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.

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.

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 %). Classe 2 : plutôt peu touchés (35 %). 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, classement ?

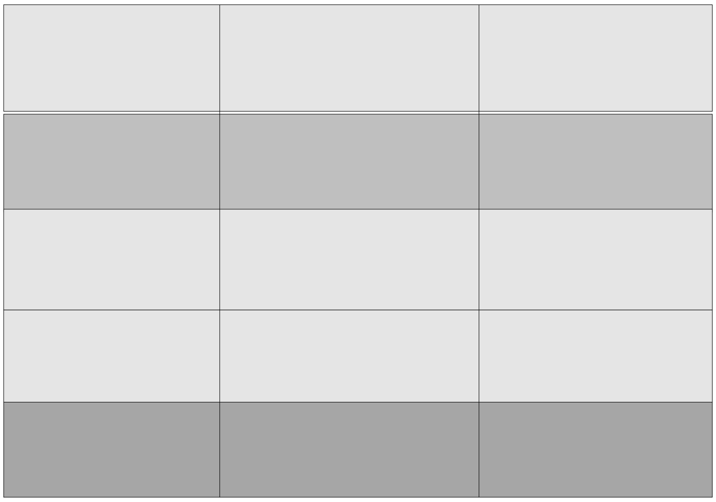
Ensemble de variables	Une variable de groupe et un ensemble de variables explicatives
$\Downarrow$	$\Downarrow$
Classification automatique	**Analyse discriminante** **Segmentation** **Discrimination logistique**
$\Downarrow$	
Définition de classes	
$\Downarrow$	$\Downarrow$

Règles de classement

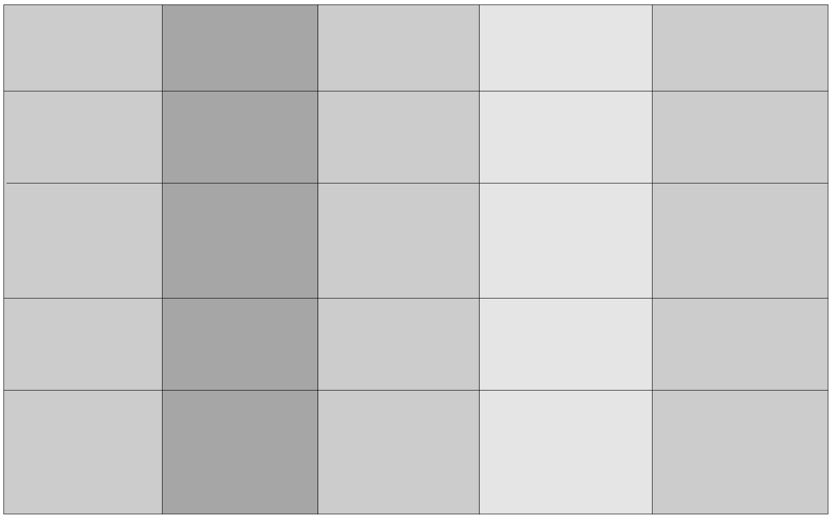
Que peut-on classifier?

- 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.

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]

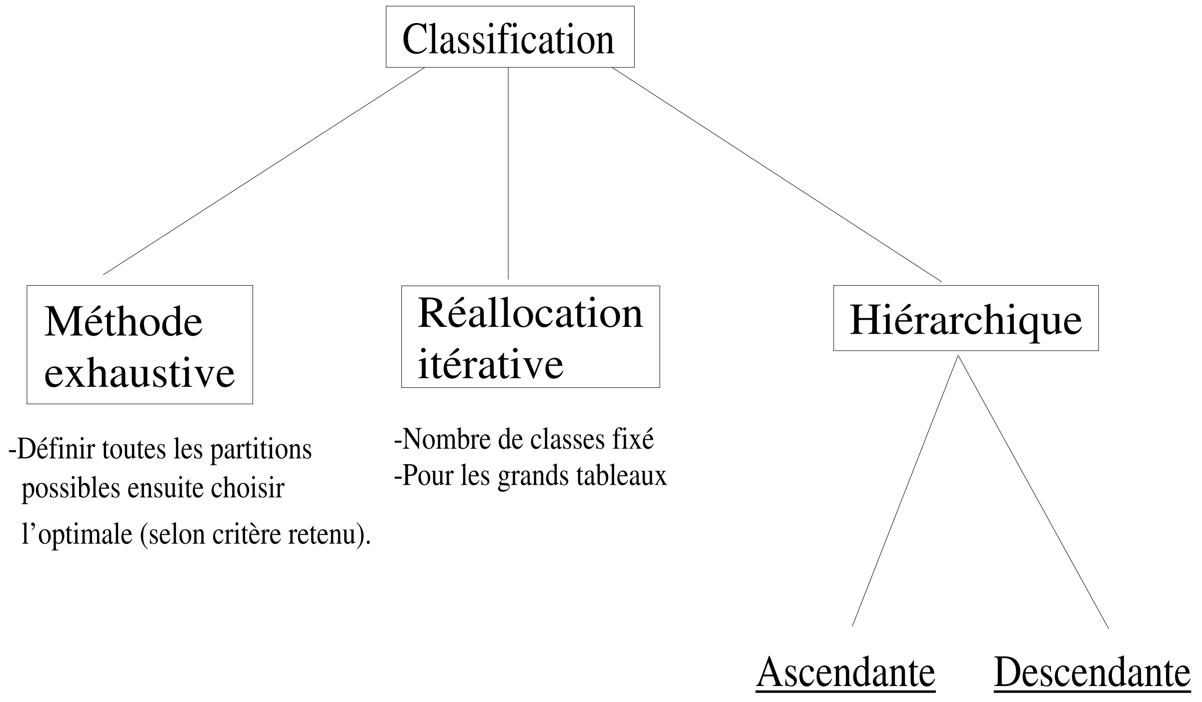
	Individu 1	Individu 2	Individu 3	Individu 4	Individu 5
Individu 1	0	4	1	3	3.16
Individu 2	4	0	4.12	5	1.41
Individu 3	1	4.12	0	2	3
Individu 4	3	5	2	0	3.6
Individu 5	3.16	1.41	3	3.6	0

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

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.

Méthodes de classification



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

Autre méthode : modélisation de classes.

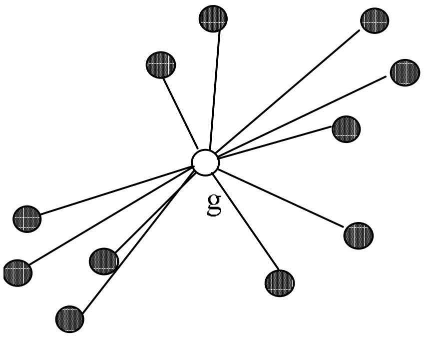
Un critère de qualité d'une partition : l'inertie inter

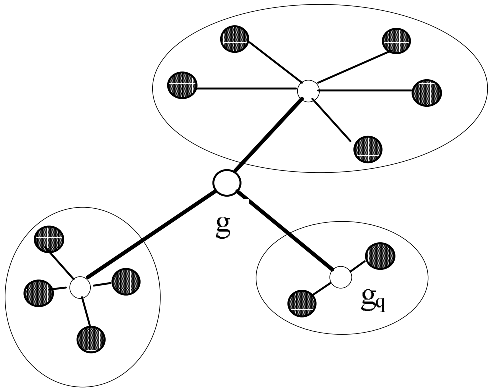
Inertie totale = inertie intra + inertie inter

Inertie totale : $I_{totale} = \sum_i m_i \|x_i - g\|^2$

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

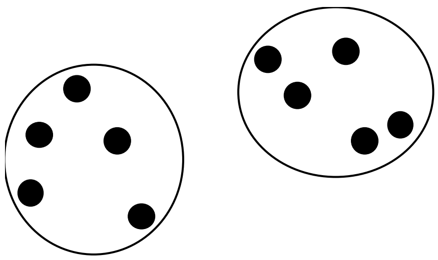
Inertie inter : $I_{inter} = \sum_q \pi_q \|g_q - g\|^2$





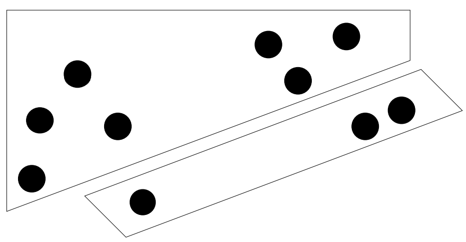
Le plus souvent, le critère retenu en classification automatique cherche à minimiser l'inertie intra (et donc à maximiser l'inertie inter).

inertie inter forte



Inertie intra-classes faible

inertie inter faible



Inertie intra-classes élevée

Classification type réallocation itérative : nuées dynamiques

Critère :

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

Démarche :

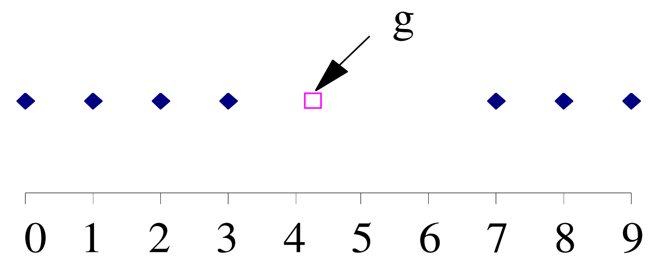
1. choix *a priori* du nombre de classes et de *noyaux* initiaux.
2. les individus changent de classe d'une itération à l'autre.
3. après réaffectation, on recalcule les nouveaux noyaux.
4. l'algorithme s'arrête lorsque 2 partitions successives sont identiques.

Remarque :

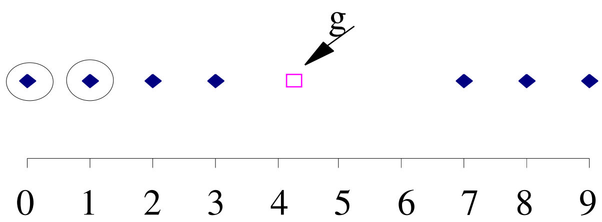
– méthode rapide, pratique pour les gros échantillons.
– le résultat dépend du choix initial des noyaux.

Classification type réallocation itérative : nuées dynamiques

Critère : minimiser l'inertie intra

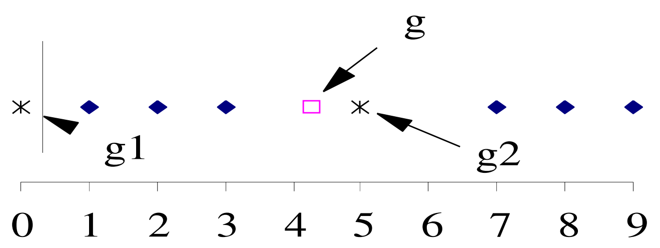


7 individus,
g : centre de gravité,
Objectif : obtenir 2 classes.

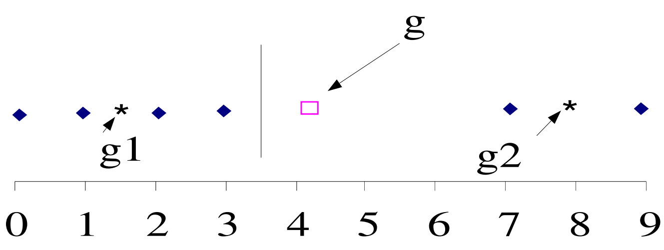


Choix des noyaux,
g : centre de gravité, coordonnée 4.286
Inertie totale : 11.35

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)



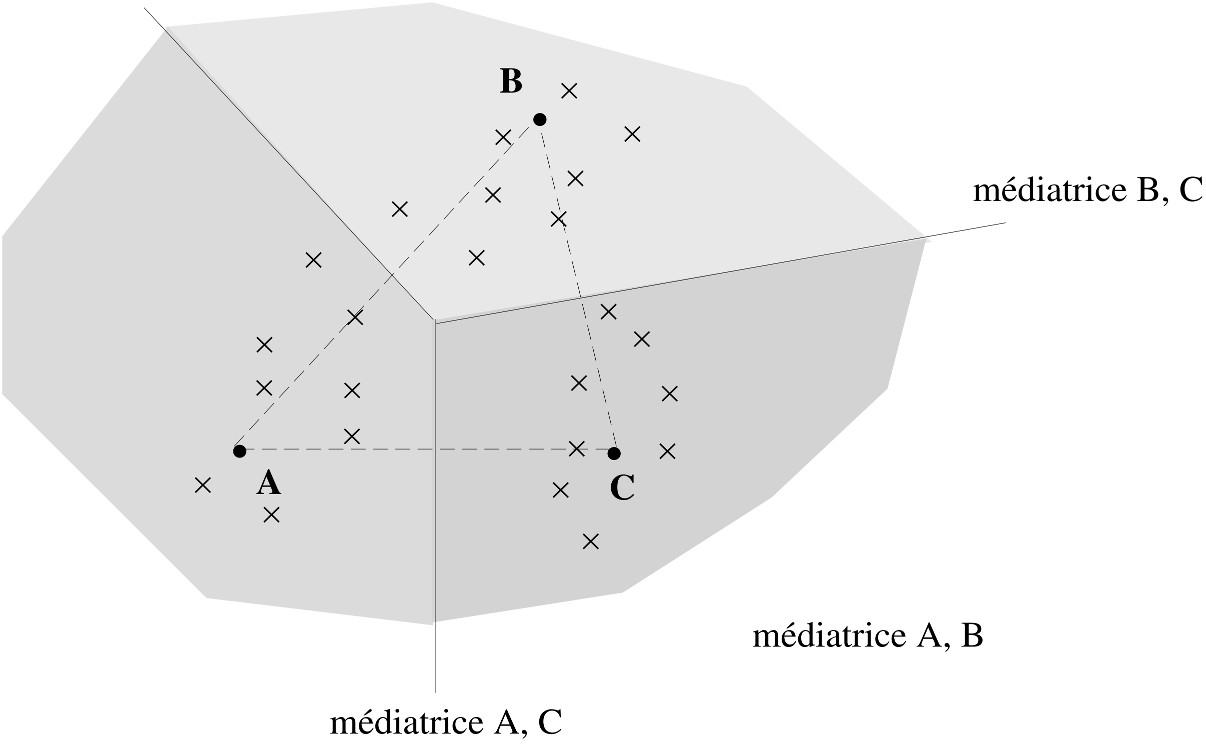
g1, g2 : centre de gravité des 2 classes Inertie intra : 1
(Iinter : 10.35)

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

Classification type réallocation itérative : nuées dynamiques

Description d'une étape

Phase 2 : réaffectation

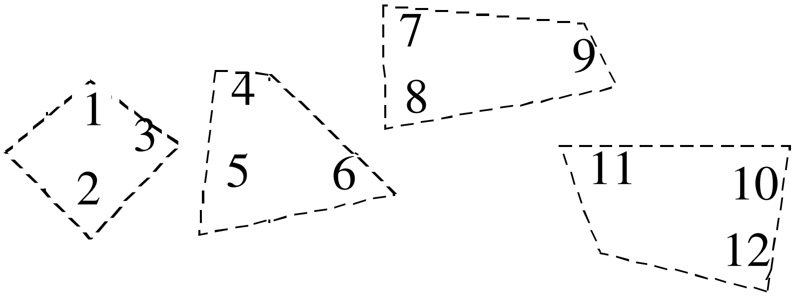


À chaque étape, les médiatrices ou hyperplans médiateurs délimitent les régions des points les plus proches des noyaux A, B et C.

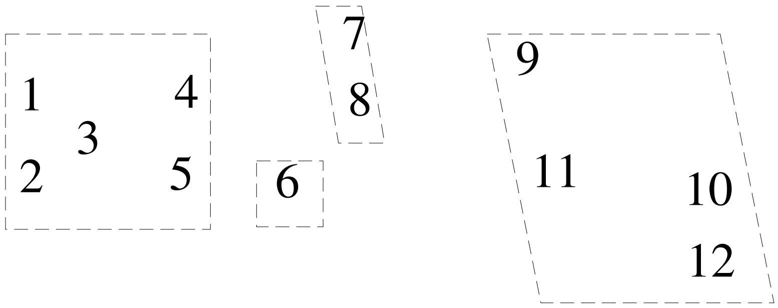
Phase 3 : Calcul des noyaux

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.



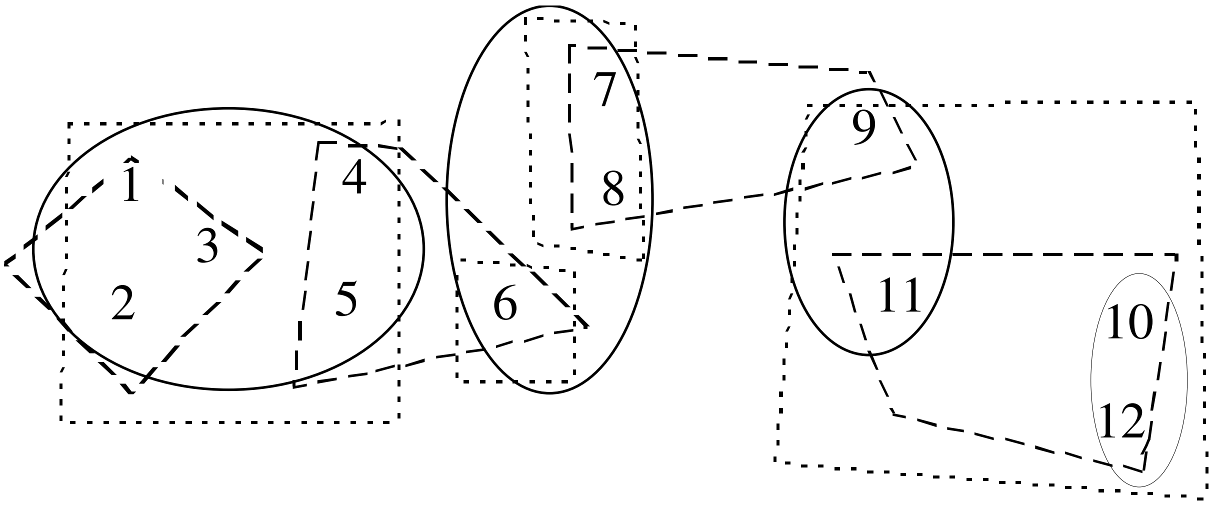
Partition 2 avec 4 noyaux.

Remarques : il peut y avoir plusieurs partitions vraisemblables.

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

Classification type réallocation itérative : nuées dynamiques

Formes fortes



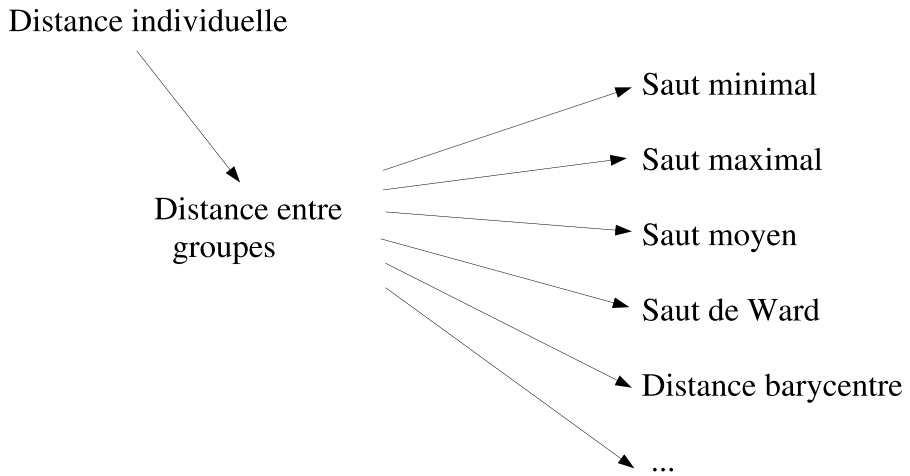
	Partition 1	Partition 2	Partition 3	F.Forte
1	A	α	a	I
2	A	α	a	I
3	A	α	a	I
4	B	α	a	II
5	B	α	a	II
6	B	β	b	III
7	C	γ	b	IV
8	C	γ	b	IV
9	C	δ	c	V
10	D	δ	d	VI
11	D	δ	c	VII
12	D	δ	d	VI

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.



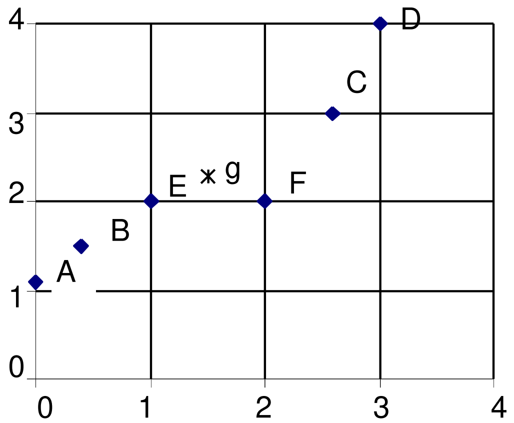
Démarche

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	$(\text{Y1-Y1g})^2$	$(\text{Y2-Y2g})^2$
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

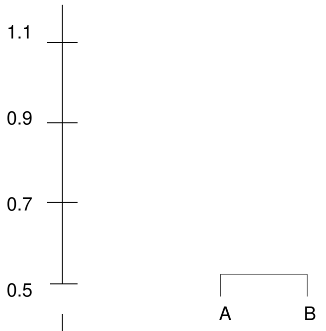
Matrice des distances euclidiennes					
1	2	3	4	5	6
0.00	0.57	3.22	4.17	1.35	2.19
0.57	0.00	2.66	3.61	0.78	1.68
3.22	2.66	0.00	1.08	1.89	1.17
4.17	3.61	1.08	0.00	2.83	2.24
1.35	0.78	1.89	2.83	0.00	1.00
2.19	1.68	1.17	2.24	1.00	0.00

CLASSIFICATION HIERARCHIQUE

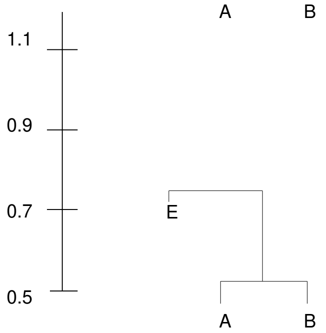
Distance entre groupe : ex. du saut minimal

Définition de la stratégie d'agrégation :

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

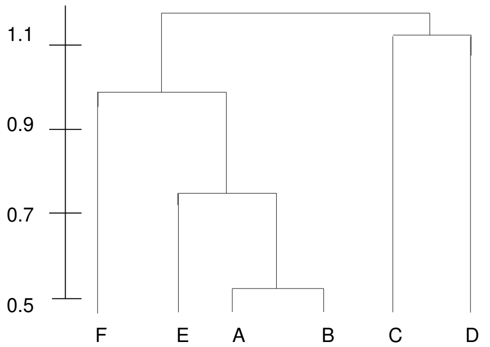


	ind A	ind B	ind C	ind D	ind E	ind F
A	0					
B	0.56	0				
C	3.20	2.66	0			
D	4.20	3.60	1.08	0		
E	1.34	0.78	1.89	2.80	0	
F	2.20	1.67	1.17	2.20	1	0



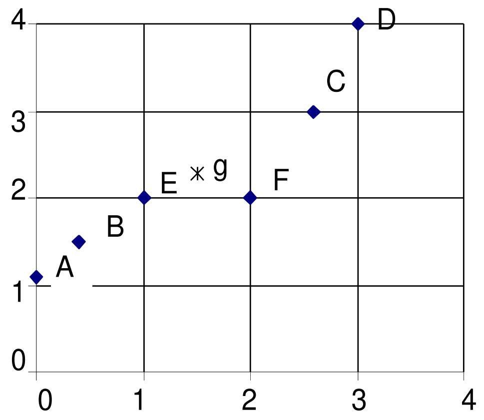
	cl 1	ind C	ind D	ind E	ind F
cl 1	0				
C	2.66	0			
D	3.60	1.08	0		
E	0.78	1.89	2.80	0	
F	1.67	1.17	2.20	1	0

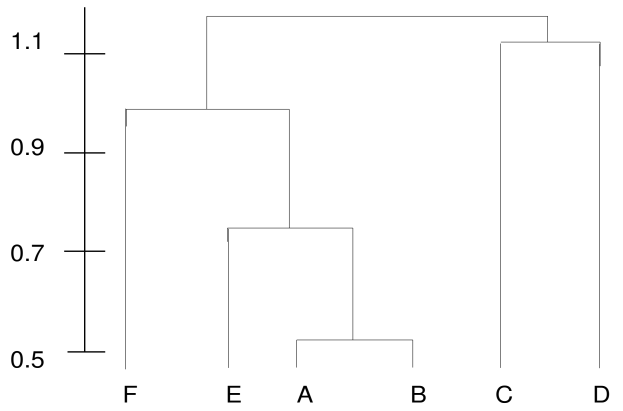
. . . .

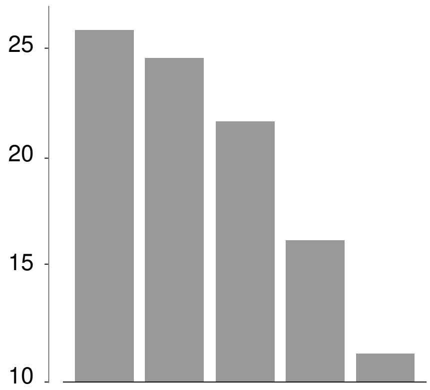


	cl 3	cl 4
cl 3	0	
cl 4	1.17	0

CLASSIFICATION HIERARCHIQUE







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

CLASSIFICATION HIERARCHIQUE
Distance entre groupes : saut de Ward

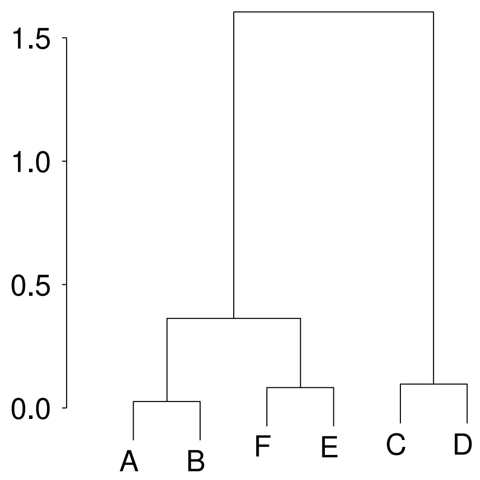
Soit d la distance euclidienne définie entre deux individus, et x et y deux groupes d'individus :

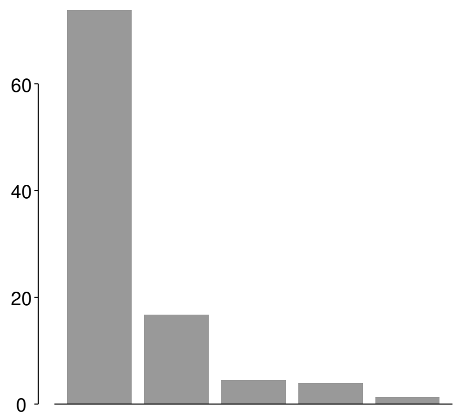
- de barycentres $\mathbf{g}_x$ et $\mathbf{g}_y$,
- de poids p_x et p_y

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



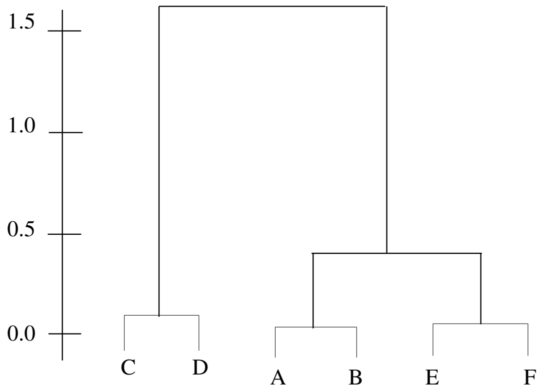


CLASSIFICATION HIERARCHIQUE :
Choix de la distance entre groupes

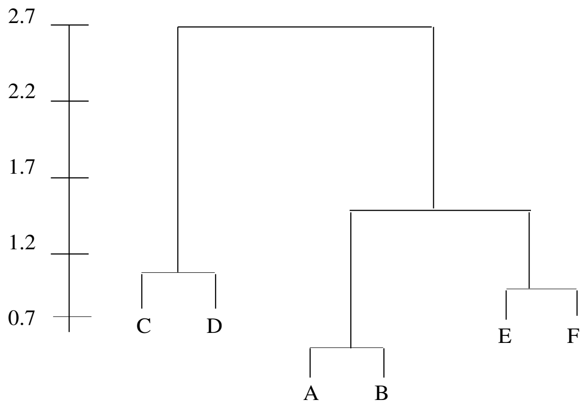
Saut de Ward :

Inertie inter avec 3 groupes :

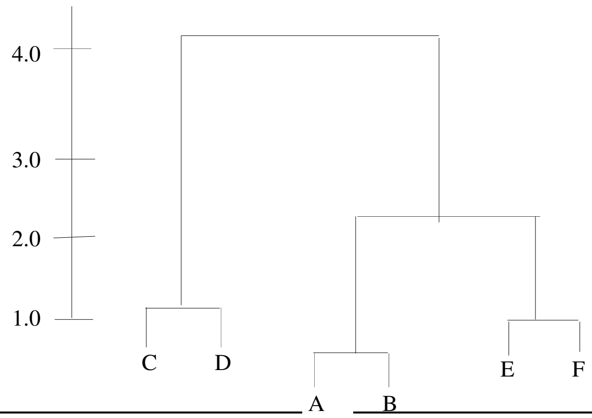
1.97



Saut moyen :



Saut maximal :



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.

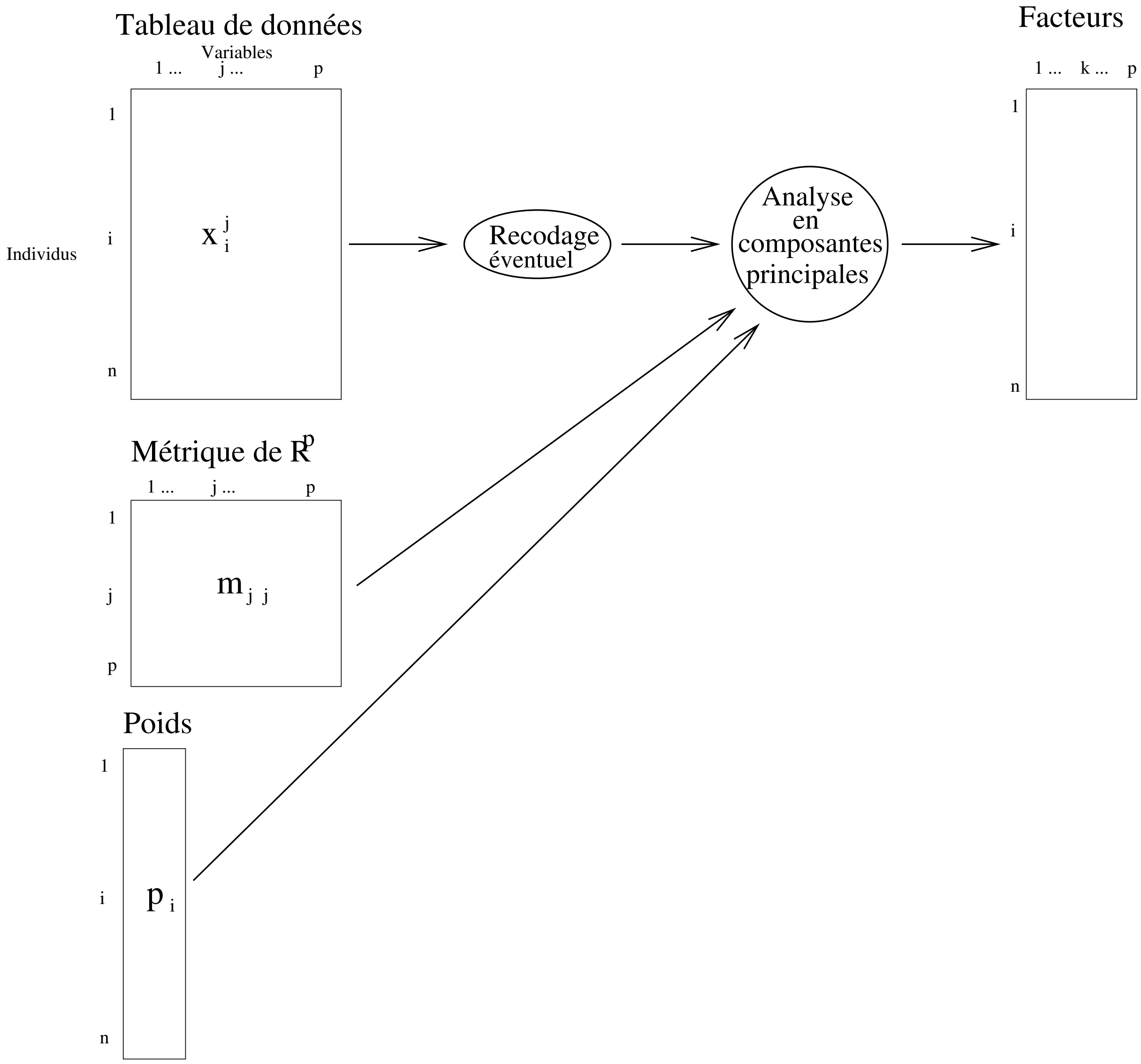
Dans la pratique, on utilise plutot le saut moyen et le saut de Ward.

CLASSIFICATION HIERARCHIQUE

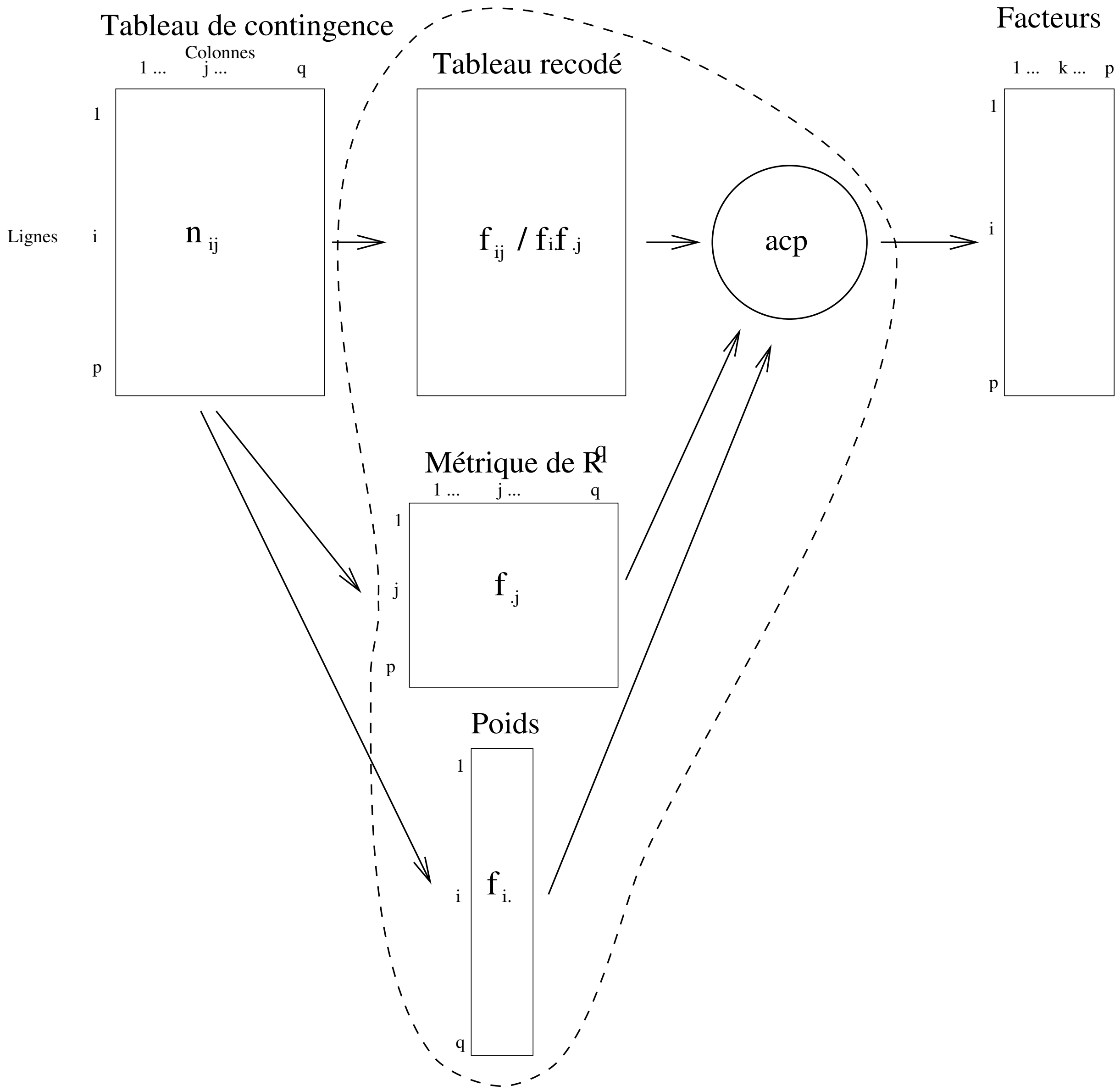
Caractéristiques :

- 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).

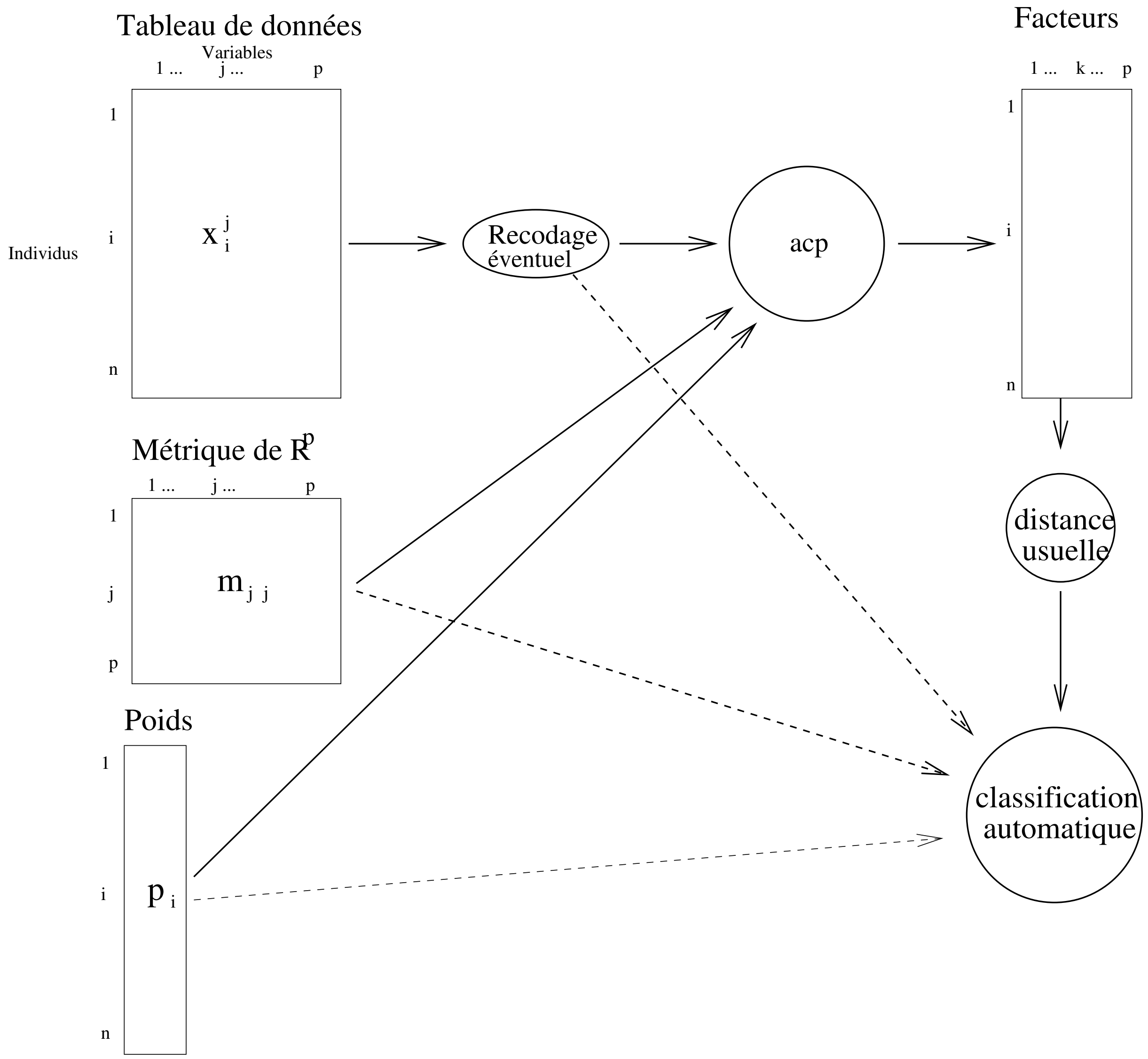
Bilan des méthodes multidimensionnelles et méthodologie : données et résultats de l'acp



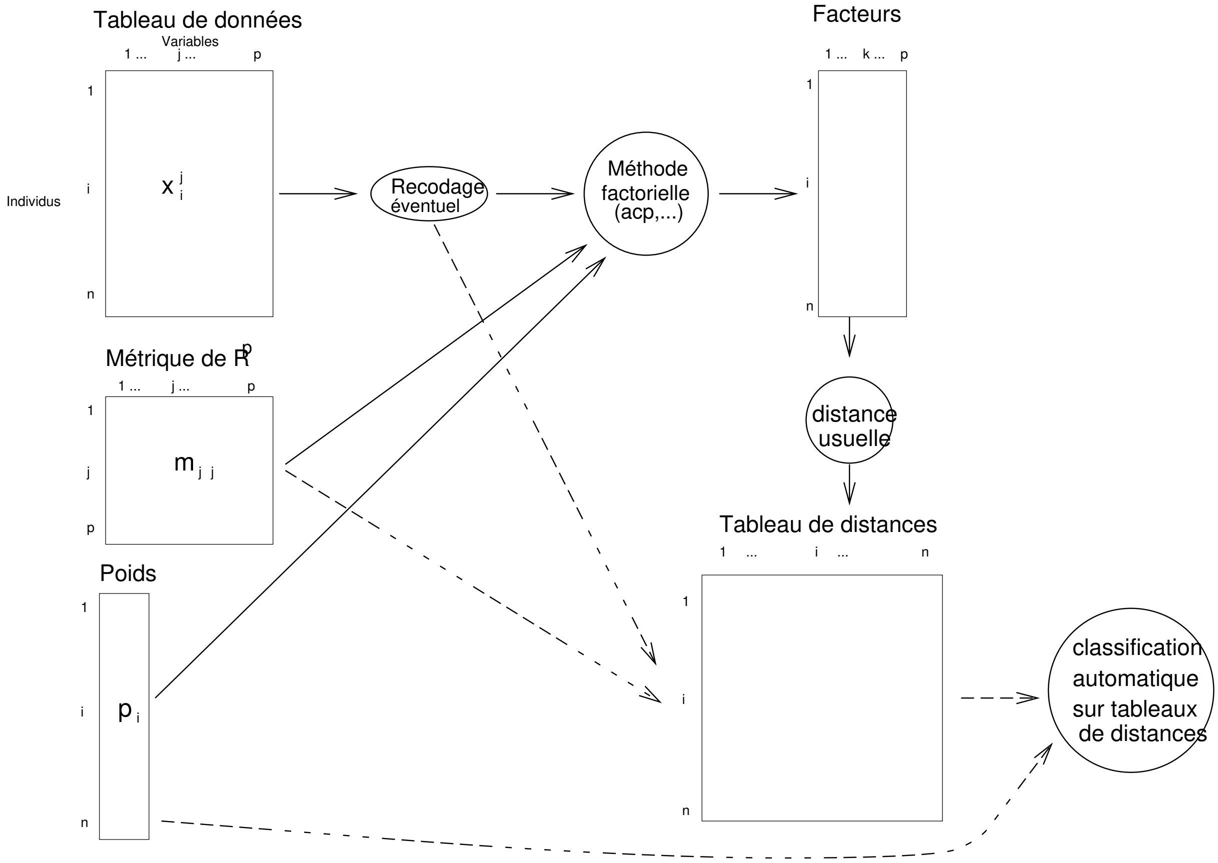
Bilan et méthodologie : l'analyse des correspondances



Bilan et méthodologie : classification après acp (ou afc)



Bilan et méthodologie : classification sur tableau de distances



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.