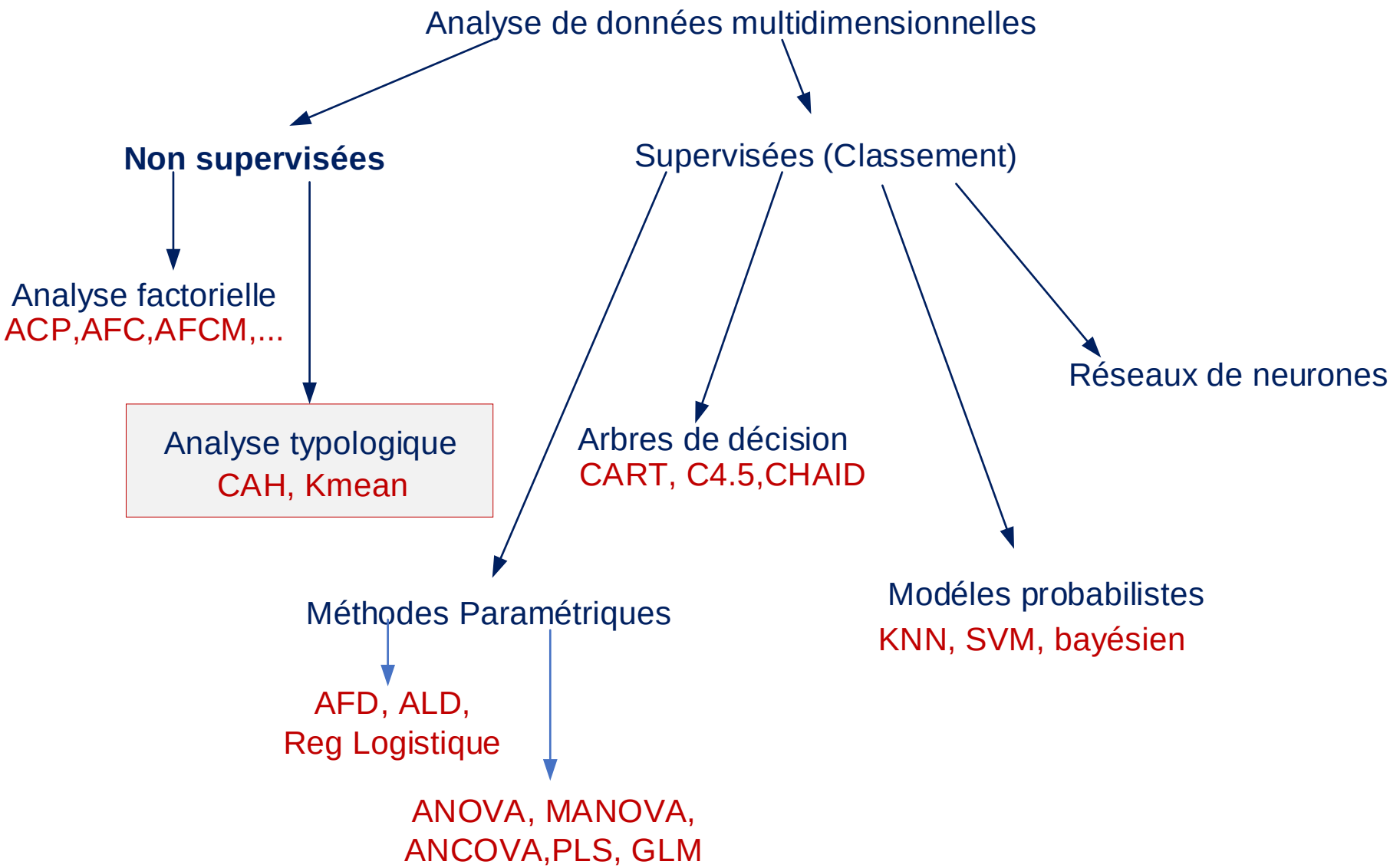
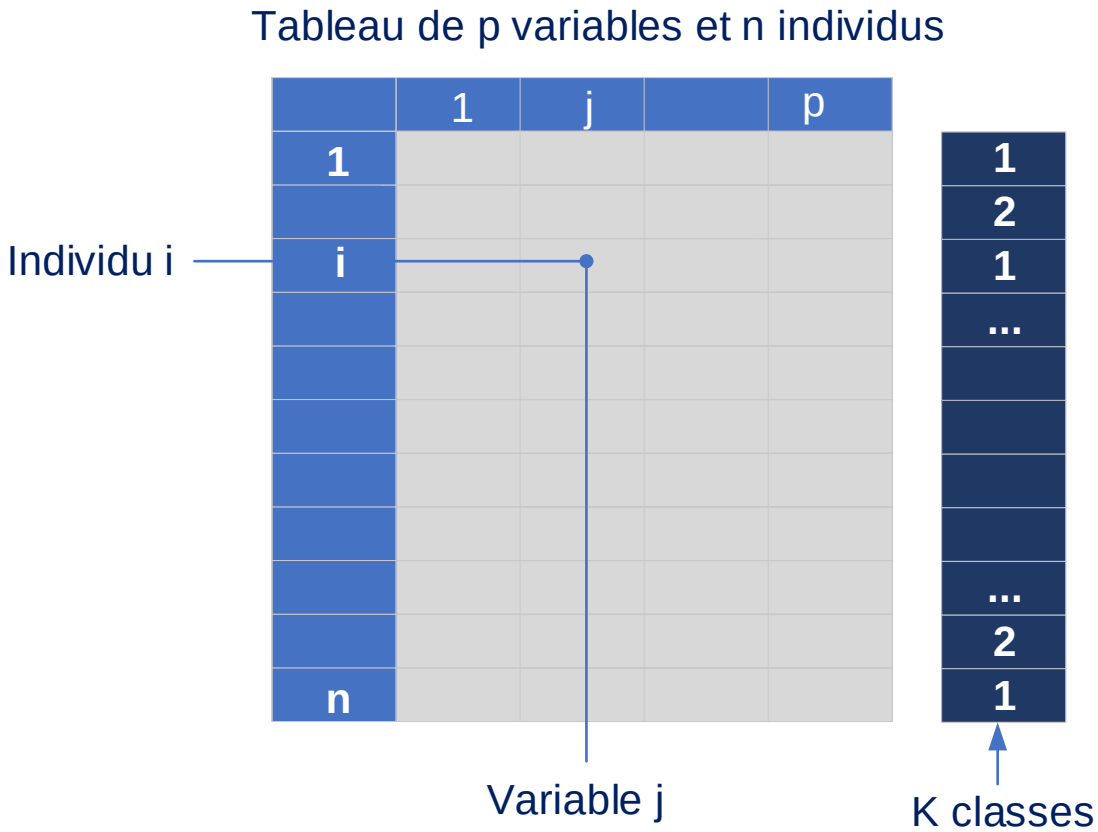
# CLASSIFICATION NON SUPERVISEE

## A. CLASSIFICATION HIERARCHISEE

## B. CENTRES MOBILES


Analyse de données multidimensionnelles
Non supervisées
Supervisées (Classement)
Analyse factorielle
ACP,AFC,AFCM,...
Analyse typologique
CAH, Kmean
Arbres de décision
CART, C4.5,CHAID
Réseaux de neurones
Méthodes Paramétriques
Modéles probabilistes
KNN, SVM, bayésien
AFD, ALD,
Reg Logistique
ANOVA, MANOVA,
ANCOVA,PLS, GLM

- **Objectif** : Effectuer des regroupements des données en k classes (k << n) de manière à rassembler dans chaque classe les individus les plus semblables



- Analyse typologique

→ Méthodes hiérarchiques  Exemple : Classification Automatique Ascendante hiérarchisée (CAH)

- Méthodes d'agrégation successives (regroupement)
- Construction d'une hiérarchie de partitions par regroupements successifs d'individus les plus proches
- Le nombre de partitions (classes) n'est pas fixé a priori

→ Méthodes de partitionnement  Exemple : Centres mobiles

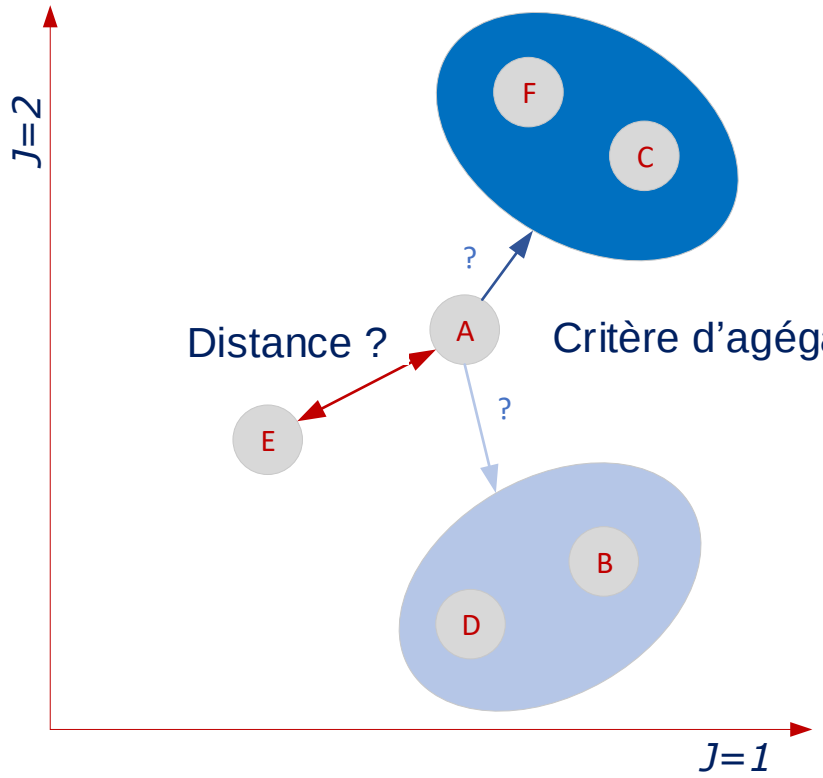
- Méthodes d'agrégation séquentielles (regroupement)
- Regroupement séquentiel autour de k centre (moyennes)
- Le nombre de partitions (classes) est fixé a priori

→ Méthodes mixtes  Partitionnement + hiérarchique

- **Objectif** : Effectuer des regroupements des données en k classes (k << n) de manière à rassembler dans chaque classe les individus les plus semblables

→ On a donc besoin définir a priori une métrique (une distance)

→ On a donc besoin, à partir d'une distance préalablement définie, de choisir un critère d'agrégation pour affecter un individus à un groupe (à une classe)



## Distance pour des données continues

La distance d est une application du produit cartésien E x E dans $R^+$ satisfaisant aux axiomes suivants

- Symétrie $d(x_i, x_{i'}) = d(x_{i'}, x_i),\ \forall x_i \in E,\ \forall x_{i'} \in E$
- Positivité $d(x_i, x_{i'}) > 0 \Leftrightarrow x_i \neq x_{i'}\ et\ d(x_{i'}, x_i) = 0 \Leftrightarrow x_i = x_{i'},\ \forall x_i \in E,\ \forall x_{i'} \in E$
- Inégalité triangulaire $d(x_i, x_{i'}) \leq d(x_i, x_{i''}) + d(x_{i''}, x_j),\ \forall x_i \in E,\ \forall x_{i'} \in E, \forall x_{i''} \in E$
- Similarité $d_{ij} = d(x_i, x_j) \quad s_{ij} = \dfrac{1}{1+d_{ij}}$

→ Définition

Soit x un individu caractérisé par p variables

$$x_i \rightarrow x_{i,1}, x_{i,2}, \ldots, x_{ip}$$

$$d_{i,i'} = \left\{ \sum_{j=1}^{p} \alpha_i \left| x_i - x_{i'} \right|^{\lambda} \right\}^{\frac{1}{\lambda}}$$

distance de Minkowski

→ Quelques exemples de distance

$$d_{i,i'} = \left\{ \sum_{j=1}^{p} \alpha_i \left| x_i - x_{i'} \right|^{\lambda} \right\}^{\frac{1}{\lambda}}$$

- distance de city-block ou distance de Manhatan

$$\lambda = 1, \alpha_j = 1 (\forall j = 1,2,....p) \rightarrow d_{ii'} = \sum_{j=1}^{p} \left| x_i - x_{i'} \right|$$

- distance Euclidienne

$$\lambda = 2, \alpha_j = 1 (\forall j = 1,2,....p) \rightarrow d_{ii'}^2 = \sum_{j=1}^{p} \left( x_i - x_{i'} \right)^2$$

- distance Euclidienne centrée réduite

$$\lambda = 2, \alpha_j = \frac{1}{s_j^2} \ , s_j^2 = \sqrt{\frac{1}{n} \sum_{j=1}^{p} \left( x_i^2 - \overline{x}_i^2 \right)} \quad (\forall j = 1,2,....p) \quad d_{ii'}^2 = \sum_{j=1}^{p} \frac{1}{s_j^2} \left( x_i - x_{i'} \right)^2$$

distance de Mahalanobis ,……

En général : distance euclidienne la plus utilisée mais pas uniquement…. Attention en fonction des distance utilisées, les résultats peuvent être différents !!!

## Distance pour des données qualitatives (dénombrements)

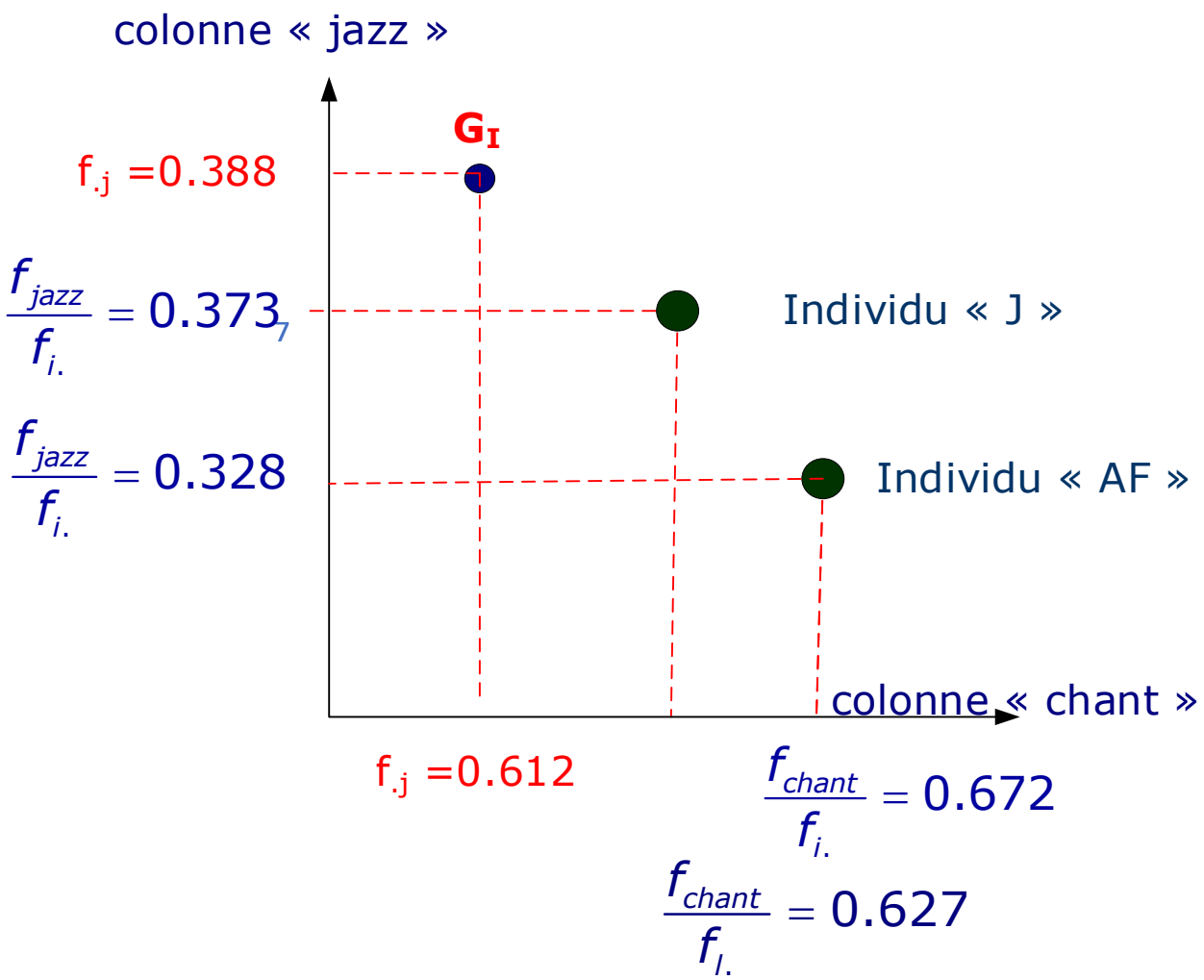
| kij | chant | jazz | somme |
|---|---|---|---|
| J | 69 | 41 | 110 |
| AF | 172 | 84 | 256 |
| AM | 133 | 118 | 251 |
| V | 27 | 11 | 38 |
| somme | 401 | 254 | 655 |

(69/655)

| fij | chant | jazz | somme |
|---|---|---|---|
| J | 0,105 | 0,063 | 0,168 |
| AF | 0,263 | 0,128 | 0,391 |
| AM | 0,203 | 0,180 | 0,383 |
| V | 0,041 | 0,017 | 0,058 |
| somme | 0,612 | 0,388 | 1,000 |

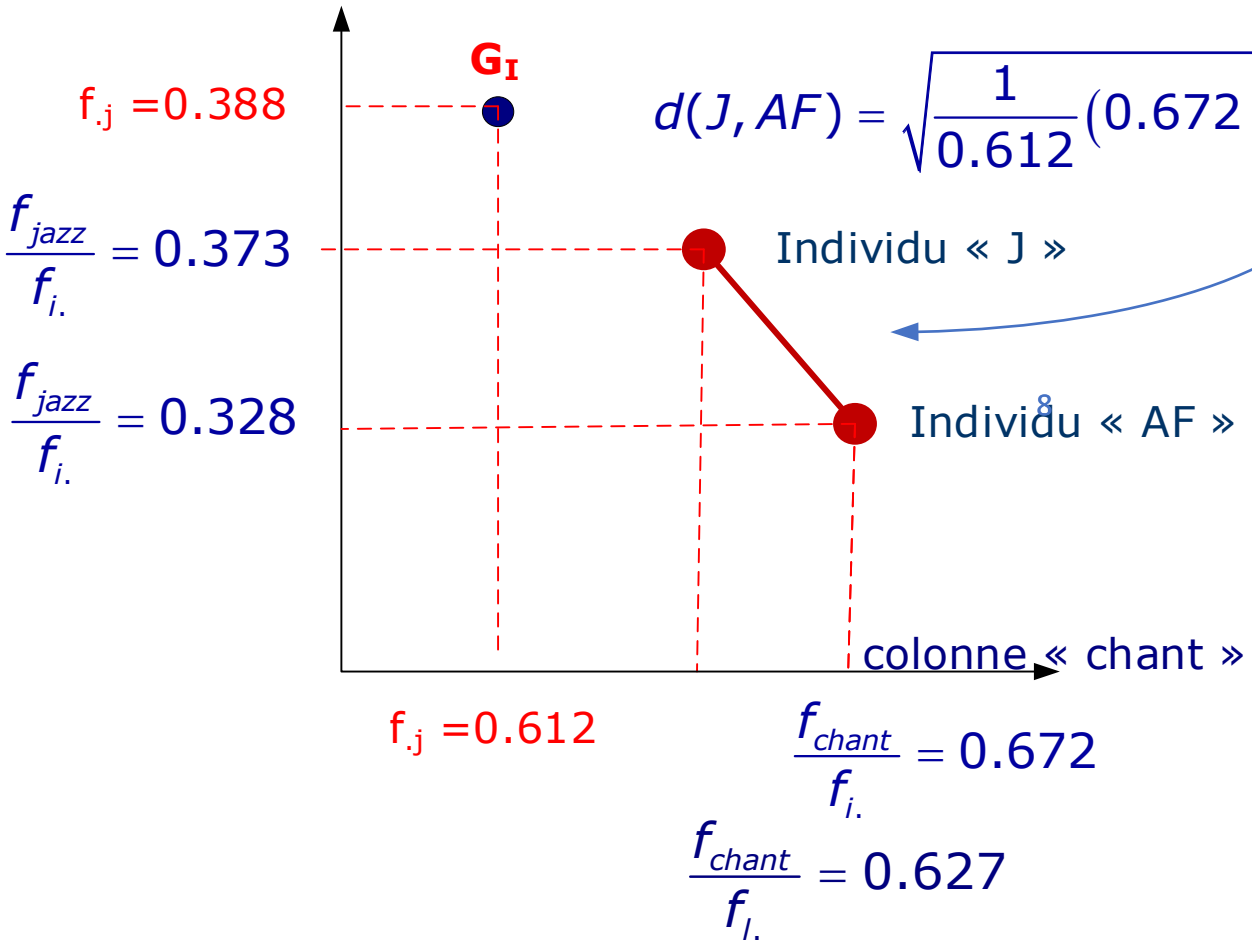
(69/110)

| fij/fi. | chant | jazz | somme |
|---|---|---|---|
| J | 0,627 | 0,373 | 1,000 |
| AF | 0,672 | 0,328 | 1,000 |
| AM | 0,530 | 0,470 | 1,000 |
| V | 0,711 | 0,289 | 1,000 |



Distance entre deux individus

colonne « jazz »




$$d(J, AF) = \sqrt{\frac{1}{0.612}(0.672 - 0.627)^2 + \frac{1}{3.88}(0.328 - 0.373)^2}$$

généralisation

$$d(i, I) = \sum_{i=1}^{p} \frac{1}{f_{.j}} \left( \frac{f_{i,j}}{f_{i.}} - \frac{f_{I,j}}{f_{I.}} \right)^2$$

## Distance pour des données disjonctives

| | barres de céréales | crèmes dessert | gateau de riz |
|---|---|---|---|
| **Chocolat** | oui | non | oui |
| **Beurre** | non | non | oui |
| **Liquide** | non | oui | non |
| **Parfum mandarine** | non | non | oui |
| **Emballage métal** | non | oui | oui |
| **Mini dose** | oui | oui | non |
| **Sucre** | oui | oui | oui |
| **Riz** | oui | non | oui |
| **Edulcorant** | non | non | oui |
| **Colorant** | non | non | oui |

9

- Le nombre de points communs est appelé coïncidence
- Cette mesure permet de construire un mesure quantitative de la similarité entre les objets (variables)

| | barres de céréales | crèmes dessert | gateau de riz | Barres de céréales | crèmes dessert | |
|---|---|---|---|---|---|---|
| **Chocolat** | 1 | 0 | 1 | oui | non | non coincidence |
| **Beurre** | 0 | 0 | 1 | non | non | coincidence négative |
| **Liquide** | 0 | 1 | 0 | non | oui | non coincidence |
| **Parfum mandarine** | 0 | 0 | 1 | non | non | coincidence négative |
| **Emballage métal** | 0 | 1 | 1 | non | oui | non coincidence |
| **Mini dose** | 1 | 1 | 0 | oui | oui | coincidence |
| **Sucre** | 1 | 1 | 1 | oui | oui | coincidence |
| **Riz** | 1 | 0 | 1 | oui | non | non coincidence |
| **Edulcorant** | 0 | 0 | 1 | non | non | coincidence négative |
| **Colorant** | 0 | 0 | 1 | non | non | coincidence négative |

| | | Barre de céréales | |
|---|---|---|---|
| | | oui | non |
| crème déssert | oui | 2 | 2 |
| | non | 2 | 4 |
| gâteau de riz | oui | 3 | 5 |
| | non | 2 | 0 |

→ *Calcul des indices de similarité*

*N : nombre total de comparaisons = nombre de composants*
*C+ : nombre de coincidences positives*
*C- : nombre de coincidences négatives*
*C+- : nombre de coincidences positives et négatives*

→ *Quantification (quelques indices…les plus communs)*

*Indice de Russel* $S = C^{+} / N$

*Indice de Jaccard* $S = C^{+} / (N - C^{-})$

*Indice de Sokal* $S = C^{+-} / N$

| | S(BC-CD) | S(BC-GR) |
|---|---|---|
| Russel | 20% | 30% |
| Jaccard | 33% | 30% |
| Sokal | 60% | 30% |

*La similarité entre des objets dépend donc de l'indice choisi !!*

| | 1 | 0 | |
|---|---|---|---|
| 1 | a | b | a+b |
| 0 | c | d | c+d |
| | a+c | b+d | n |

$$S_1 = \frac{a}{a+b+c}$$ Indice de communauté de Jaccard

$$S_2 = \frac{a+d}{n}$$ Indice de Sokal & Michener

$$S_3 = \frac{a}{a+2(b+c)}$$ Indice de Sokal & Sneath

$$S_4 = \frac{a+d}{a+2(b+c)+d}$$ Indice de Rogers et Tanimoto

$$S_5 = \frac{2a}{2a+b+c}$$ Indice de Sorensen

$$S_6 = \frac{a-b-c+d}{n}$$ Indice de Gower & Legendre

$$S_7 = \frac{a}{\sqrt{(a+b)(a+c)}}$$ Indice de Ochiai

$$S_8 = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$ Indice de Sockal & Sneath

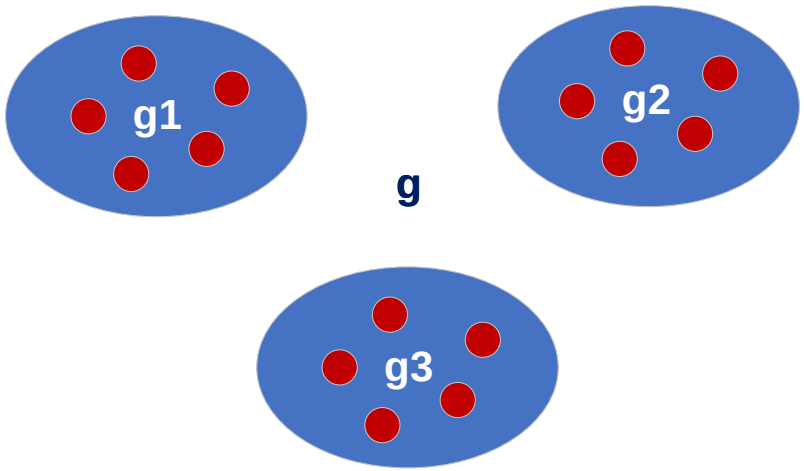
$$S_9 = \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$$ Phi de Pearson

→ Tous ces indices sont < 1

La distance associée à ces indices est définie par :

$$D_k = \sqrt{1 - S_k}$$

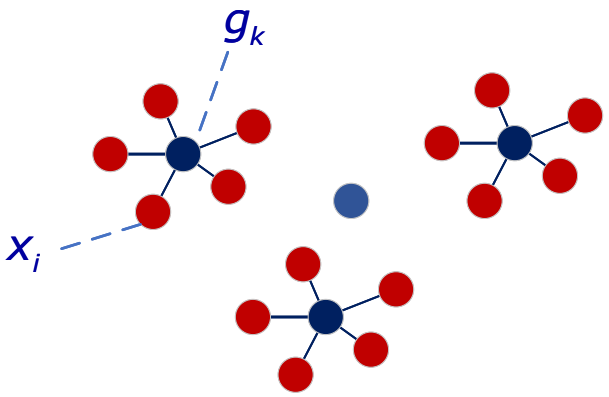
- N points dans un espace euclidien
- La distance entre les points est une distance euclidienne
- g1,g2,g3 sont les centres de gravité des partitions (classes)
- g est le centre de gravité du nuage de point




→ Soit la dispersion (appelée inertie) des points par rapport à leur centre de gravité

- On définit l'Inertie intra-classe $I_w = I_1 + I_2 + I_3$
- On définit l'Inertie inter classe $I_B$

  dispersion des classes par rapport au centre de gravité du nuage de points
- On définit l'Inertie totale du nuage de point $I_{Total} = I_W + I_B$

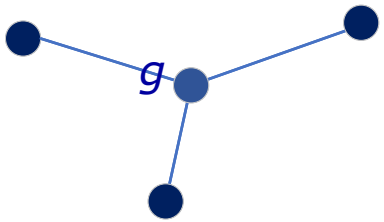



Inertie intra classe

$$I_W = \sum_{k=1}^{K} I_k$$

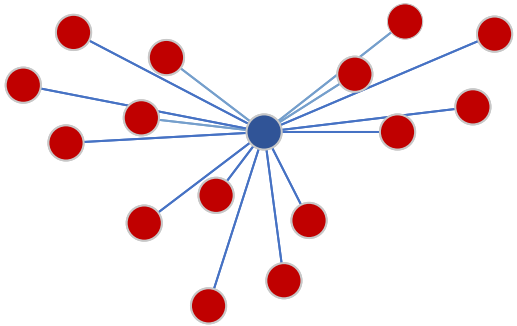
Inertie d'une classe k

$$I_W = \frac{1}{n} \sum_{i=1}^{n_k} n_k d^2(x_i, g_k)$$



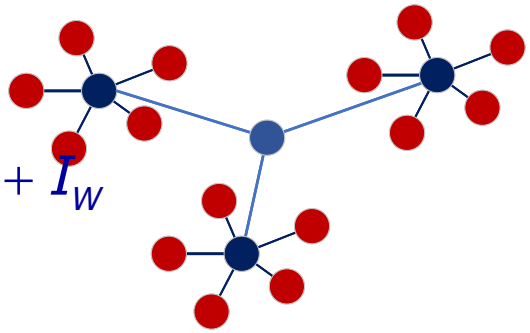

Inertie inter classe

$$I_B = \frac{1}{n} \sum_{k=1}^{K} n_k d^2(g_k, g)$$



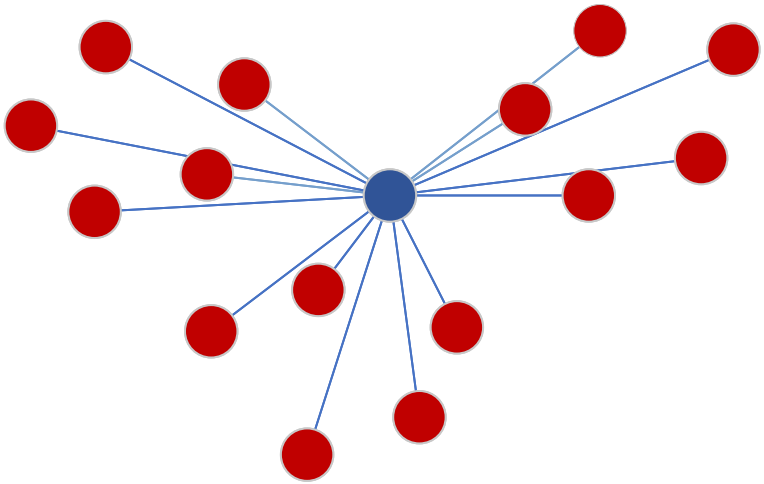
Inertie totale

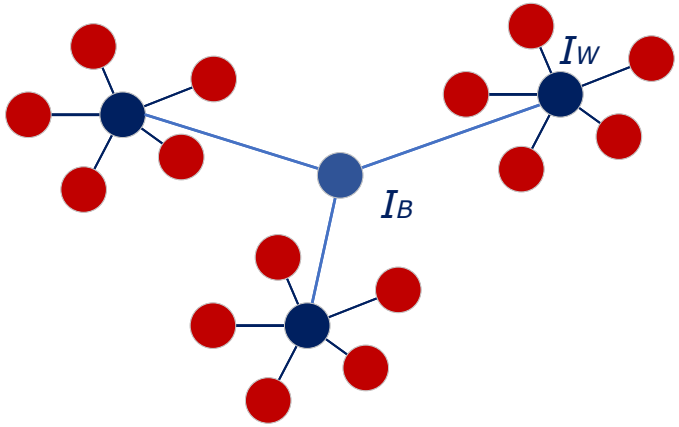
$$I_{Total} = \frac{1}{n} \sum_{i=1}^{n} d^2(x_i, g) = I_B + I_W$$



- Rmq: une classe est homogène si son inertie est faible

→ *Classification : maximiser l'inertie inter groupe (IB) et minimiser l'inertie intra groupe (IW)*



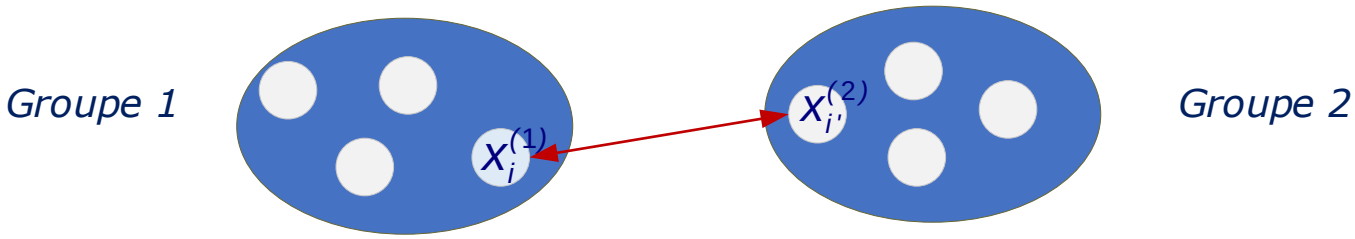



*Inertie totale est bien évidemment la même quelque soit la classification !!*

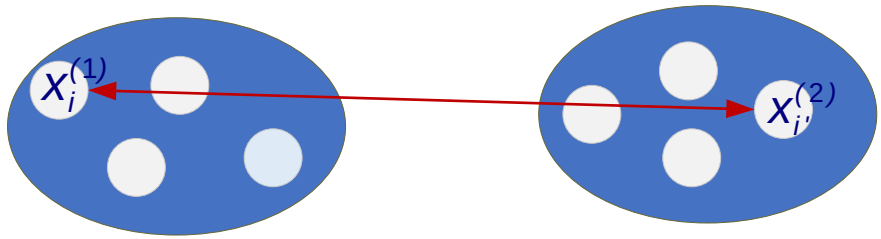
$$I_{Total} = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g) = I_B + I_W$$

## La notion d'agrégation correspond « au regroupement des classes »

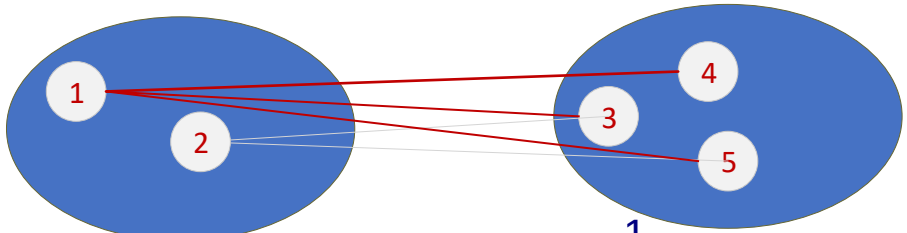
→ *Critère du saut minimal* $d(G_1,G_2) = min(x_i^{(1)}, x_{i'}^{(2)}) \quad x^{(k)} \in G$




→ *Critère du saut maximal* $d(G_1,G_2) = max(x_i^{(1)}, x_{i'}^{(2)}) \quad x^{(k)} \in G$



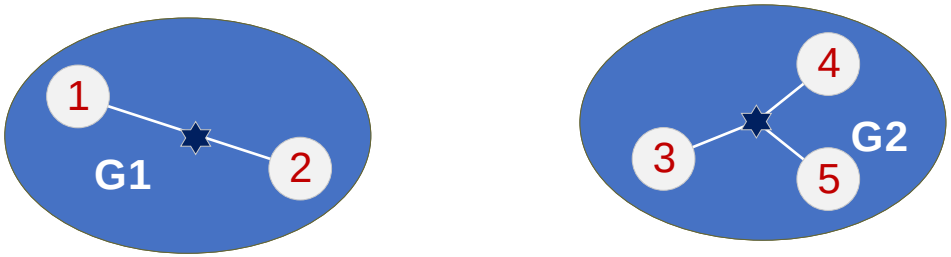

→ *Critère du saut moyen* $d(G_1,G_2) = \frac{1}{n_{G_1} n_{G_2}} \sum (x_i^{(1)}, x_{i'}^{(2)}) \quad x^{(k)} \in G$



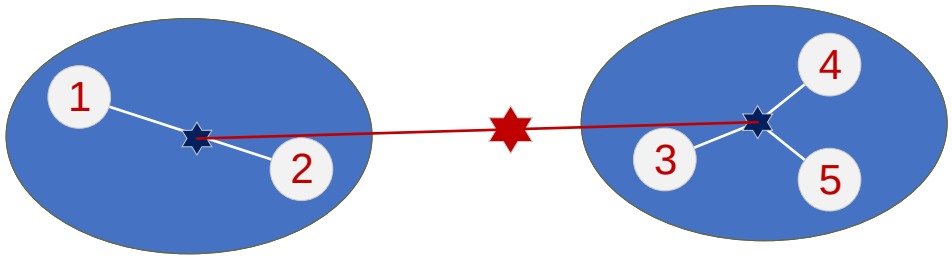

$$d(G_1,G_2) = \frac{1}{2x3}(d_{1,3} + d_{1,4} + d_{1,5} + d_{2,3} + d_{2,4} + d_{2,5})$$

→ Critère de Ward (l'un des plus utilisé…)

Les classes (groupes) sont représentées par leur centre de gravité ✸



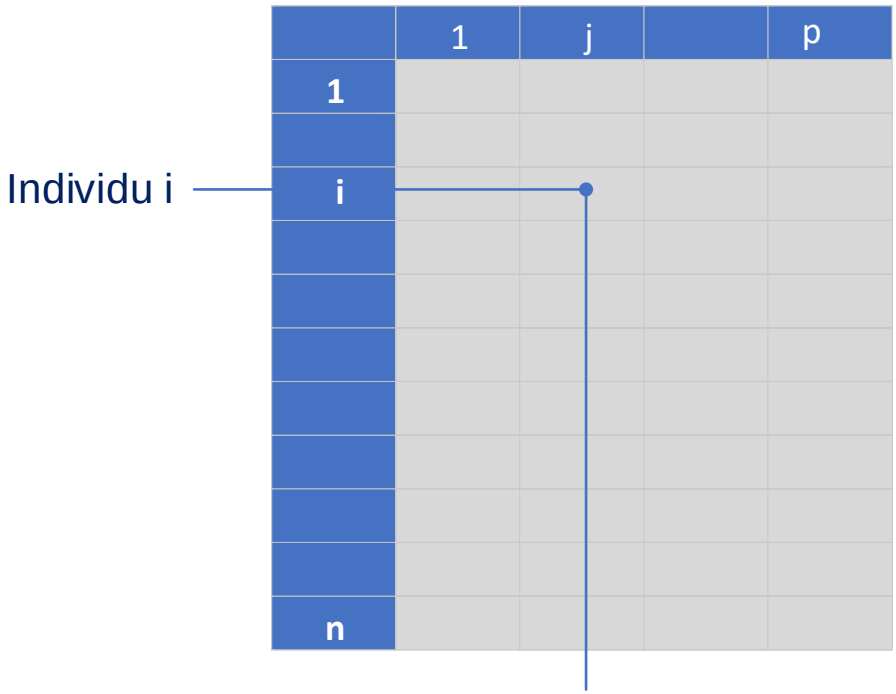

- La fusion de 2 classes est représentée par le remplacement des 2 points par leur centre de gravité muni de la somme des masses (barycentre)

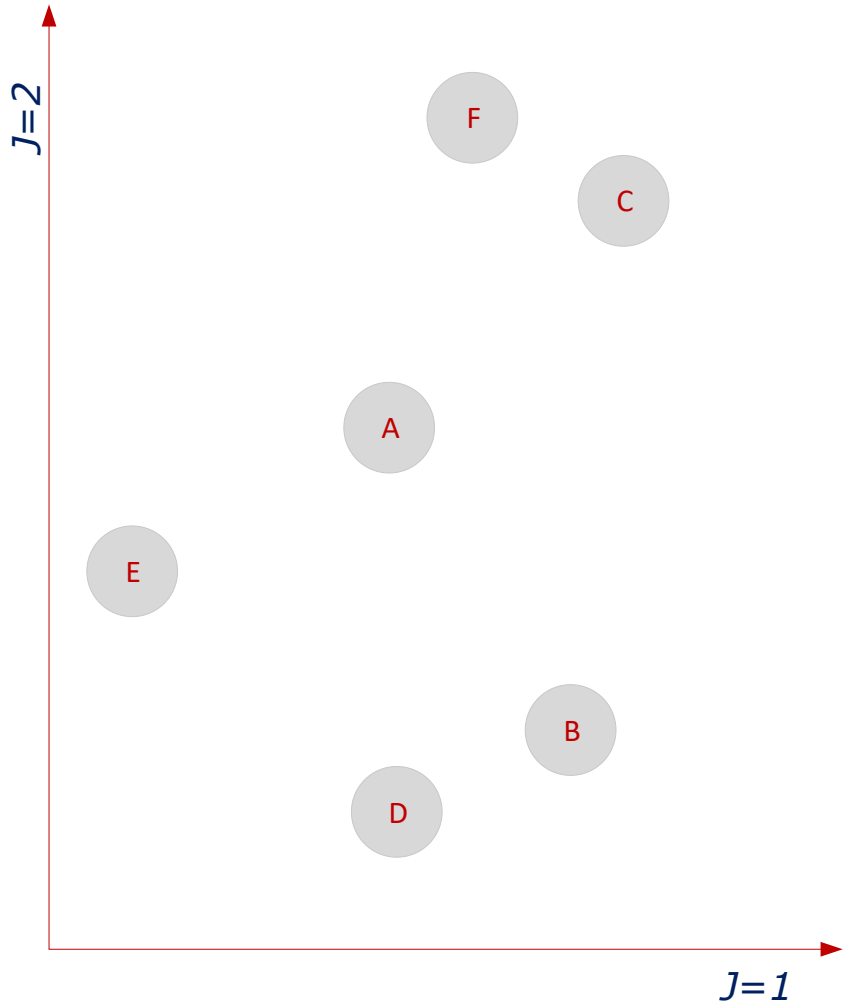



$$d_W = \frac{n_1 n_2}{n_1 + n_2} d^2(G_1, G_2) = \Delta_w$$

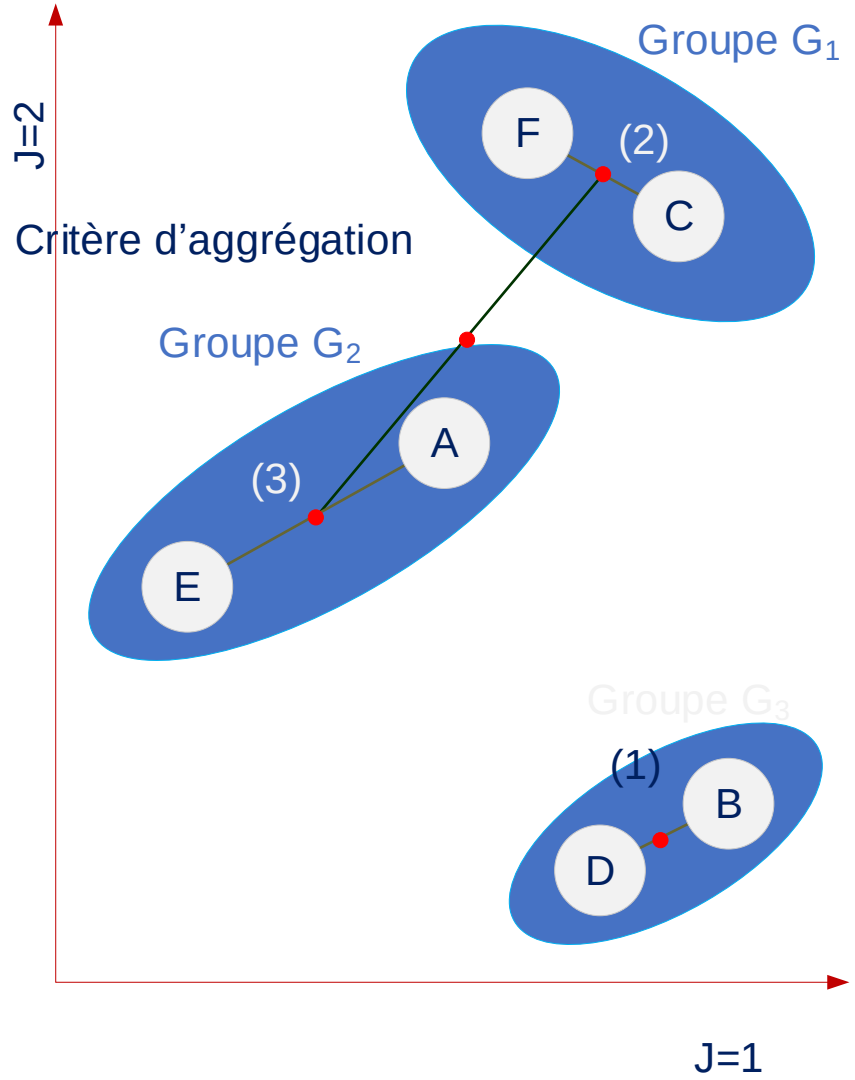
n1 et n2 sont les masses respectives des classes 1 et 2 (nb de points)

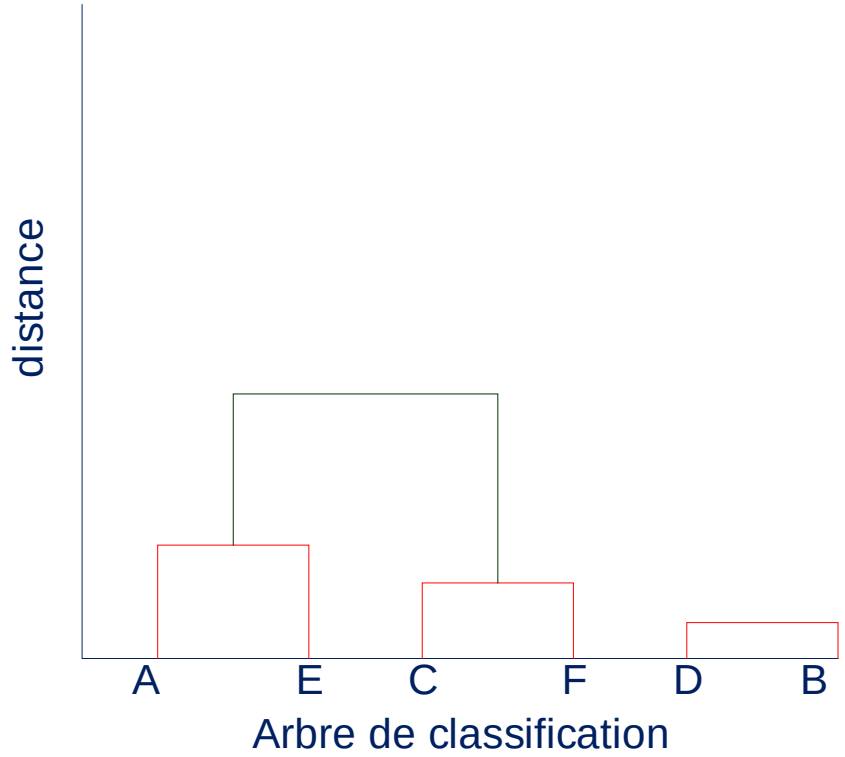
Exemple sur 2 variables



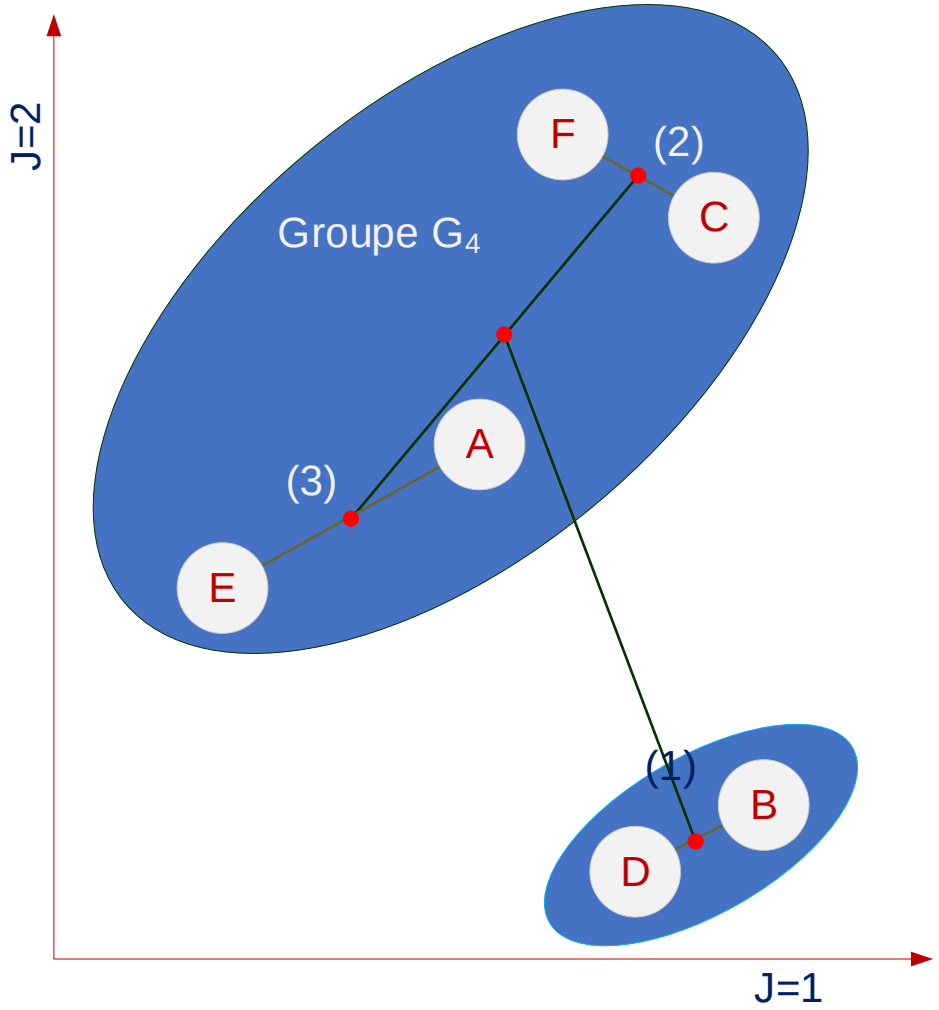



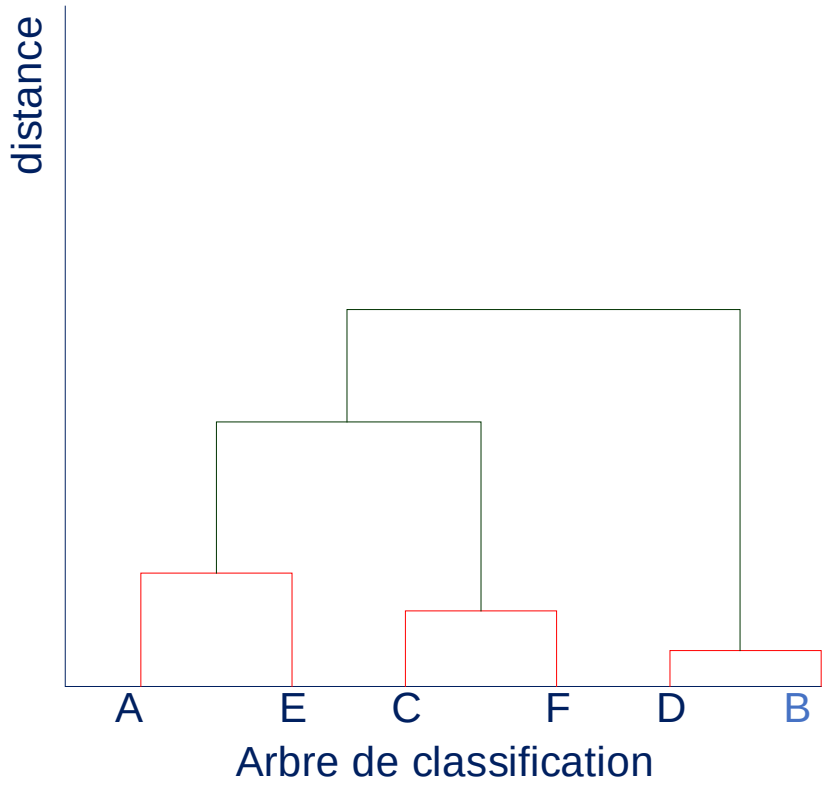
1. calcul des distances entre les points
2. aggrégation (critère)



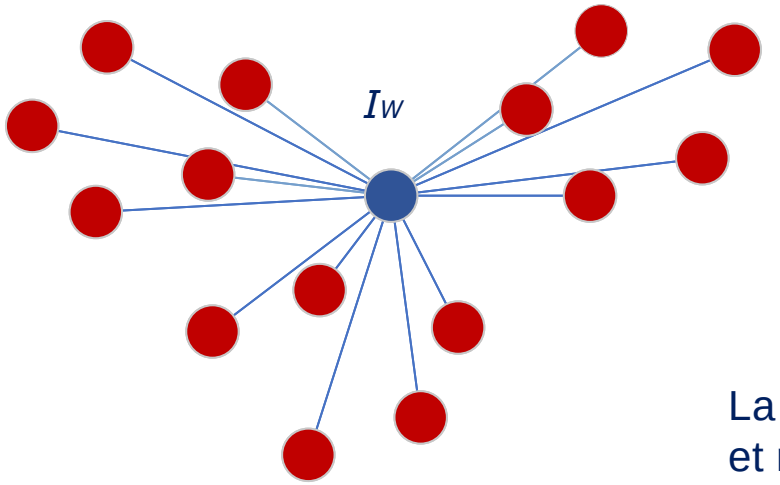
La hauteur des branches correspond à la distance entre les éléments regroupés



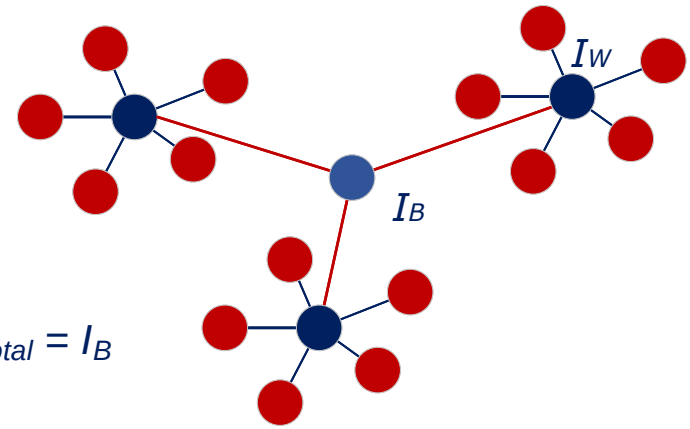
$$I_{Total} = I_W + I_B$$

Au départ, pas de classes l'inertie inter - classe est nulle, L'inertie intra-classe est maximale (estimée à partir du centre de gravité du nuage de points)




$I_{Total} = I_W$

La classification revient à minimiser l'inertie intraclasse et maximiser l'inertie interclasse… ce qui revient au même




$$d_W = \frac{n_1 n_2}{n_1 + n_2} d^2(G_1, G_2) = \Delta_w \longleftrightarrow I_{Total} = I_B$$

## Etape 1

→ a. Calcul des distances euclidiennes carrées entre le points

| | A | B | C | B | E |
|---|---|---|---|---|---|
| A | 0,00 | | | | |
| B | 16,00 | 0,00 | | | |
| C | 1,00 | 17,00 | 0,00 | | |
| D | 9,00 | 25,00 | 4,00 | 0,00 | |
| E | 10,00 | 2,00 | 9,00 | 13,00 | 0,00 |

Le « poids » de chaque point est égal à 1

$n_A=n_B=n_C=n_D=n_E=1$

$p_A=p_B=p_C=p_D=p_E=1$

→ b. Calcul des critères de Ward entre les points (à ce stade, les point sont « confondus » avec les centres de gravité)

| | A | B | C | B | E |
|---|---|---|---|---|---|
| A | 0,00 | | | | |
| B | 8,00 | 0,00 | | | |
| C | **0,50** | 8,50 | 0,00 | | |
| D | 4,50 | 12,50 | 2,00 | 0,00 | |
| E | 5,00 | 1,00 | 4,50 | 6,50 | 0,00 |

Correspond à l'inertie inter « pondérée par les masses »

$$\Delta = \frac{p_i p_{i'}}{p_i + p_{i'}} d^2(G_i, G_{i'})$$

→ c. agrégation des points dont la perte d'inertie est minimale et création d'un nouveau groupe

$$\Delta_{A,C} = \frac{p_A p_c}{p_A + p_c} d^2(A,C) = \frac{1*1}{1+1} * 1 = 0.5$$

On crée donc un nouveau groupe (6) qui inclue les points A et C

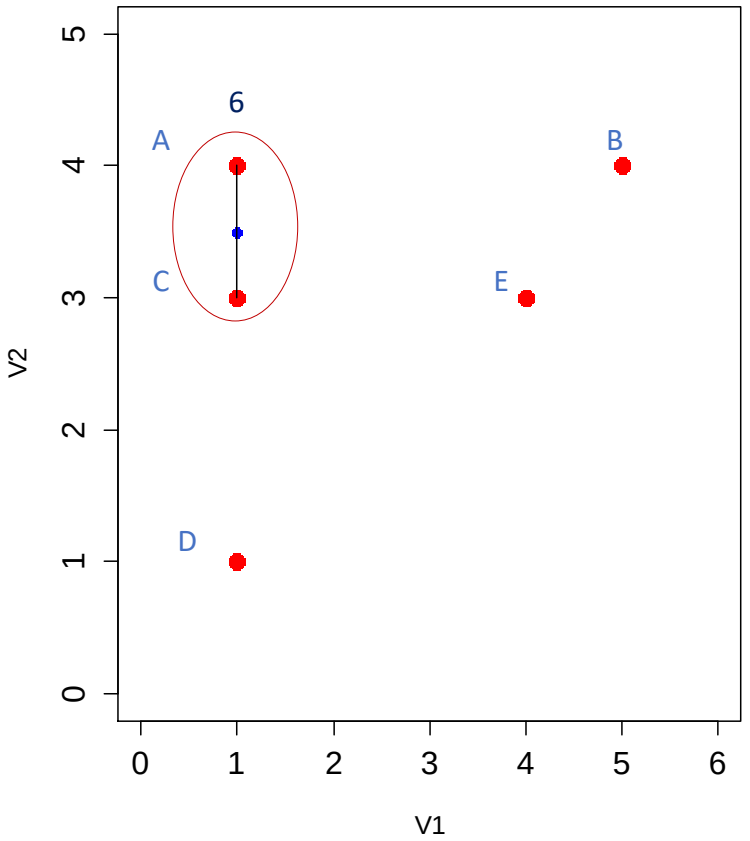
→ *d. agrégation des points dont la perte d'inertie est minimale et création d'un nouveau groupe*

- *(i) calcul du centre de gravité du nouveau groupe (6)*

$$\begin{cases} G_x = \dfrac{n_A x_A + n_C x_C}{n_A + n_C} = \dfrac{1*1+1*1}{1+1} = 1 \\ G_y = \dfrac{n_A y_A + n_C y_C}{n_A + n_C} = \dfrac{1*4+1*3}{1+1} = 3.5 \end{cases}$$

- *(ii) poids du groupe 6* $n_6 = n_A + n_C = 2$

| | V1 | V2 | nb.elem |
|---|---|---|---|
| "6" | 1,00 | 3,50 | 2,00 |
| B | 5,00 | 4,00 | 1,00 |
| D | 1,00 | 1,00 | 1,00 |
| E | 4,00 | 3,00 | 1,00 |



## Etape 2

→ a. on calcule directement les pertes d'inertie
… qui corerspondent aux critères inter-classe

Exemple :Calcul pour la première colonne

| | "6" | B | D | E |
|---|---|---|---|---|
| "6" | 0,00 | 0,00 | 0,00 | 0,00 |
| B | 10,83 | 0,00 | 0,00 | 0,00 |
| D | 4,17 | 12,50 | 0,00 | 0,00 |
| E | 6,17 | 1,00 | 6,50 | 0,00 |

$$\Delta_{(6,B)} = \frac{p_6 * p_B}{p_6 + p_B} d^2(6,B) = \frac{2*1}{2+1}\left((1-5)^2 + (3.5-4)^2\right) = 10.83$$

$$\Delta_{(6,D)} = \frac{p_6 * p_D}{p_6 + p_D} d^2(6,D) = \frac{2*1}{2+1}\left((1-1)^2 + (3.5-1)^2\right) = 4.16$$

$$\Delta_{(6,E)} = \frac{p_6 * p_E}{p_6 + p_E} d^2(6,E) = \frac{2*1}{2+1}\left((1-4)^2 + (3.5-3)^2\right) = 6.16$$

23

→ b. agrégation et création d'un nouveau groupe

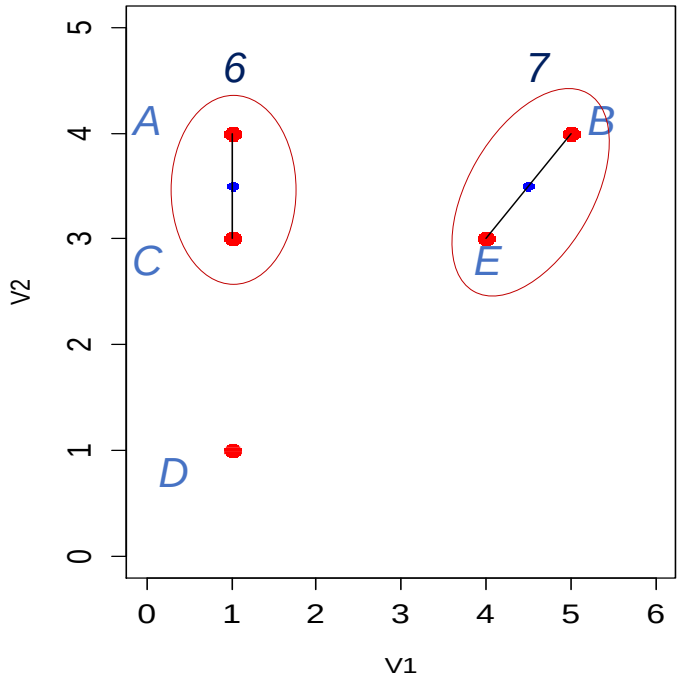
On crée donc un nouveau groupe qui inclue les points B et E (groupe 7)

- (i) calcul du centre de gravité du nouveau groupe (7)

$$\begin{cases} G_x = \dfrac{n_B x_B + n_E x_E}{n_B + n_E} = \dfrac{1*5+1*4}{1+1} = 4.5 \\ G_y = \dfrac{n_B y_B + n_E y_E}{n_B + n_E} = \dfrac{1*4+1*3}{1+1} = 3.5 \end{cases}$$

- *(ii) poids du groupe 7* $n_7 = n_D + n_E = 2$

| | V1 | V2 | nb.elem |
|---|---|---|---|
| "7" | 4,50 | 3,50 | 2,00 |
| "6" | 1,00 | 3,50 | 2,00 |
| D | 1,00 | 1,00 | 1,00 |




## Etape 3

→ *a. calcul des critères*

| | "7" | "6" | D |
|---|---|---|---|
| "7" | 0,00 | 0,00 | 2,00 |
| "6" | 12,25 | 0,00 | 2,00 |
| D | 12,33 | 4,16 | 0,00 |

*Exemple :Calcul pour la première colonne*

$$\begin{cases} \Delta_{(6,7)} = \dfrac{p_6 * p_7}{p_6 + p_7} d^2(6,7) = \dfrac{2*2}{2+2}\left((1-4.5)^2 + (3.5-3.5)^2\right) = 12.25 \\ \Delta_{(7,D)} = \dfrac{p_7 * p_D}{p_7 + p_D} d^2(7,D) = \dfrac{2*1}{2+1}\left((1-4.5)^2 + (1-3.5)^2\right) = 12.33 \end{cases}$$

→ b. agrégation des points dont la perte d'inertie est minimale et création d'un nouveau groupe

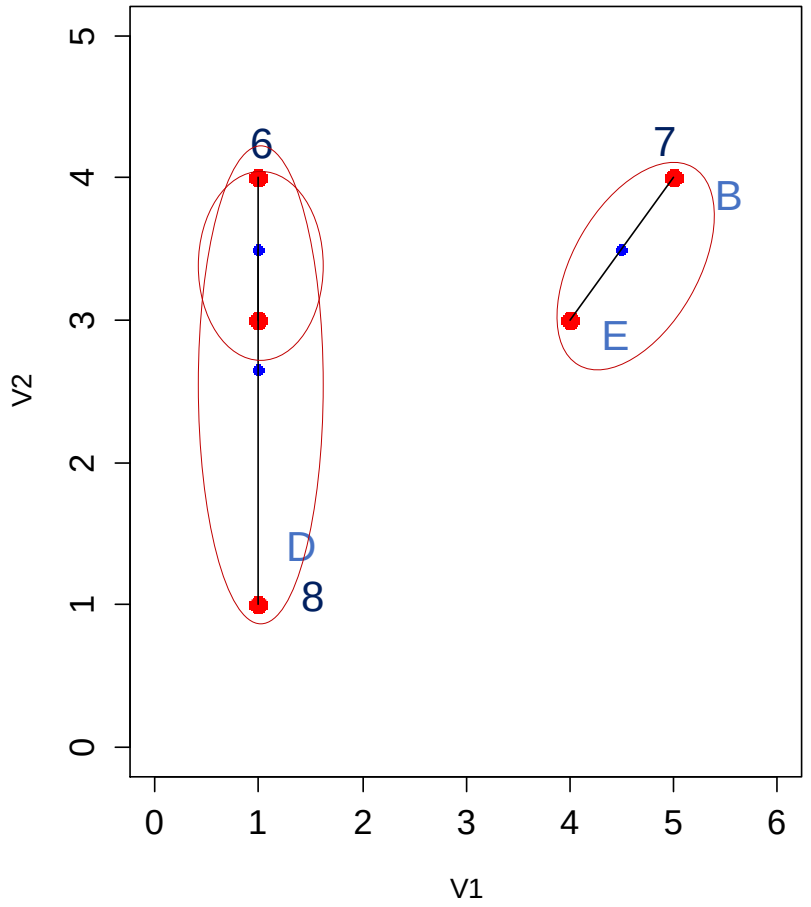
On crée donc un nouveau groupe qui inclue les points D et groupe 6 (groupe 8)

- (i) calcul du centre de gravité du nouveau groupe (7)

$$\begin{cases} G_x = \dfrac{p_6 x_6 + p_D x_D}{p_6 + p_D} = \dfrac{2\times 1 + 1\times 1}{2+1} = 1 \\ G_y = \dfrac{p_6 y_6 + p_D y_6}{p_6 + p_D} = \dfrac{2\times 3.5 + 1\times 1}{2+1} = 2.67 \end{cases}$$

- (ii) poids du groupe 8 $\quad n_8 = n_D + n_6 = 3$

| | V1 | V2 | nb.elem |
|---|---|---|---|
| "8" | 1,00 | 2,66 | 3,00 |
| "7" | 4,50 | 3,50 | 2,00 |



## Dernière étape

→ On regroupe 7 et 8

- (i) calcul du centre de gravité du nouveau groupe (9)

$$\begin{cases} G_x = \dfrac{p_7 x_7 + p_8 x_8}{p_7 + p_8} = \dfrac{2 \times 4.5 + 3 \times 1}{2 + 3} = 2.4 \\ G_y = \dfrac{p_7 y_7 + p_8 y_8}{p_7 + p_8} = \dfrac{2 \times 3.5 + 3 \times 2.67}{2 + 3} = 3 \end{cases}$$

Le centre de gravité du groupe 9 correspond au centre de gravité du nuage de points

| | V1 | V2 |
|---|---|---|
| A | 1,00 | 4,00 |
| B | 5,00 | 4,00 |
| C | 1,00 | 3,00 |
| D | 1,00 | 1,00 |
| E | 4,00 | 3,00 |
| moyenne | 2,40 | 3,00 |

- (ii) calcul de perte d'inertie ( critère de Ward)

$$\left\{ \Delta_{(8,7)} = \frac{p_8 * p_7}{p_8 + p_7} d^2(8,7) = \frac{3*2}{3+2}\left((1-4.5)^2 + (2.25-3.5)^2\right) = 16.57 \right.$$

## Bilan des Inerties

→ Inertie Totale $I_{Total} = \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, g) = I_B + I_W$

| | V1 | V2 | ni.d²(xi,g) |
|---|---|---|---|
| A | 1,00 | 4,00 | 2,96 |
| B | 5,00 | 4,00 | 7,76 |
| C | 1,00 | 3,00 | 1,96 |
| D | 1,00 | 1,00 | 5,96 |
| E | 4,00 | 3,00 | 2,56 |
| moyenne | 2,40 | 3,00 | **21,20** |

Au départ, l'inertie intra est égale à l'inertie totale

| | Inter | Intra |
|---|---|---|
| | 0,00 | 21,20 |
| Etape1 | 0,50 | 20,70 |
| Étape 2 | 1,00 | 19,70 |
| Etape 3 | 4,17 | 15,53 |
| Etape 4 | 15,53 | 0,00 |
| | 21,20 | 21,20 |

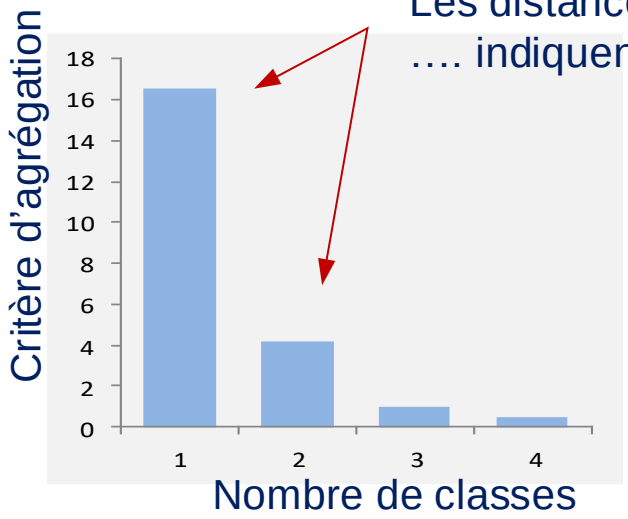
*Etape initiale : IB = 0 et IW = ITot*

*Etape finale : IB = ITot et IW = 0*

- Le choix du « bon » nombre de classes reste ouvert et dépend de la problématique posée. Il est donc indissociable de l'interprétation des classes et de l'étude !!!

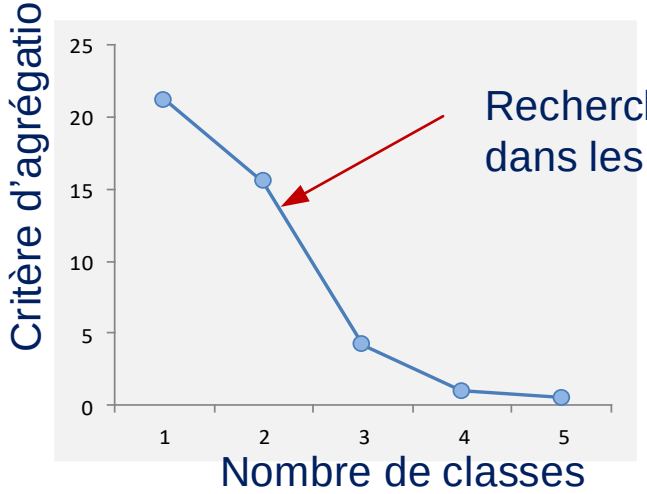
→ Hauteur des paliers d'agrégation

Les distances d'agrégation les plus élevées
…. indiquent que les groupes à fusionner sont très éloignés



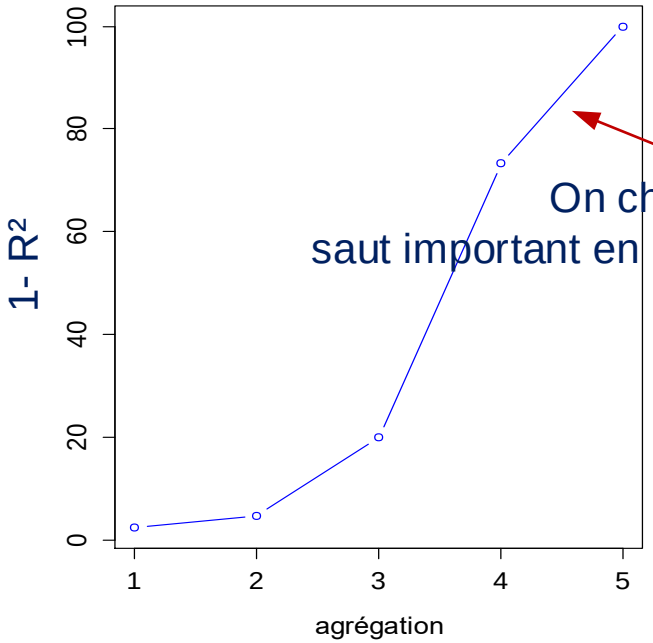

→ Evolution de l'inertie intra-classe ou inter-classe

Recherche du coude significatif d'un changement structural dans les données



→ Proportion d'inertie expliquée par les classes $R^2$ et (ou) $1- R^2$

$$R^2 = \frac{I_{inter}}{I_{Tot}} = \frac{\sum_{j=1}^{q}\sum_{i=1}^{n_q} n_i d^2(G,G_j)}{\sum_{j=1}^{q} n_i d^2(i,G)}$$




On choisit le dernier
saut important en passant de k à k+1 classes

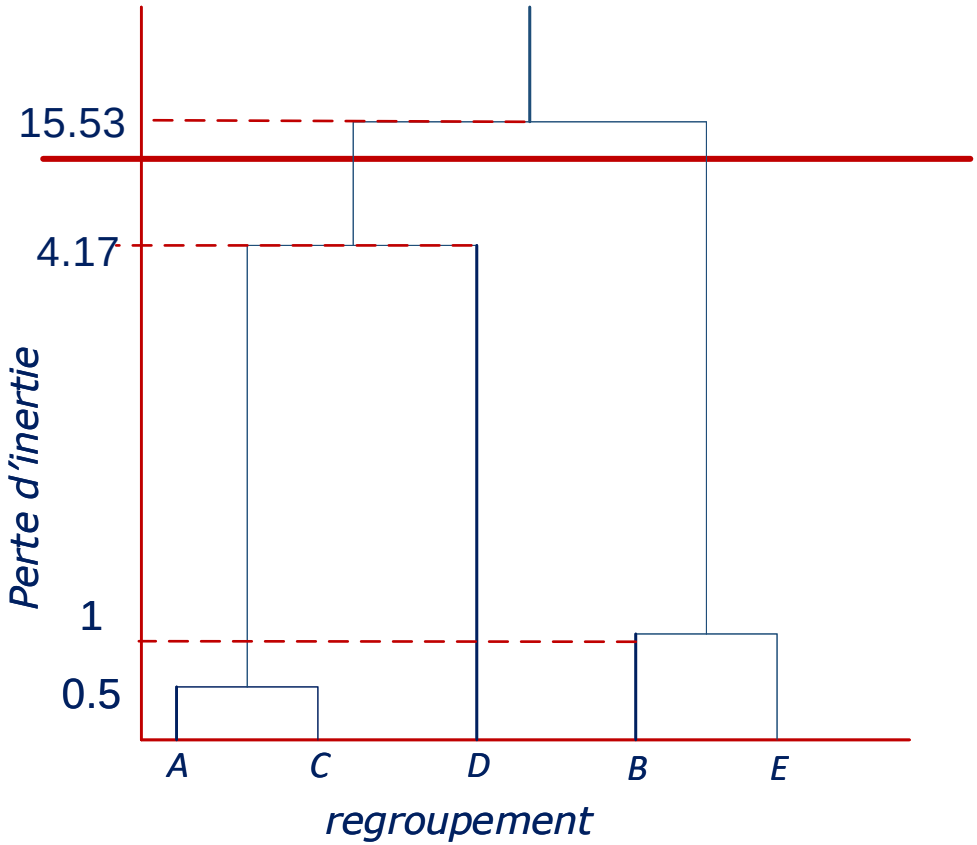
→ Critère de classification cubique

- CCC > 2 : bonne classification
- 0 < CCC < 2 : classification à vérifier
- CCC < 0 : classes trop petites (ou individus hors normes)

$$CCC = k \ln\left(\frac{1-E(R^2)}{1-R^2}\right)$$

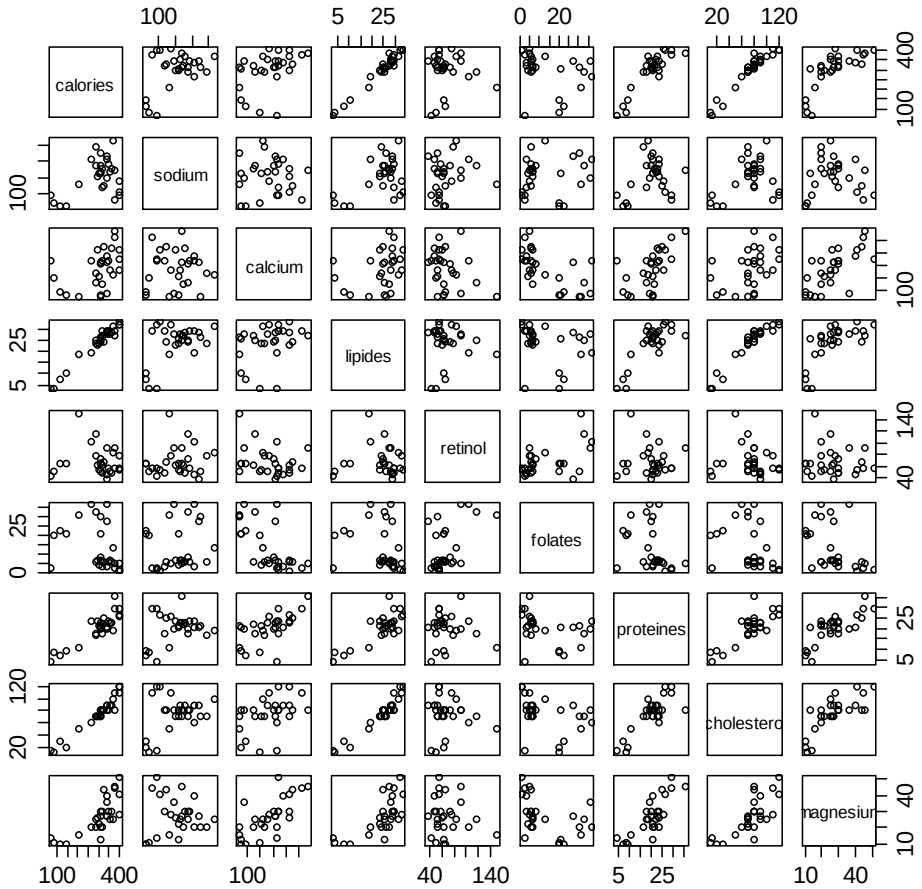
k = nombre de classes

| Inter | Intra | $1- R^2$ | $R^2$ |
|---|---|---|---|
| 0,00 | 21,20 | 0,00 | 100,00 |
| 0,50 | 20,70 | 2,36 | 97,64 |
| 1,00 | 19,70 | 4,72 | 95,28 |
| 4,17 | 15,53 | 19,67 | 80,33 |
| 15,53 | 0,00 | 73,25 | 26,75 |
| 21,20 | 0,00 | 100,00 | 0,00 |

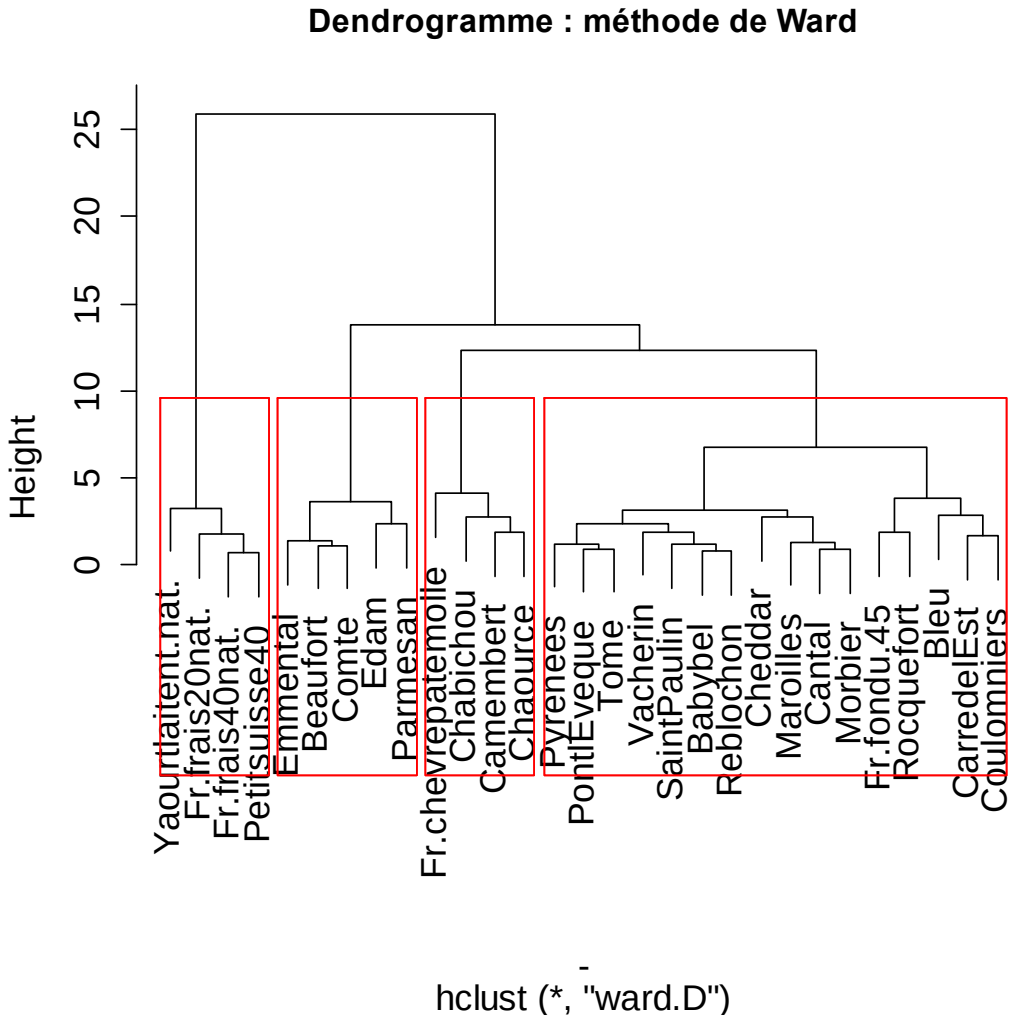


Cette analyse retranscrit une démarche de classification automatique d'un ensemble de fromages (29 observations) décrits par leur propriété nutritive (ex. protéines, lipides, etc. ; 9 variables). L'objectif est d'identifier des groupes de fromage homogènes, partageant des caractéristiques similaires.

```
# Classification Ascendante Hiérachique
# première approche
# les function standard ne permettent pas de calculer les inerties
rm (list = ls())
root  <- 'G'
dir   <- ':\\ENSEIGNEMENTS\\STAT_AF\\ANALYSE_FACTORIELLE\\R\\DATA\\'
file  <-  'EX_CAH.txt'
df<-read.table(paste(root,dir,file,sep = ''),header  = TRUE,
    row.names = 1, sep ="\t", dec = '.')
# Statistiques descriptives et graphiques
summary(df)
pairs(df)
# Classification CAH
# centrage et réduction des données
df.stand <- scale(df,center = TRUE,scale = TRUE)
# calcul de la distance euclidienne 'attention df.dist est un objet..
# de type « Distance »
"df.dist    <- dist(df.stand,method = "euclidean")
# conversion en matrice si exportation ou calculs
df.dis.mat <- as.matrix(df.dist)
```



```
# CAH avec critère de Ward
df.cah <- hclust(df.dist,method = 'ward.D')
print(df.cah)
##
## Call:
## hclust(d = df.dist, method = "ward.D")
##
## Cluster method   : ward.D
## Distance         : euclidean
## Number of objects: 29
plot(df.cah, main = 'Dendrogramme : méthode de Ward',
  cex.main = 0.9, xlab = " - ")
# On condidère 4 groupes  : découpage en 4 classes
# affectation des différentes variables aux classes
gp.cah     <- sort(cutree(df.cah,k = 4)) ; print(gp.cah)
classe.cah <- rect.hclust(df.cah,k = 4)
# graphique des Classes
rect.hclust(df.cah,k =4)
```



# CLASSIFICATION NON SUPERVISEE

## B. CENTRES MOBILES

- **Objectif** :

  Rendre Iw minimale (groupe homogène)

  Rendre IB maximale (séparation inter-groupe)

- **Méthode** : On part d'une partition arbitraire en K classes que l'on améliore progressivement jusqu'à convergence du critère choisi

  → Etape 1 : On choisit k individus comme centres initiaux des classes,

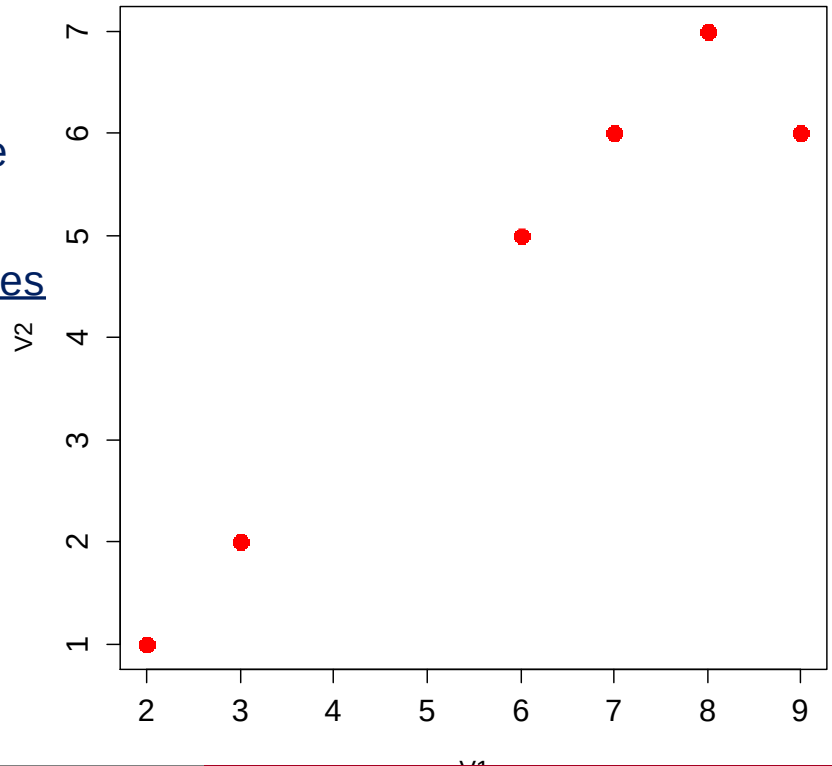
  → Etape 2: On calcule la distance entre chaque individus et chaque centre et l'on affecte l'individu au centre le plus proche,

  → Etape 3 : On recalcule le barycentre des classes,

  → Etape 4 : On réitère les étapes, 2,3 jusqu'à convergence

- **Exemple** : Classification du tableau suivant en deux groupes

| | V1 | V2 |
|---|---|---|
| A | 2,00 | 1,00 |
| B | 3,00 | 2,00 |
| C | 6,00 | 5,00 |
| D | 8,00 | 7,00 |
| E | 7,00 | 6,00 |
| E | 9,00 | 6,00 |



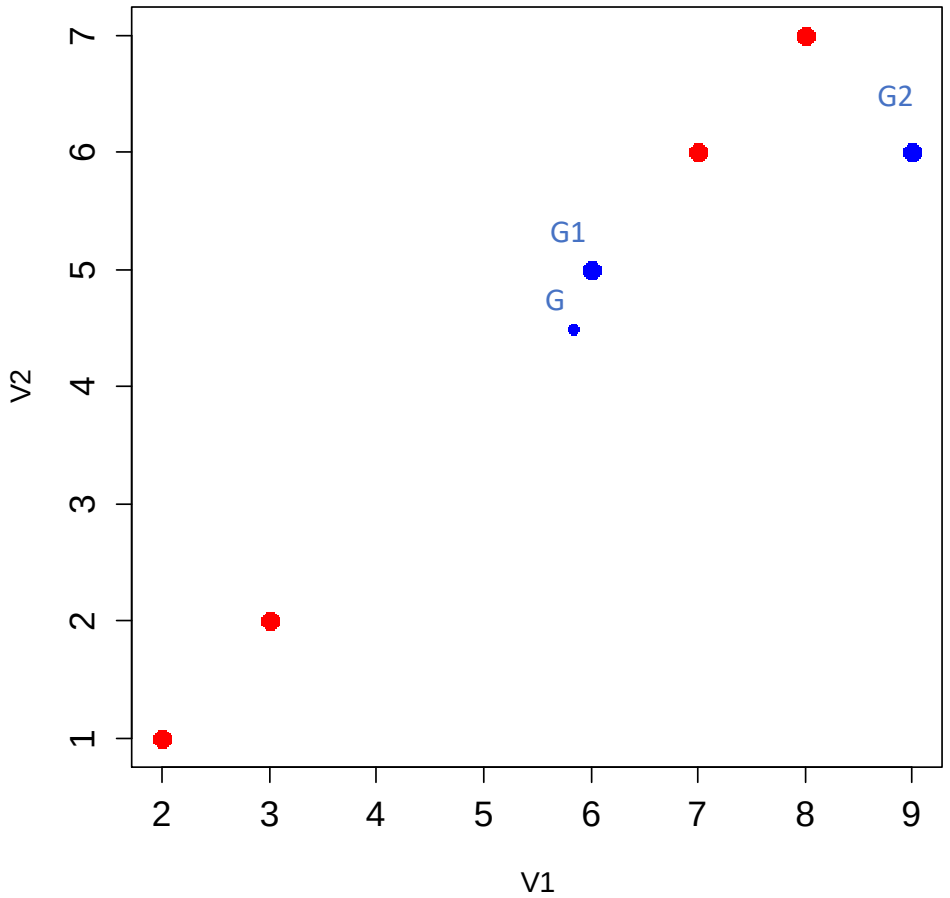
## Etape 1

→ Nombre de groupe fixé a priori

→ Les centres initiaux des groupes sont choisis au hasard parmi les points

| | V1 | V2 | ni | dé(xi,G) |
|---|---|---|---|---|
| A | 2,00 | 1,00 | 1 | 26,94 |
| B | 3,00 | 2,00 | 1 | 14,28 |
| C | 6,00 | 5,00 | 1 | 0,28 |
| D | 8,00 | 7,00 | 1 | 10,94 |
| E | 7,00 | 6,00 | 1 | 3,61 |
| E | 9,00 | 6,00 | 1 | 12,28 |
| C. gravité G | 5,83 | 4,50 | | |
| Inertie Tot | 68,33 | | | |

| | | V1 | V2 | ni |
|---|---|---|---|---|
| | A | 2,00 | 1,00 | 1 |
| | B | 3,00 | 2,00 | 1 |
| G1 | C | 6,00 | 5,00 | 1 |
| | D | 8,00 | 7,00 | 1 |
| | E | 7,00 | 6,00 | 1 |
| G2 | F | 9,00 | 6,00 | 1 |
| | C. gravité | 5,83 | 4,50 | |



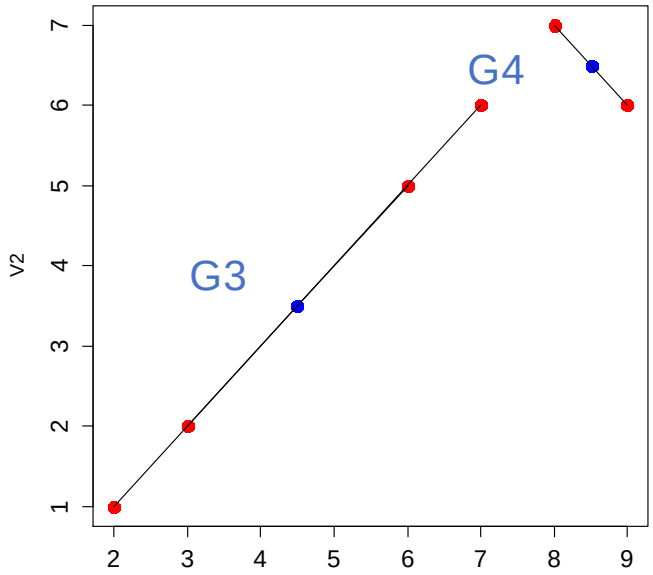
## Calcul des distances entre chaque individu et chaque centre

| | | ni | $d^2(xi,G1)$ |
|---|---|---|---|
| | A | 1 | 32,00 |
| | B | 1 | 18,00 |
| G1 | C | 1 | 0,00 |
| | D | 1 | 8,00 |
| | E | 1 | 2,00 |
| | F | 1 | 10,00 |

| | | ni | $d^2(xi,G2)$ |
|---|---|---|---|
| | A | 1 | 74,00 |
| | B | 1 | 52,00 |
| | C | 1 | 10,00 |
| | D | 1 | 2,00 |
| | E | 1 | 4,00 |
| G2 | F | 1 | 0,00 |

## Affectation des individus au centre le plus proche

| | $d^2(xi,G1)$ | $d^2(xi,G2)$ | Affectation |
|---|---|---|---|
| A | 32,00 | 74,00 | G1 |
| B | 18,00 | 52,00 | G1 |
| C | 0,00 | 10,00 | G1 |
| D | 8,00 | 2,00 | G2 |
| E | 2,00 | 4,00 | G1 |
| F | 10,00 | 0,00 | G2 |




## Calcul des coordonnées des nouveaux centres

| | V1 | V2 | d(xi,G3) |
|---|---|---|---|
| A | 2,00 | 1,00 | 12,50 |
| B | 3,00 | 2,00 | 4,50 |
| C | 6,00 | 5,00 | 4,50 |
| E | 7,00 | 6,00 | 12,50 |
| C. gravité G3 | 4,50 | 3,50 | |
| Iw | 34,00 | | |

| | V1 | V2 | d(xi,G3) |
|---|---|---|---|
| D | 8,00 | 7,00 | 0,50 |
| F | 9,00 | 6,00 | 0,50 |
| C. gravité G4 | 8,50 | 6,50 | |
| Iw | 1,00 | | |

## Etape 2

→ Calcul des distances entre chaque individu et chaque centre (étape 2)

| | V1 | V2 |
|---|---|---|
| A | 2,00 | 1,00 |
| B | 3,00 | 2,00 |
| C | 6,00 | 5,00 |
| D | 8,00 | 7,00 |
| E | 7,00 | 6,00 |
| E | 9,00 | 6,00 |
| G3 | 4,50 | 3,50 |
| G4 | 8,50 | 6,50 |

| | V1 | V2 | d(xi,G3) |
|---|---|---|---|
| A | 2,00 | 1,00 | 12,50 |
| B | 3,00 | 2,00 | 4,50 |
| C | 6,00 | 5,00 | 4,50 |
| D | 8,00 | 7,00 | 24,50 |
| E | 7,00 | 6,00 | 12,50 |
| E | 9,00 | 6,00 | 26,50 |
| G3 | 4,50 | 3,50 | |

| | V1 | V2 | d(xi,G4) |
|---|---|---|---|
| A | 2,00 | 1,00 | 72,50 |
| B | 3,00 | 2,00 | 50,50 |
| C | 6,00 | 5,00 | 8,50 |
| D | 8,00 | 7,00 | 0,50 |
| E | 7,00 | 6,00 | 2,50 |
| E | 9,00 | 6,00 | 0,50 |
| G4 | 8,50 | 6,50 | |

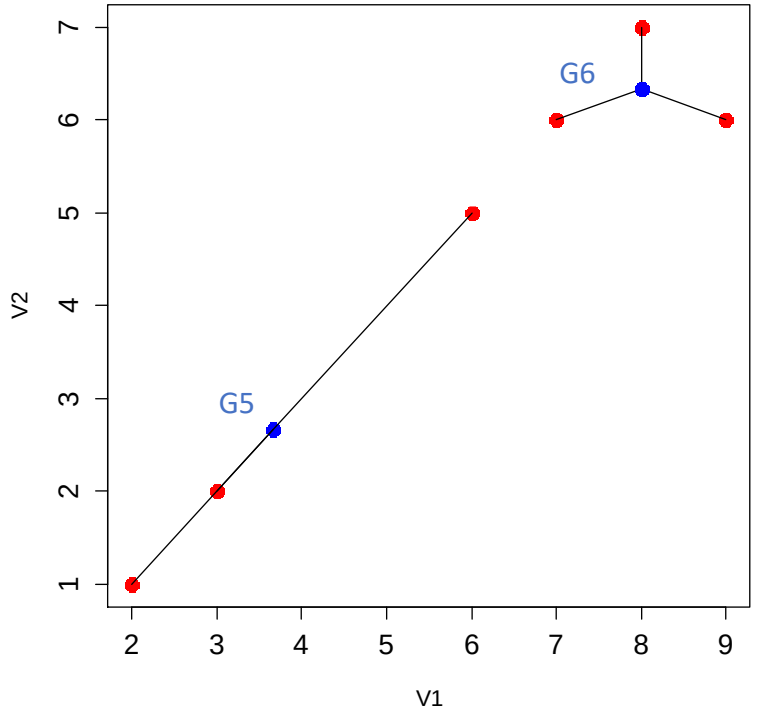
→ *Affectation des individus au centre le plus proche*

| | d(xi,G3) | d(xi,G4) | Affectation |
|---|---|---|---|
| A | 12,50 | 72,50 | G3 |
| B | 4,50 | 50,50 | G3 |
| C | 4,50 | 8,50 | G3 |
| D | 24,50 | 0,50 | G4 |
| E | 12,50 | 2,50 | G4 |
| E | 26,50 | 0,50 | G4 |

## Calcul des coordonnées des nouveaux centres

| | V1 | V2 | d(xi,G5) |
|---|---|---|---|
| A | 2,00 | 1,00 | 5,56 |
| B | 3,00 | 2,00 | 0,89 |
| C | 6,00 | 5,00 | 10,89 |
| C. Grav G5 | 3,67 | 2,67 | |
| Iw | 17,33 | | |

| | V1 | V2 | d(xi,G6) |
|---|---|---|---|
| D | 8,00 | 7,00 | 0,44 |
| E | 7,00 | 6,00 | 1,11 |
| E | 9,00 | 6,00 | 1,11 |
| C. Grav G6 | 8,00 | 6,33 | |
| Iw | 2,67 | | |



## Etape 3

→ Calcul des distances entre chaque individu et chaque centre (étape 2)

| | V1 | V2 |
|---|---|---|
| A | 2,00 | 1,00 |
| B | 3,00 | 2,00 |
| C | 6,00 | 5,00 |
| D | 8,00 | 7,00 |
| E | 7,00 | 6,00 |
| F | 9,00 | 6,00 |
| G5 | 3,67 | 2,67 |
| G6 | 8,00 | 6,33 |

| | V1 | V2 | d(xi,G5) |
|---|---|---|---|
| A | 2,00 | 1,00 | 5,58 |
| B | 3,00 | 2,00 | 0,90 |
| C | 6,00 | 5,00 | 10,86 |
| D | 8,00 | 7,00 | 37,50 |
| E | 7,00 | 6,00 | 22,18 |
| F | 9,00 | 6,00 | 39,50 |
| G5 | 3,67 | 2,67 | |

| | V1 | V2 | d(xi,G6) |
|---|---|---|---|
| A | 2,00 | 1,00 | 64,41 |
| B | 3,00 | 2,00 | 43,75 |
| C | 6,00 | 5,00 | 5,77 |
| D | 8,00 | 7,00 | 0,45 |
| E | 7,00 | 6,00 | 1,11 |
| E | 9,00 | 6,00 | 1,11 |
| G6 | 8,00 | 6,33 | |

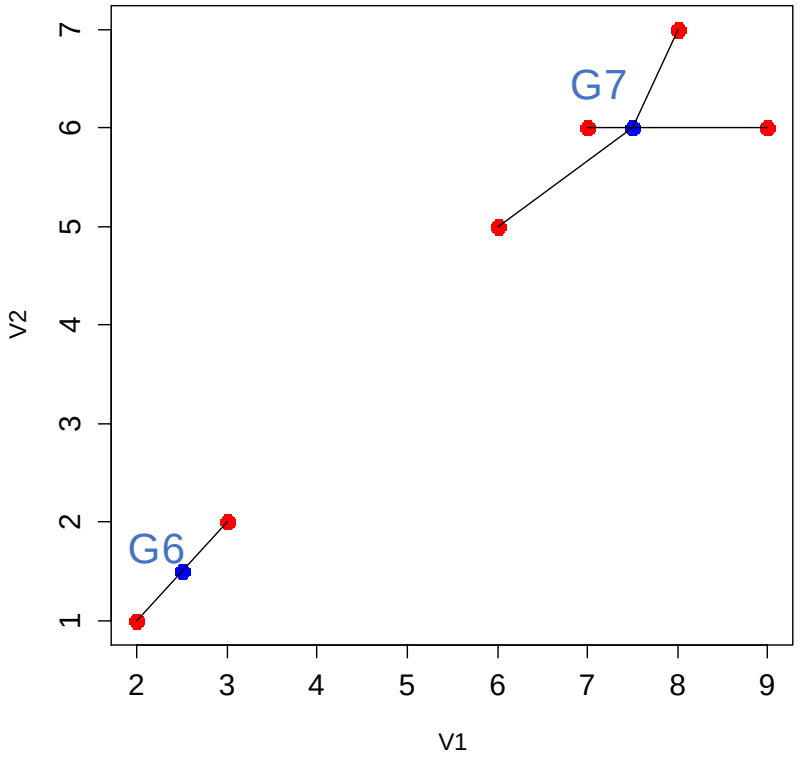
→ Affectation des individus au centre le plus proche

| | d(xi,G5) | d(xi,G6) | Affectation |
|---|---|---|---|
| A | 5,58 | 64,41 | G5 |
| B | 0,90 | 43,75 | G5 |
| C | 10,86 | 5,77 | G6 |
| D | 37,50 | 0,45 | G6 |
| E | 22,18 | 1,11 | G6 |
| F | 39,50 | 1,11 | G6 |

## Calcul des coordonnées des nouveaux centres

| | V1 | V2 | d(xi,G8) |
|---|---|---|---|
| C | 6,00 | 5,00 | 3,25 |
| D | 8,00 | 7,00 | 1,25 |
| E | 7,00 | 6,00 | 0,25 |
| F | 9,00 | 6,00 | 2,25 |
| C. Grav G6 | 7,50 | 6,00 | |
| Iw | 7,00 | | |

| | V1 | V2 | d(xi,G7) |
|---|---|---|---|
| A | 2,00 | 1,00 | 0,50 |
| B | 3,00 | 2,00 | 0,50 |
| C. Grav G7 | 2,50 | 1,50 | |
| Iw | 1,00 | | |




| | Iw (intra) | Ib (inter) | $R^2$ |
|---|---|---|---|
| Itération 1 | 35,00 | 33,33 | 0,49 |
| Itération 2 | 20,00 | 48,33 | 0,71 |
| Itération 3 | 8,00 | 60,33 | 0,88 |

```
classe.kmean <- kmeans(df.stand,centers = 5)
#--- Afichage des résultats
        print(classe.kmean)
#-- Correspondance entre CAH et kmean
        tab.sim <- table(gp.cah,classe.kmean$cluster)
         print(tab.sim)
 # -> Le fonction kmean ne fournit pas de d'outils d'aide à la détection du nombre de classes
 # il faut donc la programmer en utilisant les inerties
 inertie.expl <- rep(0,10)
 for (k in 2:10)
          {      clus <- kmeans(df.stand,centers=k)
                 inertie.expl[k] <- clus$betweenss/clus$totss
           }
 plot(inertie.expl, type = 'b', ylab = 'delta intertie')
```

**Cluster means:**

| | calories | sodium | calcium | lipides | retinol | folates | proteines | cholesterol | magnesium |
|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.2560480 | 0.2338429 | -0.8070507 | -0.1640442 | 1.9632388 | 1.7946811 | -0.31882661 | -0.3393858 | -0.2399208 |
| 2 | 0.8395372 | -0.7332260 | 1.2856329 | 0.6521049 | -0.1242419 | -0.8436457 | 1.28610740 | 0.9705456 | 1.6287198 |
| 3 | -2.1572744 | -1.5213272 | -0.7167418 | -2.1998041 | -0.5136787 | 0.2955348 | -1.86341394 | -1.9945017 | -1.3884943 |
| 4 | 0.3858540 | 0.2145199 | 0.3741340 | 0.3937909 | -0.3940748 | -0.6346609 | 0.25877721 | 0.2978781 | 0.1620799 |
| 5 | 0.2661773 | 1.1118113 | -0.6790557 | 0.3761600 | -0.2060471 | 0.3673290 | -0.04822449 | 0.2506734 | -0.5417895 |

Clustering vector:

| CarredelEst | Babybel | Beaufort | Bleu | Camembert | Cantal |
|---|---|---|---|---|---|
| 5 | 4 | 2 | 5 | 1 | 4 |
| **Chabichou** | **Chaource** | **Cheddar** | **Comte** | **Coulomniers** | **Edam** |
| 1 | 1 | 4 | 2 | 5 | 2 |
| **Emmental** | **Fr.chevrepatemolle** | **Fr.fondu.45** | **Fr.frais20nat.** | **Fr.frais40nat.** | **Maroilles** |
| 2 | 1 | 5 | 3 | 3 | 4 |
| **Morbier** | **Parmesan** | **Petitsuisse40** | **PontlEveque** | **Pyrenees** | **Reblochon** |
| 4 | 2 | 3 | 4 | 4 | 4 |
| **Rocquefort** | **SaintPaulin** | **Tome** | **Vacherin** | **Yaourtlaitent.nat.** | |
| 5 | 4 | 5 | 4 | 3 | |

**Within cluster sum of squares by cluster:**

[1] 13.432538 9.871039 6.446342 12.832448 15.166537

(between_SS / total_SS = 77.1 %)

## CONCLUSION

- **Classification (CAH)**
  - ✚ Avantages
    - Pas de dépendance au choix de centres initiaux
    - Pas de fixation à priori du nombre de classes
    - Détecte des classes de forme diverse
  - ▬ Inconvénients
    - Complexité de l'algorithme
    - A chaque étape , le partitionnement n'est pas global, il dépend des classes précédentes
    - ⊗ Variation importante du dendrogramme en fonction du critère d'agrégation
- **Partitionnement (moyennes mobiles)**
  - ✚ Avantages
    - Rapide et facile à mettre en oeuvre
    - Applicable à de grands volumes de données
    - Permet de détecter facilement des individus hors norme
    - Amélioration continue de la qualité des classes
  - ▬ Inconvénients
    - Partition initiale dépend fortement des choix initiaux
    - Nombre de classes fixées à priori

41

Toutes les méthodes de classification sont sensibles aux points aberrants (très éloignés des autres). Il est donc nécessaire de les repérer, de les isoler ou de les omettre avant d'appliquer une méthode de classification

Avant analyse , TOUJOURS CENTRER LES DONNEES !!