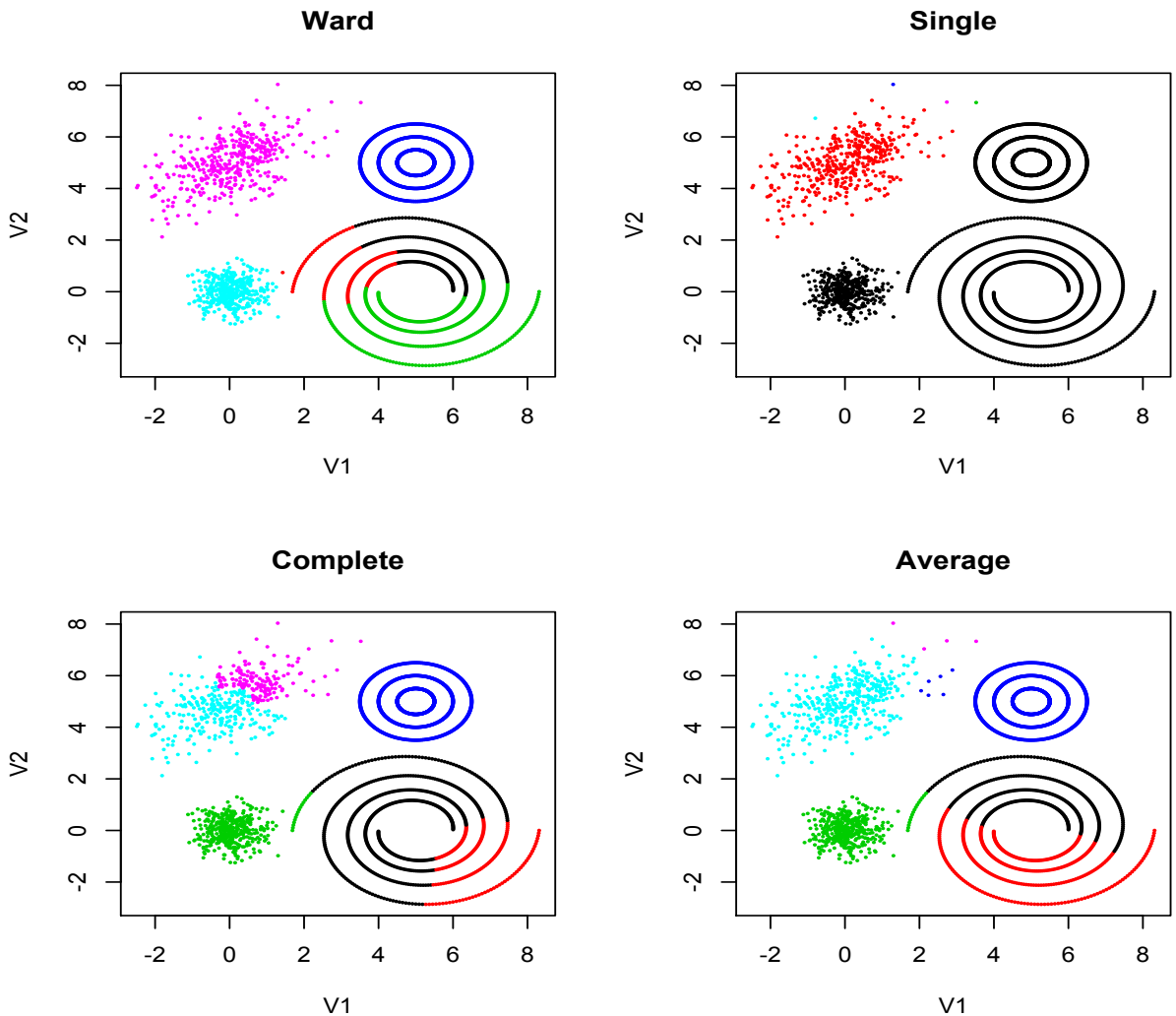
Vous pouvez retrouver les résultats présentés en utilisant les fonctions **hclust** et **cutree** de la base, ou **hclust** du package fastcluster.
Essayer les différents indices d'agrégation avec l'argument method

```
> 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)
```

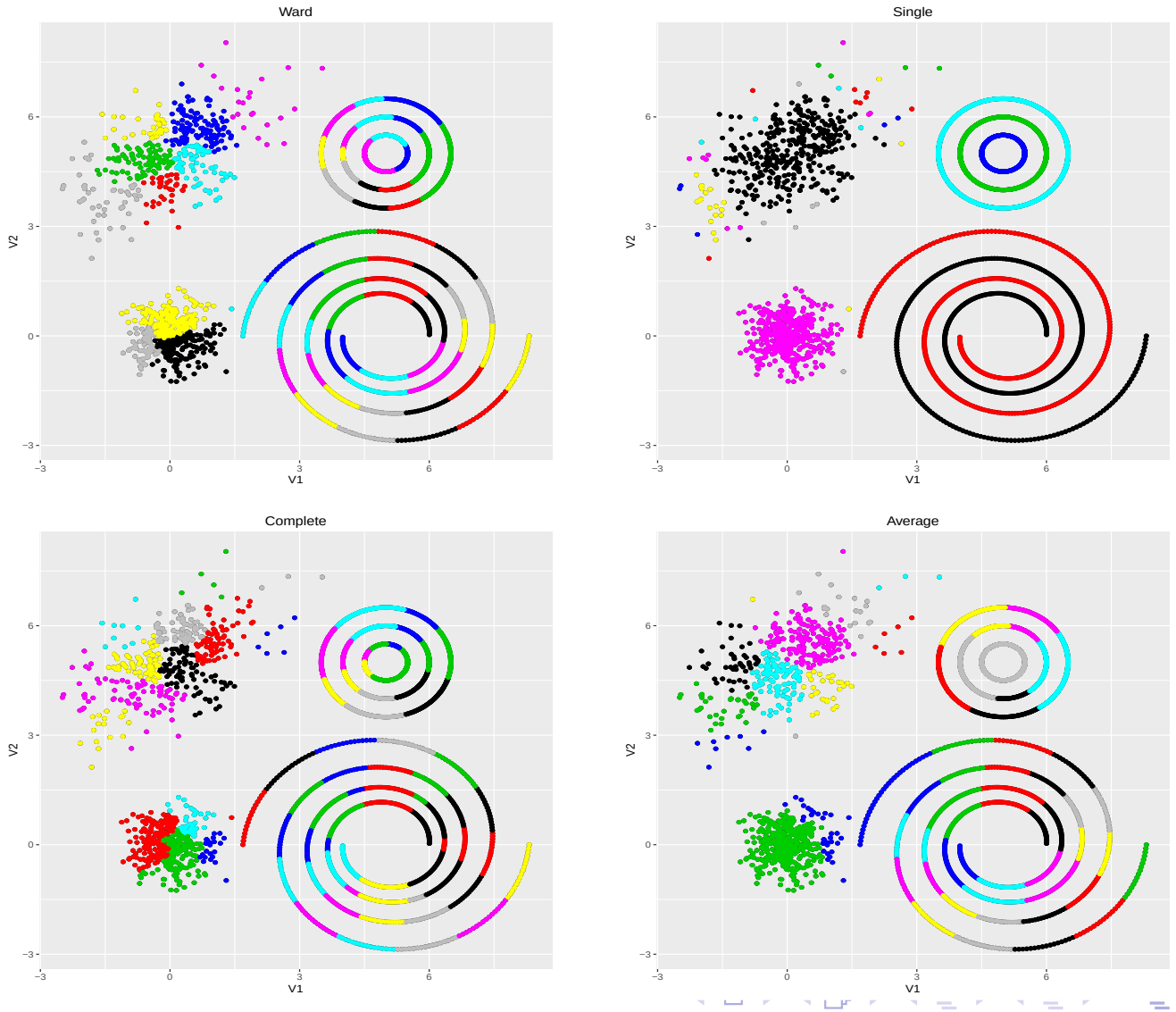
Un exemple jouet



Exemple : CAH 6 groupes



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 (containement 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.