# CLASSIFICATION

## PLAN

## La classification automatique
## (Automatic clustering)

**Définition** :

Création et caractérisation de groupes disjoints d'individus

- homogènes.
- bien séparés les uns des autres.

La classification est *automatique* :

elle est faite automatiquement à partir des valeurs des variables observées en utilisant un algorithme.

C'est une méthode *multidimensionnelle*, elle tient compte simultanément d'un ensemble de variables.



**Terminologie**

On parle de classification, de taxinomie ou taxonomie (rare).
La taxinomie est la science des classifications (des espèces vivantes, à l'origine). On suit, indifféremment, des cours de taxinomie ou de classification ; par contre, on ne peut pas faire la taxinomie des délits, on en fait une classification.
Le terme anglais de *classification* se traduit en français par *classement*.

Faire remarquer qu'on peut classifier individus ou variables.
Groupe disjoints : pas d'individus communs à 2 groupes.

**Classification automatique**

Il existe des algorithmes de classification. La classification n'est pas faite au jugé, elle découle d'un certain nombre de choix effectués *a priori*.

**C'est une méthode multidimensionnelle**

La classification automatique tient compte d'un ensemble de variables.
La « *segmentation* » est une méthode qui ressemble à la classification hiérarchique descendante, puisqu'on scinde à chaque itération la population (puis les classes obtenues par les scissions antérieures) en 2 et qu'on obtient un arbre qu'on peut couper à un endroit judicieux comme en classification hiérarchique. Elle *considère les variables explicatives l'une après l'autre et pas toutes ensemble*. De plus, la segmentation cherche en général à expliquer une variable par d'autres variables.

## La classification pour quoi faire ?

L'utilisateur peut chercher à

- confronter un modèle théorique de données avec des observations ou encore
- partitionner une population dans un but simplificateur et descriptif.

La classification

- est souvent complémentaire d'une analyse factorielle.
- peut permettre de repérer les points aberrants (classification descendante …).
- doit être suivie d'une phase d'interprétation (description des classes) et si possible d'une analyse de stabilité.



**Utilité de la classification :**

- complémentarité avec une analyse factorielle (permet de décrire le comportement de groupes d'individus et non pas d'individus isolés).

- repérer les points aberrants (ceux qui sont difficiles à classifier $\rightarrow$ ceux qui donnent des classes singleton) par les méthodes classification hiérarchique ascendante avec critère du saut minimal ou classification hiérarchique descendante.

  Pour les points aberrants, même si la classification hiérarchique descendante est efficace, il faut noter que les algorithmes n'existent pas dans des logiciels Splus ou SAS. Voir alors nuées dynamiques ou saut minimal, éventuellement.
- Exemple d'analyse de stabilité : forme forte, bootstrap.

## Classification, classement ?

### CLASSIFICATION

| QUESTIONS | RÉPONSES |
|---|---|
| **Existe-t-il des groupes et combien** dans la population étudiée ? | Il semble qu'on puisse distinguer trois groupes (ou classes). |
| **Caractéristiques** de ces classes ? | Classe 1 : plutôt jeunes et indemnes (60 %).<br>Classe 2 : plutôt peu touchés (35 %).<br>Classe 3 : principalement personnes âgées plutôt gravement touchées (5 %). |

### CLASSEMENT

| QUESTION | RÉPONSE |
|---|---|
| Une classification étant effectuée, **où** mettre ce nouvel individu ? | Dans la classe 3. |



**Classification n'est pas classement**

Des questions différentes ⇒ des méthodes différentes.

Classification = constitution et caractérisation de classes homogènes à partir d'un tableau de données. Regrouper les individus pour avoir homogénéité à l'intérieur des classes.
Classement = les classes existent déjà, on se contente d'affecter les individus à ces classes préexistantes. Une règle de classement peut être le résultat d'une classification.

**Exemple :**

Tableau de 100 individus et 5 variables (degré de gravité de 5 maladies).
La *classification* consiste *à créer un certain nombre de classes*, inconnu *a priori*, et de les caractériser.
On obtient par exemple 3 classes (personnes jeunes et indemnes, personnes un peu touchés, personnes âgées plutôt gravement touchées).
Le *classement* consiste à *affecter à ces 3 classes* d'autres individus que ceux qui ont servi à la classification, en utilisant les valeurs des variables de ces nouveaux individus.
Une règle de classement peut-être obtenue par une analyse discriminante ou par une régression logistique ou par une méthode de segmentation (où il y a une variable privilégiée par rapport aux autres : la variable nominale à expliquer).

## Classification, classement ?

| Ensemble de variables | Une variable de groupe et un ensemble de variables explicatives |
|---|---|
| ⇓ | ⇓ |
| **Classification automatique** | **Analyse discriminante**<br>**Segmentation**<br>**Discrimination logistique** |
| ⇓ | |
| Définition de classes | |
| ⇓ | ⇓ |

# Règles de classement

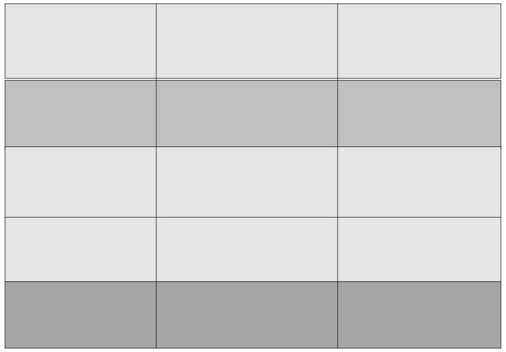


Règle de classement : en général, calcul des distances entre 1 individu et les centres des classes → affectation à la classe la plus proche.

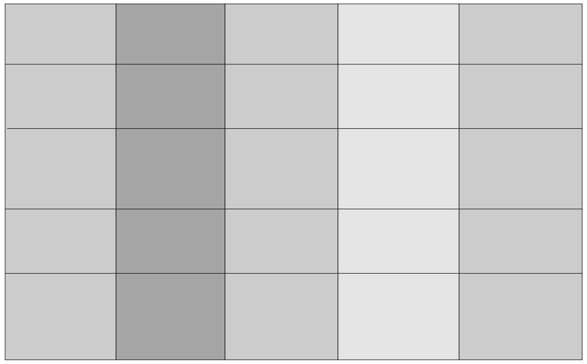
## Que peut-on classifier?

- individus ou variables

individus



variables



## Que peut-on classifier?

Des données issues de différents tableaux :

**Données quantitatives**

| | taille | poids |
|---|---|---|
| individu 1 | 130 | 40 |
| individu 2 | 110 | 44 |
| individu 3 | 150 | 56 |
| individu 4 | 195 | 87 |
| individu 5 | 122 | 29 |
| individu 6 | 154 | 75 |
| individu 7 | 201 | 76 |
| individu 8 | 154 | 65 |

**Données quantitatives et ordinales**

| | taille | poids | classe d'âge |
|---|---|---|---|
| individu 1 | 130 | 40 | 1 |
| individu 2 | 110 | 44 | 1 |
| individu 3 | 150 | 56 | 2 |
| individu 4 | 195 | 87 | 3 |
| individu 5 | 122 | 29 | 1 |
| individu 6 | 154 | 75 | 3 |
| individu 7 | 201 | 76 | 3 |
| individu 8 | 154 | 65 | 2 |

cf. ACP

**Remarque** : comme les unités sont différentes, il faut réduire les variables avant de faire la classification.



**Types de tableaux**

On peut classer des individus ou des variables.

Les données peuvent être issues de plusieurs types de tableaux :

- tableaux de données quantitatives, c'est le cas le plus courant.
- tableaux de données quantitatives et qualitatives. Attention : une variable qualitative ne peut être traitée comme une variable quantitative que si elle est ordonnée.

## Que peut-on classifier?

### tableaux de contingence

| | | ouvriers | cadres | autres |
|---|---|---|---|---|
| type | **A** | 1128 | 35 | 0 |
| | **B** | 1652 | 560 | 3 |
| de | **C** | 52 | 1392 | 1005 |
| | **D** | 2 | 2263 | 2635 |
| maladie | **E** | 0 | 65 | 95 |

cf. AFC

### tableaux de présence-absence

| | mâle | femelle | $< 50$ ans | $\geq 50$ ans |
|---|---|---|---|---|
| individu 1 | 1 | 0 | 1 | 0 |
| individu 2 | 1 | 0 | 0 | 1 |
| individu 3 | 0 | 1 | 1 | 0 |
| individu 4 | 0 | 1 | 0 | 1 |
| individu 5 | 1 | 0 | 0 | 1 |

cf.ACM

Remarque : dans ce cas, utilisation de la métrique du Khi-deux sur le tableau transformé des profils lignes.

### tableaux de distances[1]

| | Individu 1 | Individu 2 | Individu 3 | Individu 4 | Individu 5 |
|---|---|---|---|---|---|
| Individu 1 | 0 | 4 | 1 | 3 | 3.16 |
| Individu 2 | 4 | 0 | 4.12 | 5 | 1.41 |
| Individu 3 | 1 | 4.12 | 0 | 2 | 3 |
| Individu 4 | 3 | 5 | 2 | 0 | 3.6 |
| Individu 5 | 3.16 | 1.41 | 3 | 3.6 | 0 |

[1]Ne pas confondre le tableau de distances entre individus et la matrice $M$ qui définit la distance.



### Types de tableaux (suite)

- tableaux de contingence : ici on regroupe les lignes. La distance entre les lignes est mesurée par la distance du $\chi^2$ sur les profils lignes (cf. analyse des correspondances).
  Dans l'exemple, on regroupera les types de maladies.
- tableaux de données catégorielles ou présence/absence (cf. analyse des correspondances multiples).
- tableaux de distances, de similarités (le carré $R^2$ de la matrice de corrélation est une matrice de similarité) ou de dissimilarités ($E - R^2$, avec $E$ une matrice de 1, est une matrice de dissimilarité).
- Le tableau de distance est symétrique. On peut le trouver sous la forme d'une demi-matrice supérieure ou inférieure.

### Tableau de distances

Les algorithmes de classification opèrent soit sur les tableaux individus x variables (ex. nuées dynamiques, forme classique), soit directement sur les tableaux de distances. Dans le cas de gros échantillons, il faut éviter de générer des tableaux de distance.

Ce sont souvent des tableaux de distance euclidienne - distance euclidienne « *classique* » (encore appelée distance de l'arpenteur) ou distance euclidienne du Khi-deux qui sont utilisées. Mais on peut aussi utiliser des indices de similarité, c'est le cas en taxonomie bactérienne. (Ex. traité dans la journée de TP).

## Codage

- Centrer ne modifie pas les distances entre individus : ça ne modifiera pas la classification.
- Réduire uniformise l'influence des variables dans la classification.
- De plus, il peut être nécessaire d'effectuer des transformations de variables (log, ... ) de façon à adapter le calcul de la distance entre individus à la représentation qu'en a l'utilisateur.



**À propos du centrage et de la réduction**

C'est un problème très important et qui ne doit pas être pris à la légère sans examen antérieur des données.

La réduction d'une variable contenant une donnée aberrante écrasera les différences entre les individus *normaux* et accentuera l'aberrance de l'individu *anormal*. Il faut donc une analyse antérieure à la standardisation pour repérer ces points aberrants (*outliers* en Anglais).

D'autre part, il n'y a pas toujours lieu de standardiser : une différence de 0.1 point sur un pourcentage de 0.5 % est-il aussi important qu'une différence de 5 points sur un pourcentage de 50 % (cas de l'analyse d'un vote dans deux régions) ?

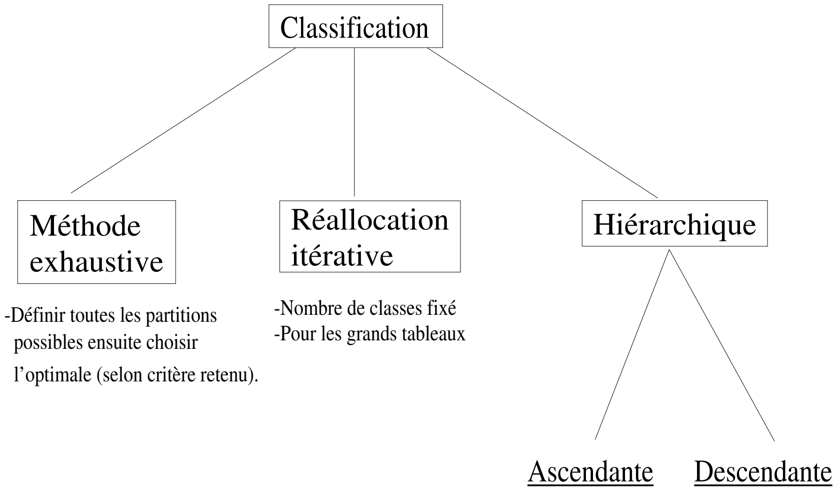
Plusieurs façons de réduire :

- écart-type,
- intervalle interquartile, valeurs absolues des écarts,
- amplitude, valeurs maximales,
- moyenne, médiane,
- valeur de référence ...

Le but est de rendre les données a-dimensionnelles (indépendantes de l'unité). En général, on utilise les paramètres de dispersion (écart-type, intervalle interquartile).

Le passage par tableau de distance est facultatif. La classification peut opérer directement sur les données recodées avec une notion de distance.

## Méthodes de classification




**Remarque** : il est possible d'utiliser différentes méthodes successivement (classification mixte).

Autre méthode : modélisation de classes.



– Il faut d'abord choisir un critère mesurant la qualité d'une partition.

– Si l'on a $n$ individus à classifier, la partition en $n-1$ classes obtenue sera la meilleure possible, mais rien ne permet d'indiquer que la meilleure partition en $n-2$ classes dérive de cette partition en $n-1$ classes (par analogie, une régression pas à pas atteint elle aussi un optimum local).

  Comment trouver la meilleure partition en $k$ classes, avec $k < n$ ? Ainsi pour $n = 12$ et $k = 3$, il y a 86526 partitions possibles.

  Formule de calcul : $\frac{1}{k!} \times \sum_{i}^{k} {C_k}^i \times (-1)^{k-i} \times i^n$ avec $k$ classes et $n$ individus.

  Les différentes méthodes de classification automatique sont les suivantes :

  – classification exhaustive (en pratique souvent impossible),
  – réallocation itérative autour de centres mobiles (*nuées dynamiques*),
  – regroupements successifs en partant de la partition discrète (1 classe = 1 singleton) jusqu'à l'obtention d'une classe unique (méthode hiérarchique ascendante) ou inversement à partir d'une classe unique, divisions successives ... (méthode hiérarchique descendante).

    Certaines méthodes hiérarchiques admettent une modélisation *a priori* des classes.

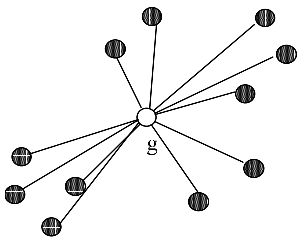
  Seules une méthode de réallocation itérative et la classification hiérarchique *ascendante* seront abordées ici, en ajoutant la combinaison des 2, qui cumule les avantages respectifs de ces deux méthodes. C'est la *classification mixte*.

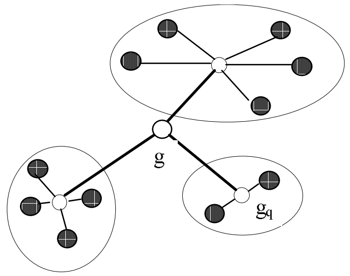
## Un critère de qualité d'une partition : l'inertie inter

Inertie totale = inertie intra + inertie inter

Inertie totale : $I_{totale} = \sum_i m_i \|x_i - g\|^2$ Inertie intra : $I_{intra} = \sum_q \sum_{i \in G_q} p_i \|x_i - g_q\|^2$

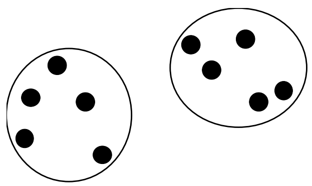
Inertie inter : $I_{inter} = \sum_q \pi_q \|g_q - g\|^2$



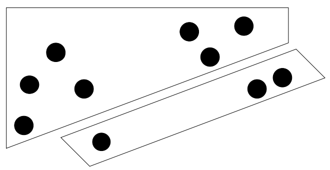

Le plus souvent, le critère retenu en classification automatique cherche à minimiser l'inertie intra (et donc à maximiser l'inertie inter).

inertie inter forte



Inertie intra-classes faible

inertie inter faible



Inertie intra-classes élevée



$g$ centre de gravité global ($g = \sum_i p_i x_i$ et $\sum_i p_i = 1$)

$q$ numéro du groupe

$G_q$ ensemble des indices du groupe $q$

$g_q$ barycentre du groupe $q$ ($g_q = \frac{1}{\pi_q} \sum_{i \in G_q} p_i x_i$) avec $\pi_q = \sum_{i \in G_q} p_i$

Remarquons que $g = \sum_q \pi_q g_q$

On remplace le nuage des individus par le nuage des barycentres, le nuage se contracte et perd de l'inertie. *L'inertie intra* est l'inertie perdue par le regroupement, *l'inertie inter* est l'inertie conservée ; la qualité de la partition peut être mesurée par le pourcentage d'inertie conservée (ne permet de comparer que des partitions de même nombre de classes).

Faire remarquer que ce sont les mêmes points dans les deux cas.

## Classification type réallocation itérative : nuées dynamiques

**Critère :**

minimiser l'inertie intra ou maximiser l'inertie inter.

**Démarche :**

1. choix *a priori* du nombre de classes et de *noyaux* initiaux.
2. les individus changent de classe d'une itération à l'autre.
3. après réaffectation, on recalcule les nouveaux noyaux.
4. l'algorithme s'arrête lorsque 2 partitions successives sont identiques.

**Remarque :**

- méthode rapide, pratique pour les gros échantillons.
- le résultat dépend du choix initial des noyaux.

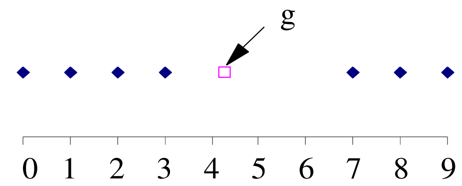


**Quelques propriétés**

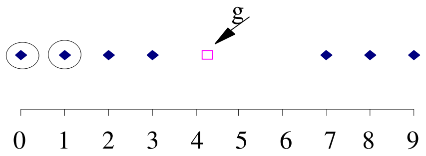
- Le critère est compatible avec celui de l'analyse factorielle : maximiser l'inertie du nuage résumé.
- Rapidité : au bout de 5 ou 6 itérations, il y a généralement stabilité des classes, ce qui est intéressant pour les gros échantillons.
- Économie : ne nécessite pas le calcul des distances inter points.
- Défaut : on ne sait pas trouver facilement le bon nombre de classes.

## Classification type réallocation itérative : nuées dynamiques

### Critère : minimiser l'inertie intra

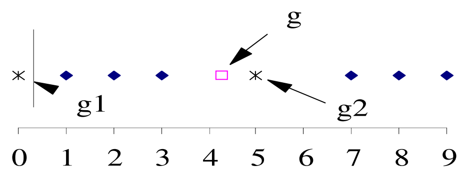



7 individus,
g : centre de gravité,
Objectif : obtenir 2 classes.

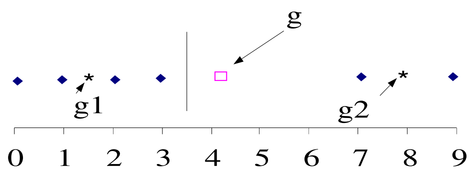



Choix des noyaux,
g : centre de gravité, coordonnée 4.286
Inertie totale : 11.35

Calcul des inerties intra (et inter) à partir des centres de gravité des 2 classes :




g1, g2 : centre de gravité des 2 classes Inertie intra : 8.29
(Iinter : 3.06)




g1, g2 : centre de gravité des 2 classes Inertie intra : 1
(Iinter : 10.35)

Stabilisation après minimisation de l'inertie intra (minimum local).



Petit exemple illustratif. Il est assez curieux car il n'y a qu'une seule variable. C'est la distance usuelle qui est retenue, elle permet de visualiser les notions d'individus proches.
L'objectif est de faire calculer les inerties inter et intra.

Dans l'exemple présenté ici, le "choix" des noyaux initiaux est assez curieux (!), ce sont les 2 premiers individus. Faire calculer éventuellement au stagiaire l'inertie totale : $(18.4 + 10.8 + \ldots + 22.2)/7$.
Sur le 2ème graphique sont portés les 2 centres de gravité des 2 classes (coordonnées 0 et 5). Calcul de l'inertie intra et inter.
$I_{inter} = (0 + 16 + 9 + 4 + 4 + 9 + 16)/7$ et $I_{intra} = (18.37 + 0.51 \times 6)/7$.
Bien vérifier que $I_{totale} = I_{inter} + I_{intra}$.
Ensuite, sont reportés les 2 étapes nécessaires à la stabilisation.

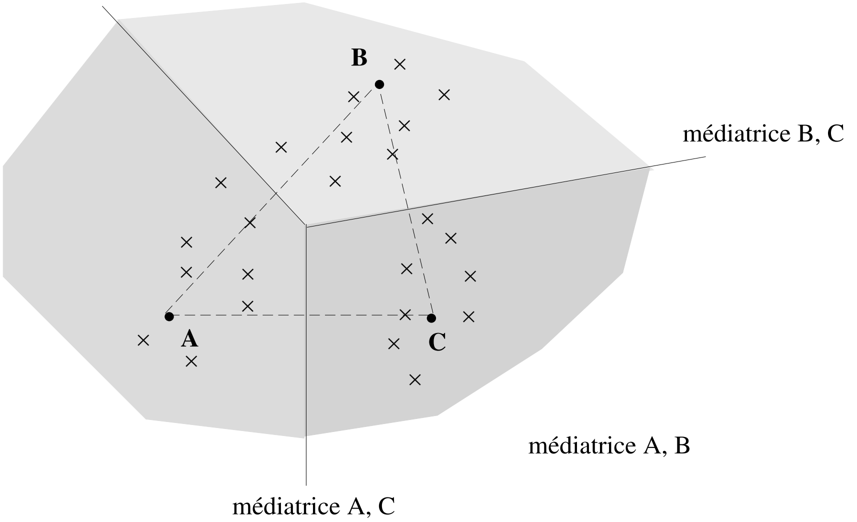
**Remarque :** à partir d'une partition de départ, on minimise l'inertie intra. On obtient donc en général seulement un minimum local !

**Exercice :** réaliser la seconde étape ou choisir deux nouveaux noyaux et effectuer l'algorithme entier.

## Classification type réallocation itérative : nuées dynamiques

### Description d'une étape

**Phase 2 :** réaffectation

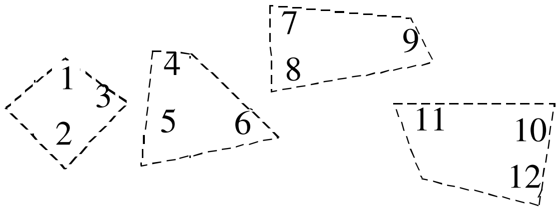



À chaque étape, les médiatrices ou hyperplans médiateurs délimitent les régions des points les plus proches des noyaux A, B et C.

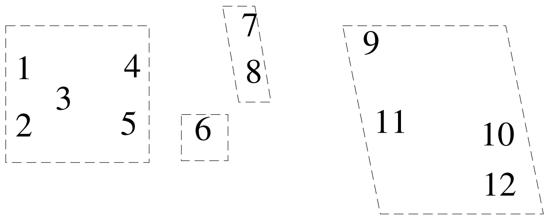
**Phase 3 :** Calcul des noyaux

Après réaffectation, les nouveaux noyaux A, B et C sont les barycentres des nouvelles classes.

## Classification type réallocation itérative : nuées dynamiques




Partition 1 avec 4 noyaux.




Partition 2 avec 4 noyaux.

**Remarques** : il peut y avoir plusieurs partitions vraisemblables.

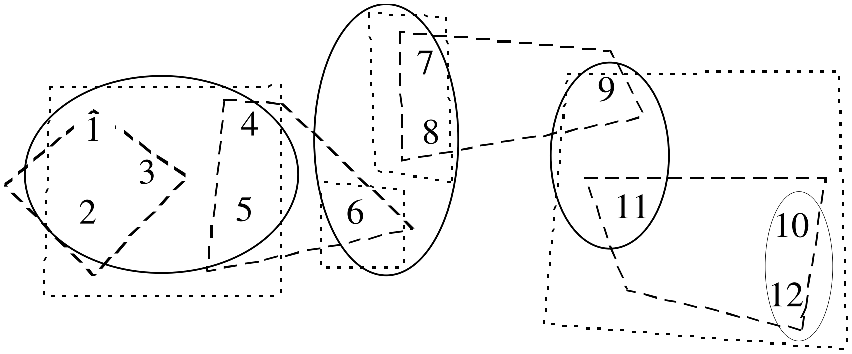
- On peut imposer *a priori* le choix des noyaux.
- On peut utiliser les formes fortes.



En fonction du choix des noyaux initiaux, on obtiendra des classifications différentes.

**Classification type réallocation itérative : nuées dynamiques**

**Formes fortes**




| | Partition 1 | Partition 2 | Partition 3 | F.Forte |
|---|---|---|---|---|
| 1 | A | $\alpha$ | a | I |
| 2 | A | $\alpha$ | a | I |
| 3 | A | $\alpha$ | a | I |
| 4 | B | $\alpha$ | a | II |
| 5 | B | $\alpha$ | a | II |
| 6 | B | $\beta$ | b | III |
| 7 | C | $\gamma$ | b | IV |
| 8 | C | $\gamma$ | b | IV |
| 9 | C | $\delta$ | c | V |
| 10 | D | $\delta$ | d | VI |
| 11 | D | $\delta$ | c | VII |
| 12 | D | $\delta$ | d | VI |



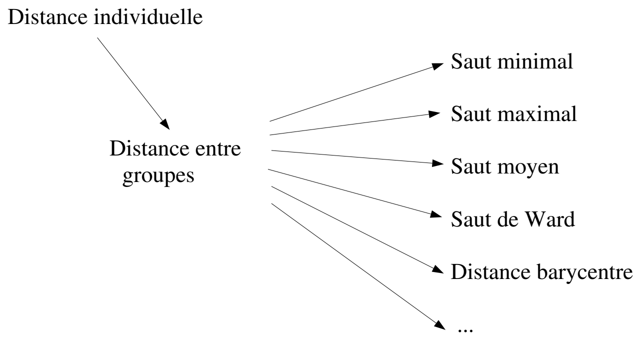
*Les formes fortes* sont obtenues par superposition (ou intersection) des différentes partitions.

Une réponse possible est de faire plusieurs classifications et de ne retenir que les individus toujours présents dans les mêmes classes. On pourra ne retenir que les formes fortes ayant plus de $x$ individus en laissant de côté les formes fortes de faible effectif. Les classes formées par les formes fortes retenues sont plus faciles à caractériser. Les individus écartés pourront être reclassés dans les formes fortes les plus proches ou mis dans les classes rebut.

Dans cet exemple, il y a 7 formes fortes. Cela correspond aux individus qui n'ont été séparés par aucune des partitions. Les formes fortes singleton ne présentent pas d'intérêt réel.
Au total, il y 611501 façons de répartir 12 individus dans 4 classes.

## CLASSIFICATION HIERARCHIQUE

- Le nombre de classes n'est pas fixé *a priori*.
- Création d'une hiérarchie de partitions (arbre ou dendrogramme).
- Il faut calculer une distance entre individus et une distance entre groupes.



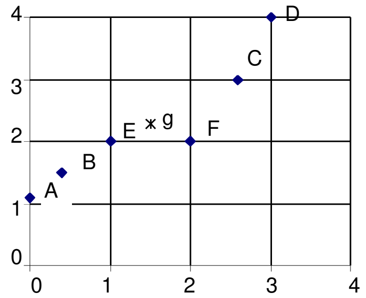

### Démarche

1. On part de la partition discrète (un groupe = un singleton).
2. On regroupe les 2 classes les plus proches (constitution d'une nouvelle classe).
3. On recalcule la distance de cette classe avec les autres (selon la distance entre groupes choisie).

On itère 2 et 3 jusqu'à constitution d'une classe unique.

## CLASSIFICATION HIERARCHIQUE

Exemple avec 6 individus et 2 variables :




| | **Y1** | **Y2** | **(Y1-Y1g)$^2$** | **(Y2-Y2g)$^2$** |
|---|---|---|---|---|
| ind A | 0 | 1.1 | 2.25 | 1.36 |
| ind B | 0.4 | 1.5 | 1.21 | 0.59 |
| ind C | 2.6 | 3 | 1.21 | 0.54 |
| ind D | 3 | 4 | 2.25 | 3.00 |
| ind E | 1 | 2 | 0.25 | 0.07 |
| ind F | 2 | 2 | 0.25 | 0.07 |
| **Moyenne** | **1.5** | **2.27** | | |

**Inertie totale : 2.176**

| Matrice des distances euclidiennes | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| 0.00 | 0.57 | 3.22 | 4.17 | 1.35 | 2.19 |
| 0.57 | 0.00 | 2.66 | 3.61 | 0.78 | 1.68 |
| 3.22 | 2.66 | 0.00 | 1.08 | 1.89 | 1.17 |
| 4.17 | 3.61 | 1.08 | 0.00 | 2.83 | 2.24 |
| 1.35 | 0.78 | 1.89 | 2.83 | 0.00 | 1.00 |
| 2.19 | 1.68 | 1.17 | 2.24 | 1.00 | 0.00 |



**Exemple de classification hiérarchique.**

Le but de l'exemple est de :

1. revenir sur les notions d'inertie et de distance entre individus,
2. construire un arbre pas à pas avec les stagiaires,
3. montrer que l'arbre final dépendra de la distance entre groupes retenue.

*Nous reviendrons sur ces 3 points dans les transparents suivants.*

**1) Revenir sur les notions d'inertie et de distance entre individus**

- Tableau avec 6 individus et 2 variables.
- Graphe avec les 6 individus et leur centre de gravité, $g$.
- Détailler le calcul de l'inertie totale. Montrer que c'est le carré moyen des écarts au point $g$ (à illustrer avec le premier graphique).
- Revenir sur le calcul de la distance (à illustrer avec le premier graphique).

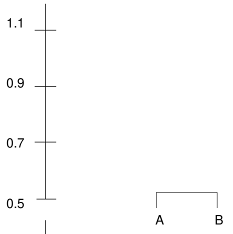
Inertie totale $= (2.25 + 1.21 + \ldots + 1.36 + 0.59 + \ldots)/6$
ce qui correspond au transparent 12 avec $p_i = 1/6$.

## CLASSIFICATION HIERARCHIQUE

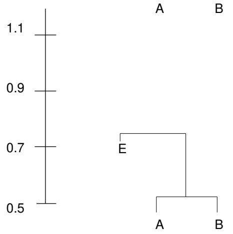
### Distance entre groupe : ex. du saut minimal

Définition de la stratégie d'agrégation :

$$d(\{x,y\},z) = min\{d(x,z), d(y,z)\}$$

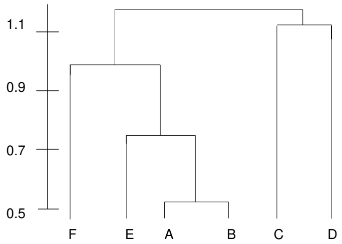



| | ind A | ind B | ind C | ind D | ind E | ind F |
|---|---|---|---|---|---|---|
| A | 0 | | | | | |
| B | 0.56 | 0 | | | | |
| C | 3.20 | 2.66 | 0 | | | |
| D | 4.20 | 3.60 | 1.08 | 0 | | |
| E | 1.34 | 0.78 | 1.89 | 2.80 | 0 | |
| F | 2.20 | 1.67 | 1.17 | 2.20 | 1 | 0 |




| | cl 1 | ind C | ind D | ind E | ind F |
|---|---|---|---|---|---|
| cl 1 | 0 | | | | |
| C | 2.66 | 0 | | | |
| D | 3.60 | 1.08 | 0 | | |
| E | 0.78 | 1.89 | 2.80 | 0 | |
| F | 1.67 | 1.17 | 2.20 | 1 | 0 |

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .




| | cl 3 | cl 4 |
|---|---|---|
| cl 3 | 0 | |
| cl 4 | 1.17 | 0 |



**2) Construire un arbre pas à pas avec les stagiaires :**

Exemple dans le plan avec la distance euclidienne classique (même exemple que précédemment) :

distance entre groupes : on choisit une métrique particulière, celle du saut minimal ("single linkage" ou "connected" en Anglais).

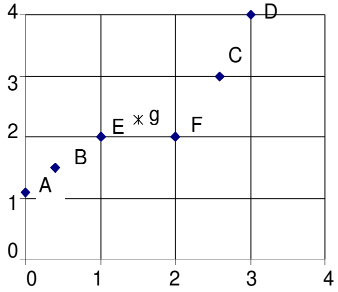
Stratégie d'agrégation du saut minimal :
x, y et z sont trois classes.
Si x et y sont regroupés en une classe h, alors :
$d(\{x,y\},z) = min\{d(x,z), d(y,z)\}$.

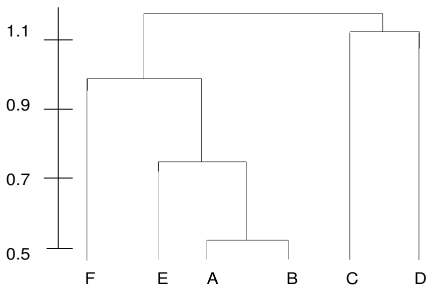
Sur le $2^{ème}$ tableau, le noeud 1 a remplacé les individus A et B. Les distances entre les points C, D, E, F entre eux ne sont pas modifiées. La distance entre un point et le noeud 1 est le minimum des distances entre ce point et les points A et B.

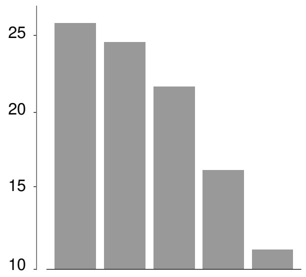
Faire remarquer que 2 points "proches" (notion de distance entre individus) ne seront pas forcément plus rapidement unis dans un arbre de classification que 2 points plus éloignés. Leur réunion dépend en effet de la stratégie d'agrégation (ici saut minimal) mais aussi de la réunion antérieure d'autres points. On verra d'autres stratégies plus loin.

On a construit ce qu'on appelle un arbre hiérarchique ou dendrogramme indicé.

## CLASSIFICATION HIERARCHIQUE




On choisit généralement le nombre de classes à l'aide de la règle du coude.



## CLASSIFICATION HIERARCHIQUE ASCENDANTE (CAH)

### Philosophie (et un peu de vocabulaire)

Il y a constitution d'un *arbre hiérarchique* ou *dendrogramme* (du Grec dendron = arbre et gramma = dessin). Les 2 individus les plus proches sont agrégés en 1 classe, puis c'est le tour des 2 individus (ou classes) suivants les plus proches, etc.

On coupera l'arbre et on indiquera la partition associée.

Une hiérarchie indicée est définie par un arbre hiérarchique indicé.

## CLASSIFICATION HIERARCHIQUE
## Distance entre groupes : saut de Ward

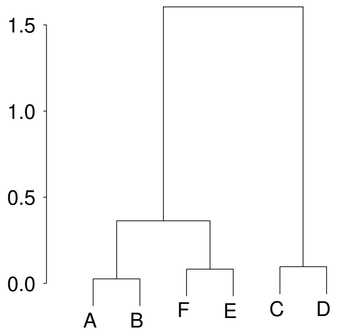
Soit $d$ la distance euclidienne définie entre deux individus, et $x$ et $y$ deux groupes d'individus :

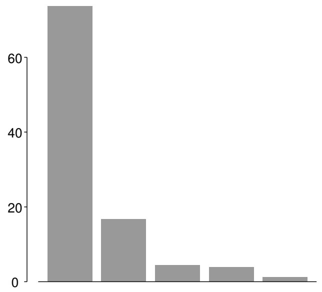
- de barycentres $\mathbf{g}_x$ et $\mathbf{g}_y$,
- de poids $p_x$ et $p_y$

la perte d'inertie par regroupement de deux groupes $x$ et $y$ s'écrit :

$$\frac{p_x p_y}{p_x + p_y} d^2(g_x, g_y)$$

| | Inertie inter | Inertie intra | Perte d'inertie |
|---|---|---|---|
| Avant tout regroupement | 2.176 | 0 | |
| Regroupement de A et B | 2.149 | 0.027 | 0.027 |
| Regroupement de E et F | 2.066 | 0.110 | 0.083 |
| Regroupement de C et D | 1.969 | 0.207 | 0.097 |
| **Regroupement de {A,B} et de {E,F}** | 1.606 | 0.570 | **0.363** |
| Regroupement de {A,B,E,F} et de {C,D} | 0 | 2.176 | 1.605 |
| Somme | | | 2.176 |






- Exemple de calcul de saut de Ward :
  $x = \{A, B\}$, $y = \{E, F\}$, $p_{\{A,B\}} = (1/6 + 1/6) = 1/3$, $p_{\{E,F\}} = 1/3$,

  $g_{\{A,B\}} = 1/2 \left( \begin{pmatrix} 0 \\ 1.1 \end{pmatrix} + \begin{pmatrix} 0.4 \\ 1.5 \end{pmatrix} \right) = \begin{pmatrix} 0.2 \\ 1.3 \end{pmatrix}$, $g_{\{E,F\}} = \begin{pmatrix} 1.5 \\ 2 \end{pmatrix}$,

  $d^2(g_{\{A,B\}}, g_{\{E,F\}}) = (1.5 - 0.2)^2 + (2 - 1.3)^2 = 2.18$,

  La perte d'inertie par regroupement de $\{A, B\}$ et $\{E, F\}$ est :

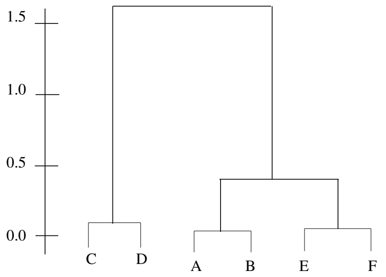
  $$\frac{p_{\{A,B\}} p_{\{E,F\}}}{p_{\{A,B\}} + p_{\{E,F\}}} d^2(g_{\{A,B\}}, g_{\{E,F\}}) = \frac{1/3 * 1/3}{(1/3) + (1/3)} * 2.18 = 2.18/6 = 0.363$$
- Dans le tableau, la somme intertie inter + inertie intra est constante et égale à l'inertie totale.
- Les indices de niveau des regroupements sont égaux aux pertes d'inertie. Ils sont aussi égaux à la hauteur des barres de l'éboulis des indices de niveau(remarquer la coïncidence entre les hauteurs des barres et les niveaux de regroupements).
- L'inertie totale est égale à la somme de toutes les pertes successives d'inertie quand on regroupe tout. C'est encore la surface de l'éboulis des valeurs propres.
- Quand on choisit $k$ classes, l'inertie conservée (ou inter inter) est représentée par la surface des $k-1$ premières barres de l'éboulis, l'inertie perdue (ou intra).

## CLASSIFICATION HIERARCHIQUE : Choix de la distance entre groupes

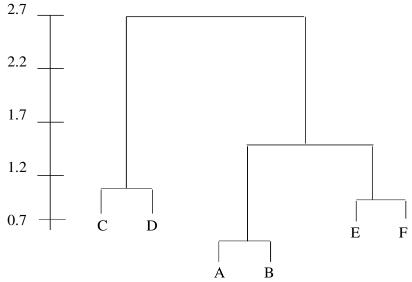
**Saut de Ward :**

**Inertie inter avec 3 groupes :**

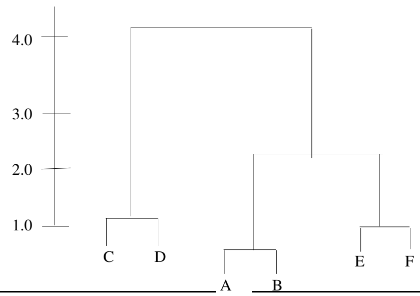
**1.97**




**Saut moyen :**




**Saut maximal :**






**3) montrer que l'arbre final dépendra de la distance entre groupes retenue :**

Présentation de 3 classifications hiérarchiques obtenues avec 3 distances entre groupes différentes : le saut de Ward, le saut moyen et le saut maximum donnent ici le même arbre (mais avec des indices différents), par contre ce sont des arbres différents de celui obtenu avec le saut minimal (transparent 20). Eventuellement, il est possible de construire avec les stagiaires l'arbre prenant le saut maximal comme critère d'agrégation.

On peut aussi faire recalculer les inerties inter associées aux trois arbres dans le cas où 3 classes sont demandées. (Inertie inter de 1.97 pour saut de Ward, saut moyen et saut maximal, et de 1.7 dans la cas du saut minimal). Pour cet exemple, les inerties inter pour le saut de Ward, saut moyen et saut maximal sont égales, mais ce n'est pas une généralité !
L'inertie inter n'est un "bon" critère que pour le saut de Ward, les autres critères ayant des objectifs différents (ex : saut maxi = groupe de faible diamètre).

**CLASSIFICATION HIERARCHIQUE :**
**Les différentes distances entre groupes, avantages et inconvénients**

| Critère | Caractéristiques |
|---|---|
| saut minimal | favorise le chaînage, peut être très sensible à la position d'un seul point. |
| saut maximal | supprime le chaînage, produit des classes compactes. Très sensible à la position d'un seul point. |
| saut moyen | tend à produire des classes de même variance. |
| centre de gravité | il est moins sensible aux points aberrants. |
| saut de Ward | compatible avec une analyse factorielle (même notion d'inertie). A tendance à faire des classes d'effectifs égaux. |

Dans la pratique, on utilise plutot le saut moyen et le saut de Ward.



L'effet de chaînage provoqué par l'agrégation avec le critère du saut minimal est dû au fait que la stratégie d'agrégation repose sur le lien entre 1 seul individu d'une classe A et 1 seul d'une classe B (on ne tient pas compte de l'ensemble des individus). Les individus d'une même classe ne sont pas forcément proches mais sont *reliés* entre eux par des individus intermédiaires.

Plus une classe est nombreuse, plus l'inertie intra augmente et donc plus la perte d'inertie inter est accentuée. C'est la raison pour laquelle le critère de Ward qui minimise la perte d'inertie inter a tendance à faire des classes d'effectifs égaux.

Eclaircir la notion de compatibilité entre analyse factorielle et classifications :

si l'on utilise la même distance inter points,

- l'acp ou l'afc détermine les sous espaces sur lesquels en projection la perte d'inertie est minimum,
- la classification avec saut de Ward, ou les algorithmes kmeans et forgy cherchent les regroupements qui minimisent la perte d'inertie (regroupement = contraction du nuage = perte d'inertie). Par regroupement on entend le fait de remplacer un groupe de points par leur barycentre en lui affectant un poids égal à la somme des poids.

*Une classif après afc sera proposée dans la journée TP.*

Proximité visuelle : distance euclidienne classique (arpenteur).

## CLASSIFICATION HIERARCHIQUE

**Caractéristiques :**

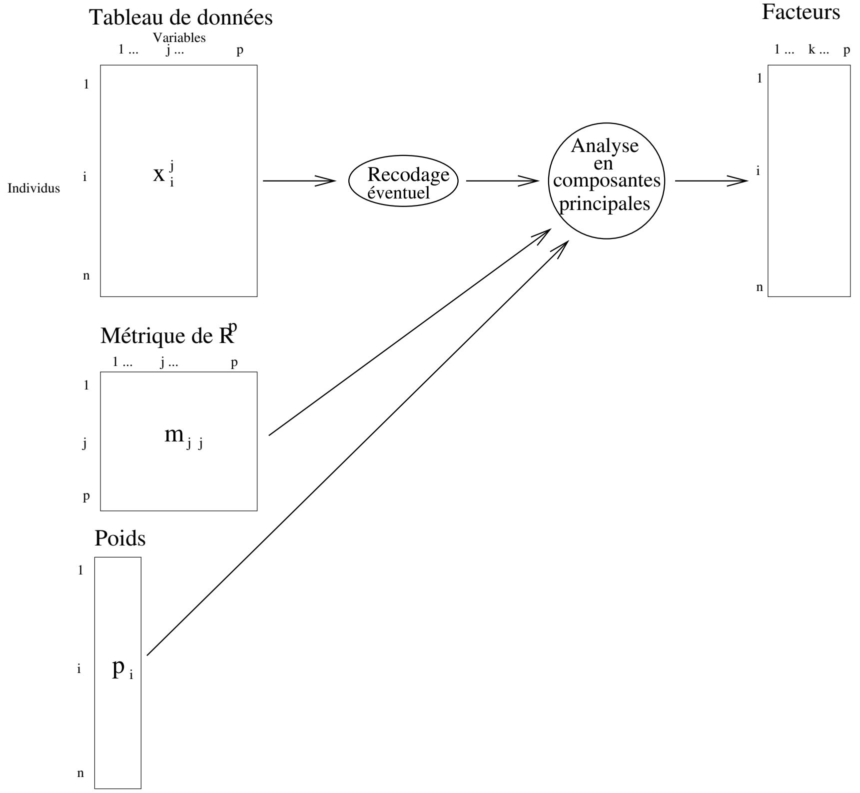
- unicité de l'arbre (à des permutations près des paires descendant d'un même nœud).
- choix raisonné du nombre de classes
- algorithme coûteux si $n$ grand ($n$ généralement inférieur à 1000).



**Propriétés et différence par rapport à la classification type nuées dynamiques**

Unicité (à un jeu de données ne correspond qu'un arbre). Parler plutôt de quasi-unicité car l'unicité d'une CAH n'est pas absolue à cause des ex æquo (« ties » en Anglais). Dans le cas de données discrètes avec peu de variables, ce n'est pas si rare. Le choix d'une des 2 agrégations ex æquo modifiera les distances entre la nouvelle classe formée et les autres. L'autre agrégation ex æquo ne sera pas peut-être effectuée au même niveau ⇒ l'arbre sera différent.
Mais cette quasi-unicité n'est pas qu'un avantage : une population est généralement répartie de façon continue, il existe donc plusieurs classifications possibles. La classification type *nuées dynamiques* reflète cela.

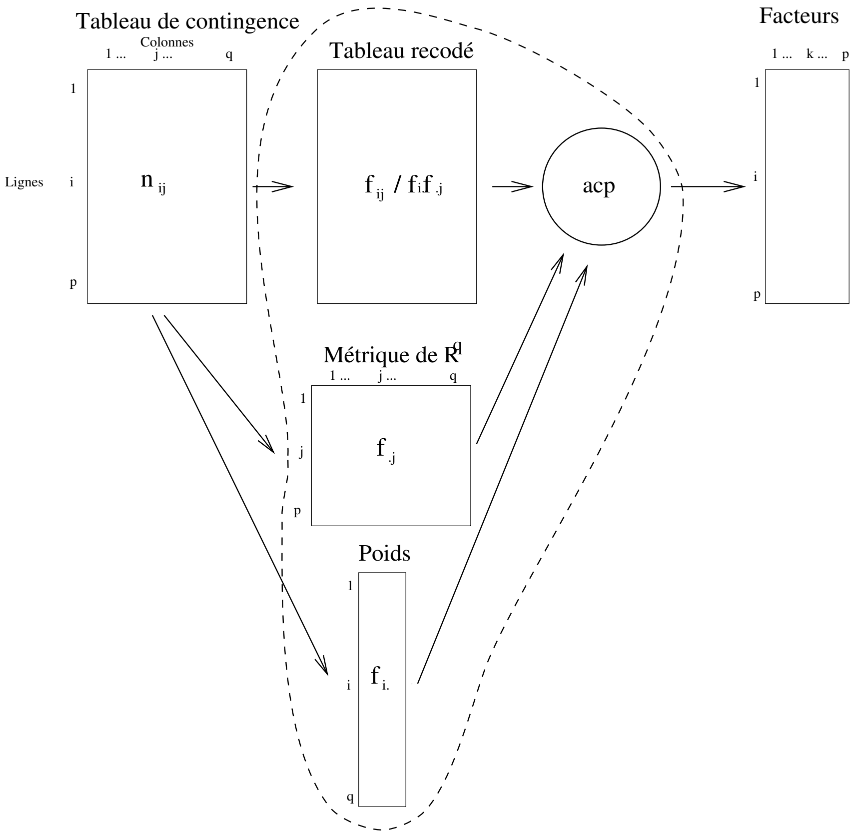
## Bilan des méthodes multidimensionnelles et méthodologie : données et résultats de l'acp






– L'analyse en composantes principales admet en entrée :
  – un tableau de données individus x variables $n$ lignes $p$ colonnes.
  – une métrique dans $\mathbf{R}^p$ (qui peut-être la métrique usuelle).
  – un ensemble de poids $p_i$ ($p_i > 0$, $\sum_i p_i = 1$) associés aux individus (on peut avoir $p_i = 1/n \forall i$).
– Elle fournit les nouvelles coordonnées des individus (les facteurs), les valeurs propres et les vecteurs propres. Les indices variés (qualité, contributions) sont calculables à partir de ces résultats.
– Dans le diagramme, les rectangles sont des données ou tableaux de résultats, les ronds des méthodes.

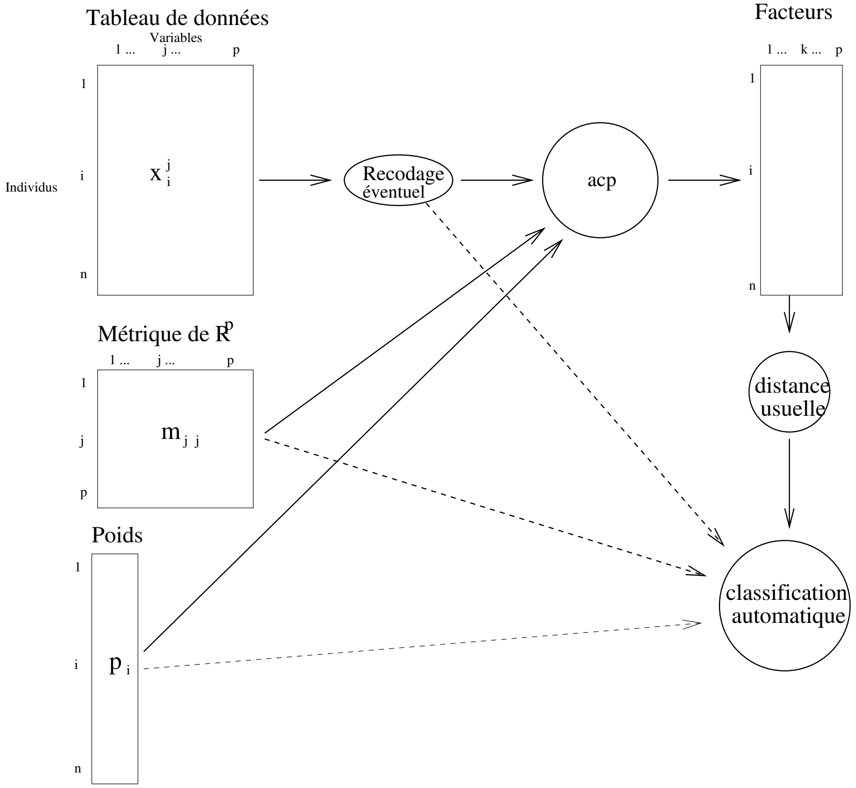
## Bilan et méthodologie : l'analyse des correspondances






L'analyse des correspondances réalise une acp particulière, et n'admet en entrée que la table de contingence (elle fait elle-même le recodage et le calcul des métriques et du système de pondération). La méthode afc est ici symbolisée par le contour en pointillés.

## Bilan et méthodologie : classification après acp (ou afc)






– Les données d'une classification automatique sont :

1. soit un tableau de données (éventuellement recodé), une métrique (le plus souvent usuelle, dans les programmes de classification automatique) et pour certaines méthodes (comme la méthode de Ward) un système de poids,
2. soit les composantes principales d'une acp (ou afc) avec *systématiquement* la métrique usuelle. En effet, la métrique souhaitée est déjà intégrée dans les composantes principales de l'acp ou de l'afc réalisée : la distance entre deux lignes du tableau recodé au sens de la métrique $M$ est égale à la distance (au sens usuel) entre les deux lignes associées des composantes principales.

On rajoute en entrée de la méthode de classification le système de poids si nécessaire.

– On peut remplacer ici l'acp par l'afc. La classification permet alors de regrouper les lignes du tableau de contingence ayant des profils lignes proches.

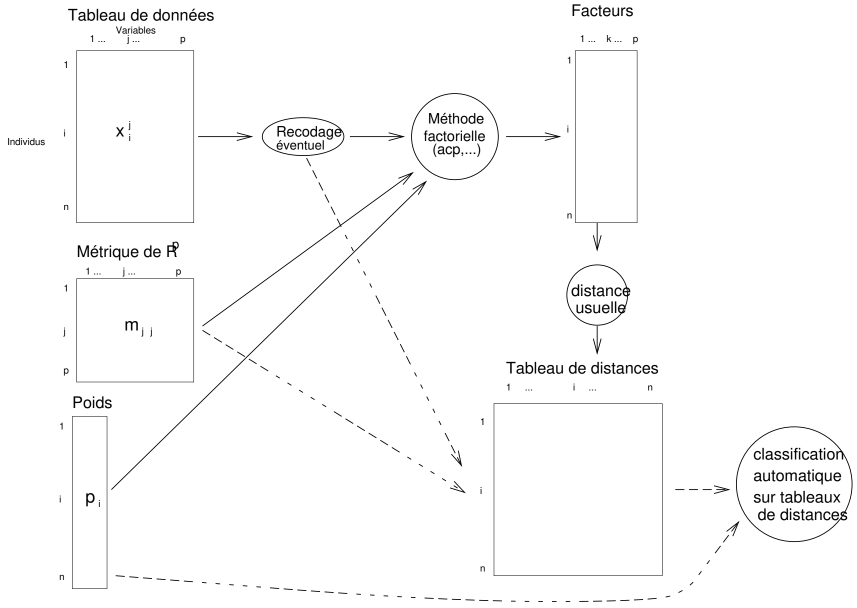
– *En pratique*, puisque les programmes de classification n'admettent pas en général de métrique $M$, si on utilise les entrées "tableau de données + métrique non usuelle", il est nécessaire de modifier le tableau de données pour y intégrer la métrique (comme lorsque l'on passe de la DVS à la DVSG). C'est pourquoi il est plus sûr et plus commode d'utiliser comme entrée les composantes principales de l'acp ou de l'afc. On effectue alors une *classification sur composantes principales*. Ainsi, on est sûr que la distance entre deux individus, utilisée en classification automatique, est la même que celle utilisée dans la méthode factorielle.

– Il est souvent intéressant d'accompagner une méthode factorielle d'une classification et inversement :

  – La classification après acp ou afc permet d'appréhender la qualité du résumé fourni par un plan : si les classes calculées sur l'ensemble des dimensions sont bien séparées sur le plan, alors le résumé fourni par le plan est bon. A l'inverse, si les classes représentées sur le plan sont très mal séparées, alors le plan ne résume qu'une petite partie de l'inertie. Enfin (surtout si $n$ est grand), il est souvent plus facile de décrire les différences entre des groupes d'individus qu'entre les individus eux-mêmes.
  – De façon inverse, l'acp (ou l'afc) permet de situer les classes les unes par rapport aux autres. Si toutes les classes sont ordonnées sur le premier axe, alors la classification est *unidimensionnelle* et les classes peuvent être hiérarchisées : un seul facteur permet de discriminer et d'ordonner les classes. Si elles apparaissent séparées sur le plan, la classification est bidimensionnelle. Enfin, si elles sont projetées en se recouvrant partiellement, alors il est probable qu'il faille au moins trois facteurs pour les séparer.

– *Filtrer le bruit* On peut n'effectuer la classification que sur les $k$ premiers facteurs. Cela peut se justifier par le fait que l'on suppose alors que les $r - k$ derniers facteurs constituent du bruit qui est ainsi filtré. On peut fixer le nombre $k$ de facteurs en utilisant l'éboulis des valeurs propres (règle du coude). Mais il faut éviter de lui donner une valeur trop petite car on supprimerait ainsi des facteurs d'intérêt, même si on ne souhaite pas les interpréter. En pratique, on peut souvent garder tous les facteurs dans la classification

## Bilan et méthodologie : classification sur tableau de distances






Certains programmes de classification admettent en entrée, non plus le tableau individus variables mais le tableau de distances entre individus. Un des intérêts est qu'on peut ne disposer que d'un tableau de distances, mais pas du tableau individus x variables. Parfois, le tableau de distances n'est en fait qu'un tableau de dissimilarités (il ne possède pas toutes les propriétés d'une distance), mais on peut souvent "faire comme si" c'était un vrai tableau de distances. Par exemple, si on prend comme distance entre deux variables $\mathbf{x}^j$ et $\mathbf{x}^{j'}$ la quantité $1 - r^2\left(\mathbf{x}^j, \mathbf{x}^{j'}\right)$, on obtient bien un tableau de dissimilarités (deux variables sont proches si elles sont liées linéairement, et la dissimilarité d'une variable avec elle-même est nulle). Ce tableau peut être assimilé à un tableau de distances et servir de base à une classification.

## Bilan et méthodologie : la classification mixte

- Lorsque $n$ est grand
  - les méthodes hiérarchiques
  - les méthodes utilisant un tableau de distances

  sont inutilisables.

  Seules restent en lice les méthodes de réallocation itérative.
- La méthode *mixte* permet de traiter un échantillon très grand tout en gardant certains aspects agréables de la CAH (choix du nombre de classes).

*La classification mixte comporte trois phases*

1. La première consiste à remplacer les $n$ individus initiaux par $N$ noyaux ou centre de classes, obtenus par une méthode de réallocation itérative (on peut passer ainsi par exemple de $n = 10000$ à $N = 500$).
2. Dans une seconde phase, les $N$ noyaux obtenus, affectés de leur poids, subissent alors une C.A.H. avec le critère de Ward. On choisit alors le nombre final de classes, grâce à l'arbre de classication (ou l'éboulis des indices de niveau).
3. Après coupure de l'arbre, dans une troisième phase, on reprend la méthode de réallocation itérative sur l'échantillon initial, en prenant comme noyaux les barycentres des classes obtenue dans la C.A.H.. Le poids d'une classe est encore défini par la somme des poids des individus de la classe. La partition issue de la C.A.H. se trouve ainsi améliorée.



- Lorsque $n$ est grand, la classification sur tableau de distances n'est pas possible (selon les programmes et les ordinateurs, le nombre $n$ doit être inférieur à quelques centaines (avec $n = 500$ et n = 1000, on obtient respectivemetn une demi-matrice de distance d'environ 125000 à 500 000 éléments).
- La classification hiérarchique ascendante nécessite le calcul des éléments de la matrice de distance, mais les algorithmes peuvent permettre de ne pas les stocker en mémoire. Cette méthode reste inutilisable pour de grandes valeurs de $n$.
- Les méthodes de réallocation présentent l'inconvénient de fournir plusieurs partitions (en fonction des noyaux initiaux) et de ne pas permettre un choix aisé du nombre de classes. La méthodologie associée (formes fortes, classification des formes fortes, ...) est lourde.
- Le critère de Ward est revient à maximiser l'inertie inter-classe. C'est le même critère que celui des méthodes de réallocation.
- La troisième phase est une phase d'amélioration. Elle peut toujours être utilisée après une méthode hiérarchique avec critère de Ward.