Introduction à l'analyse des correspondances et à la classification

Bertrand looss
Véronique Verrier
EDF R&D
Département Management des Risques Industriels

Cours IUP SID Toulouse - M1 - 17/10/2011

Les catégories de méthodes d'analyse de données

Les méthodes descriptives :

- **L'analyse en composantes principales** cherche à représenter dans un espace de dimension faible un nuage de points représentant n individus, ou objets, décrits par p variables numériques en utilisant les corrélations existant entre ces variables.
- **L'analyse des correspondances** (AFC ou ACM) étudie les proximités entre individus décrits par deux ou plusieurs variables qualitatives ainsi que les proximités entre les modalités de ces variables.
- **Les méthodes de classification** ou de typologie procèdent par regroupement des individus en classes homogènes.

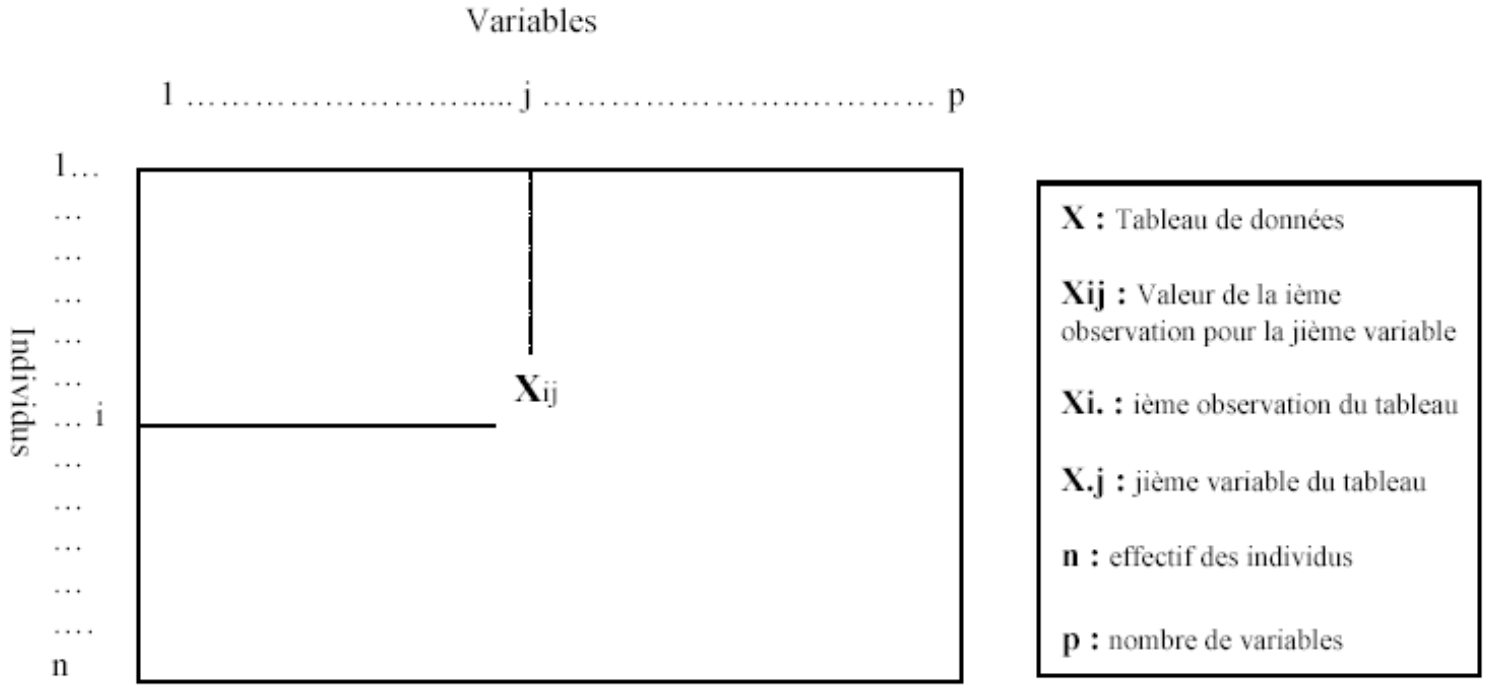
Les méthodes explicatives et prédictives :

- **La régression logistique** étudie la prévision d'une variable binaire au moyen de plusieurs autres.
- **L'analyse discriminante** étudie la prévision d'une variable qualitative par des variables numériques.
- **Les arbres de décision / régression** étudient la prévision d'une variable respectivement qualitative ou quantitative

ACP – Analyse en composantes principales

Données type : tableau rectangulaire de mesures où :

- les colonnes sont des variables quantitatives
- et dont les lignes représentent des individus statistiques



Objectif : visualiser, résumer l'information contenue dans ce tableau afin d'avoir une réprésentation permettant plus facilement l'interprétation

ACP – Analyse en composantes principales

Démarche

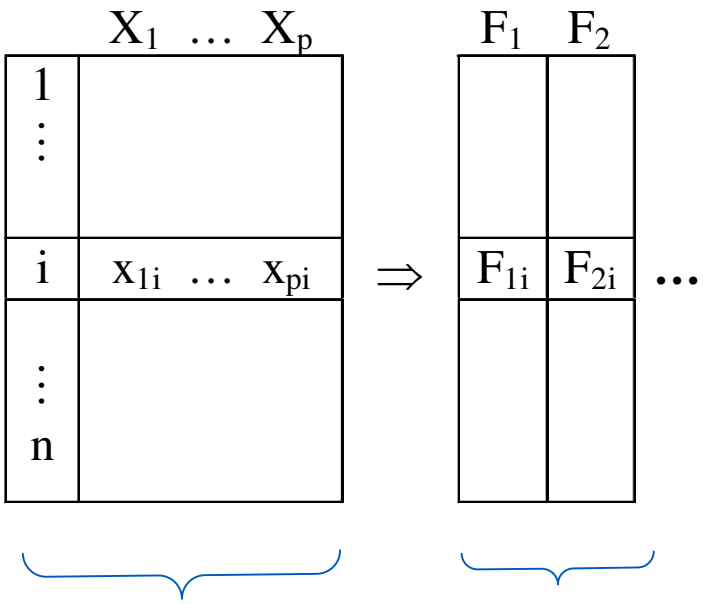
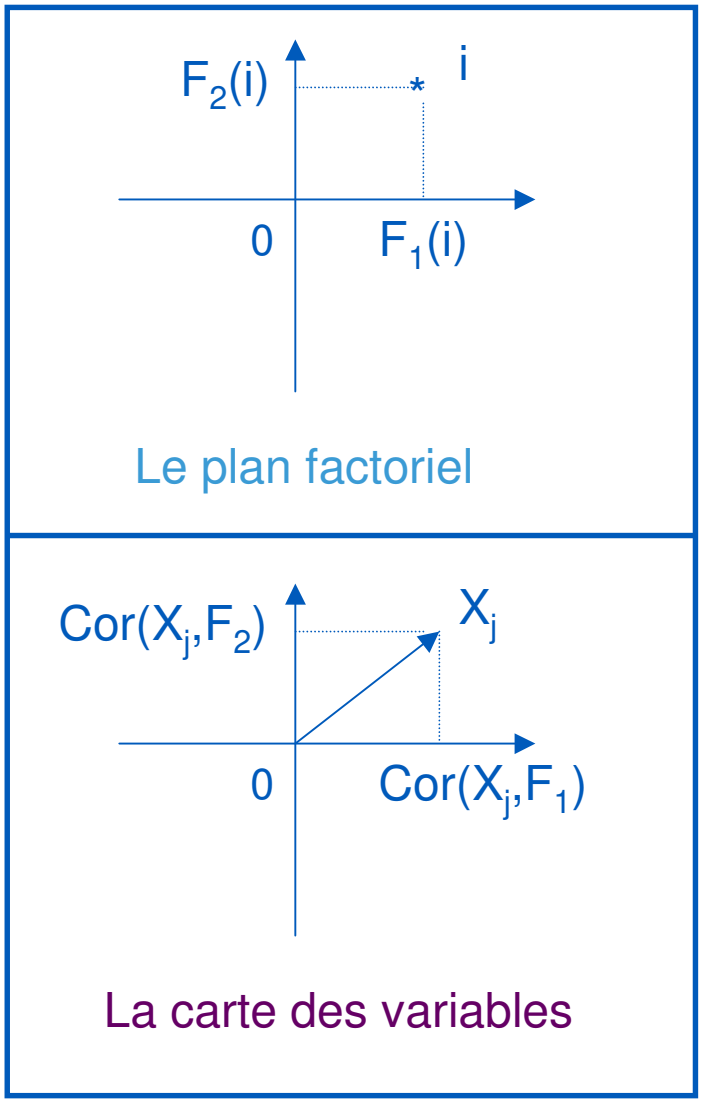


Tableau des données

Facteurs centrés-réduits résumant les données

$$F_h = \sum_{j=1}^{p} u_{hj} X_j$$

(non corrélés entre eux)



Plan du cours

1. **Introduction à l'Analyse Factorielle des Correspondances**
2. Exemple
3. Formalisation mathématique de l'AFC
4. Exemple

5. Analyse des correspondances multiples
6. Classification
7. Synthèse

8. Évaluation

Analyse factorielle des correspondances (AFC)

Etude des corrélations entre deux variables catégorielles
(dites aussi « qualitatives »)

Exemples :

- Variables nominales : sexe, catégorie socio-professionnelle
- Variables ordinales : mention à un examen, tranche d'âge

Historique :

- Principes théoriques : Fisher (1940)
- AFC développée par J-P Benzécri et ses étudiants en France (> 1965)
  « Analyse de données à la française »
- Nombreux développements ultérieurs

Objectifs de cette introduction

[Source : cours de Rémi Bachelet, EC Lille]

- Comprendre les concepts de l'AFC
- Connaître les principes de calcul
- Savoir interpréter les résultats
- Placer l'AFC par rapport à l'ACP et aux méthodes de classification

Exemple 1

Dans une entreprise, la répartition par sexe et catégorie socio-professionnelle (CSP) est la suivante :

	Ouvriers	Techniciens	Cadres
Hommes	20	40	40
Femmes	30	60	10

Y-a-t-il un lien entre le sexe S à deux modalités et la CSP à trois modalités ?

Exemple 2

Dans une entreprise, la répartition par âge et catégorie socio-professionnelle (CSP) est la suivante :

	Ouvriers	Techniciens	Administrat	Cadres sup
< 30 ans	5	8	3	12
[30 ; 40[ans	10	6	5	15
[40 ; 50[ans	15	4	4	15
>= 50 ans	6	4	4	15

Y-a-t-il un lien entre l'âge à 4 modalités et la CSP à 4 modalités ?

Exemple 3 : enquête sur les séjours-vacances des français

Données publiées par l'INSEE en 2002 et étudiées dans *Saporta, 2006*

CSP

$n = 18532$

TABLEAU 6.3 Tableau de contingence

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	41	47	13	59	17	26	4	9	19
Artisans, commerçants, chefs d'entreprise	220	260	71	299	120	42	64	35	29
Cadres et professions intellectuelles supérieures	685	775	450	1242	706	139	122	100	130
Professions intermédiaires	485	639	292	1250	398	189	273	68	193
Employés	190	352	67	813	163	92	161	49	72
Ouvriers	224	591	147	1204	181	227	306	74	114
Retraités	754	393	692	1158	223	25	195	47	115
Autres inactifs	31	34	2	225	42	33	5	6	14

Mode d'hébergement

Rsec = rés. 2ndaire

Rppa = rés. principale parents amis

Rspa = rés. 2ndaire parents amis

Y-a-t-il un lien entre la CSP à 8 modalités et le mode de vacances à 9 modalités?

Principes généraux de l'AFC

L'AFC consiste à remplacer un tableau de nombres difficile à analyser par une série de tableaux plus simples qui sont une bonne approximation de celui-ci.

Les tableaux sont simples car ils sont exprimables sous forme de graphiques

Factorielle = mise en facteur du tableau initial

Correspondance = corrélation pour des variables qualitatives

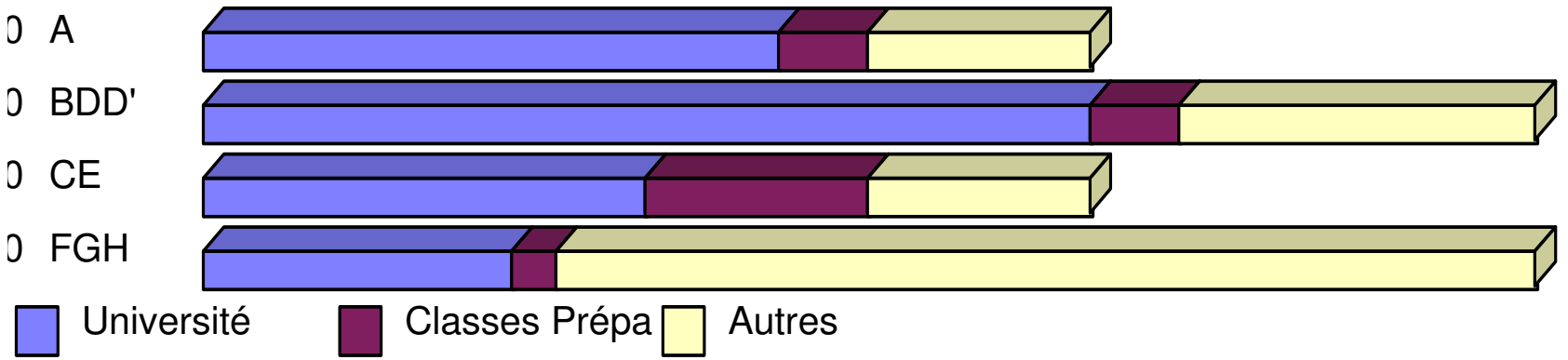
Exemple 4 : devenir des bacheliers

Stats MEN 1975 - 1975 204 489 lycéens

[Exemple issu de Bachelier, EC Lille]

	destination			
	université	***classes prépa***	***autres***	***total***
A	**13**	**2**	**5**	**20**
BDD'	**20**	**2**	**8**	**30**
CE	**10**	**5**	**5**	**20**
FGH	**7**	**1**	**22**	**30**
total	**50**	**10**	**40**	**100**

Représentation graphique par le diagramme en barre



Comment faire parler les données ?

Trouver des valeurs inattendues dans les données, c'est-à-dire des valeurs qui dévient d'une situation attendue (uniforme)

1. Évaluer ce que serait une situation d'uniformité, d'indépendance
2. Calculer en quoi la situation constatée en diffère
3. Exprimer cette différence graphiquement pour pouvoir l'analyser
4. Interpréter les graphiques obtenus
5. Optimiser la lisibilité des graphiques

Matrice « T » des données d'entrée

	destination			
	université	*classes prépa*	*autres*	*total*
A	13	2	5	20
BDD'	20	2	8	30
CE	10	5	5	20
FGH	7	1	22	30
total	50	10	40	100

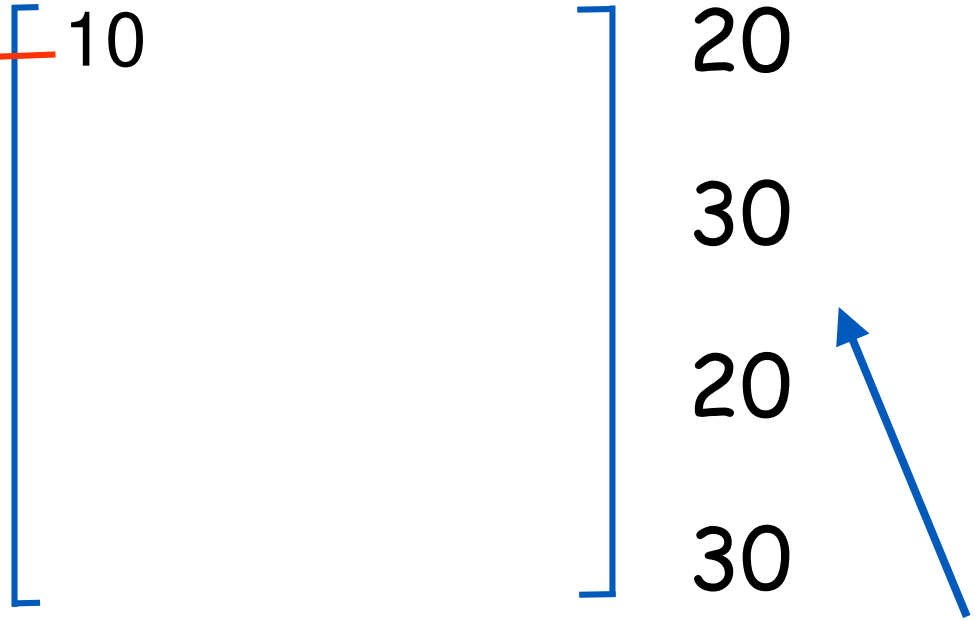
Ce tableau est aussi une matrice, appelons-la « T »

Quelle matrice aurait-on si la répartition dans les filières post-Bac ne dépendait pas du type de Bac ?

Situation d'indépendance

10 = 50 * 20%

(produit matriciel /100 puisqu'on raisonne en %)



50 10 40

On reconstitue la matrice à partir de ses marges

Situation d'indépendance

10 = 50 * 20%

(produit matriciel /100 puisqu'on raisonne en %)

10	2	8	20
15	3	12	30
10	2	8	20
15	3	12	30
50	10	40	

On reconstitue la matrice à partir de ses marges

Appellons cette matrice « T_0 »

Matrice des écarts à l'indépendance

$$T - T_0 = R$$

$$\begin{bmatrix} 13 & 2 & 5 \\ 20 & 2 & 8 \\ 10 & 5 & 5 \\ 7 & 1 & 22 \end{bmatrix} - \begin{bmatrix} 10 & 2 & 8 \\ 15 & 3 & 12 \\ 10 & 2 & 8 \\ 15 & 3 & 12 \end{bmatrix} = \begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix}$$

Quelle est la particularité de R ?

Expression simple de R

On décompose la matrice des écarts à l'indépendance en une somme de matrices

$$R = T_1 + T_2$$

Chacune de ces matrices étant mise en facteur (produit d'un vecteur ligne et d'un vecteur colonne)

$$T_1 = C_1 L_1$$

(une matrice dont la plus petite dimension est N (rang N) est décomposable au maximum en N matrices pouvant se mettre en facteurs)

Ici $T = T_0 + T_1 + T_2$

T est de rang 3, mais R est de rang 2

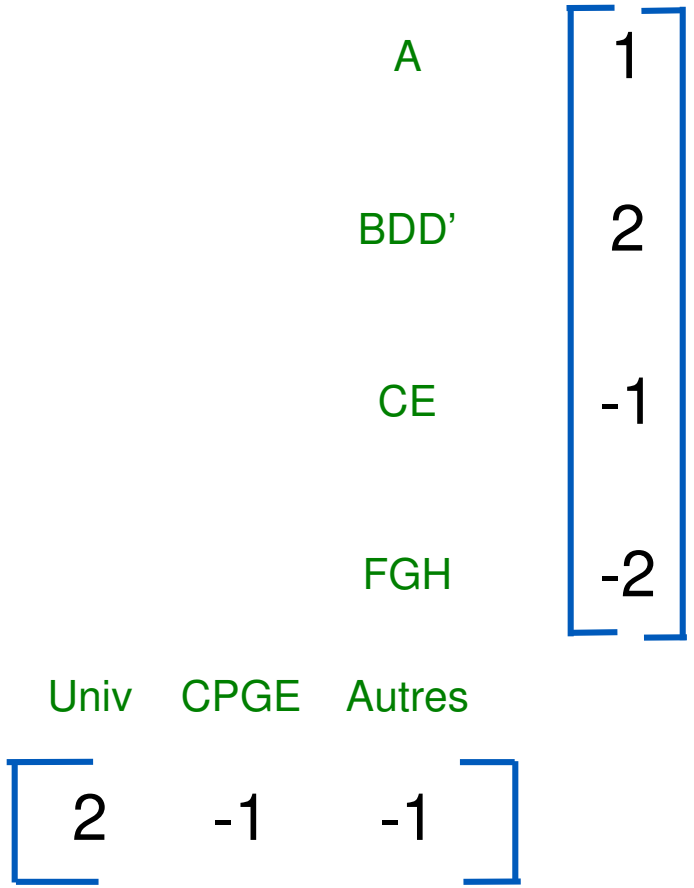
Mise en facteur de R : $R = T_1 + T_2 = C_1L_1 + C_2L_2$

$$R = \begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix} \quad T_1 = \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix} \quad C_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix} \quad T_2 = \begin{bmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{bmatrix} \quad C_2 = \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix}$$

$$L_1 = \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} \qquad L_2 = \begin{bmatrix} 2 & -1 & -1 \end{bmatrix}$$

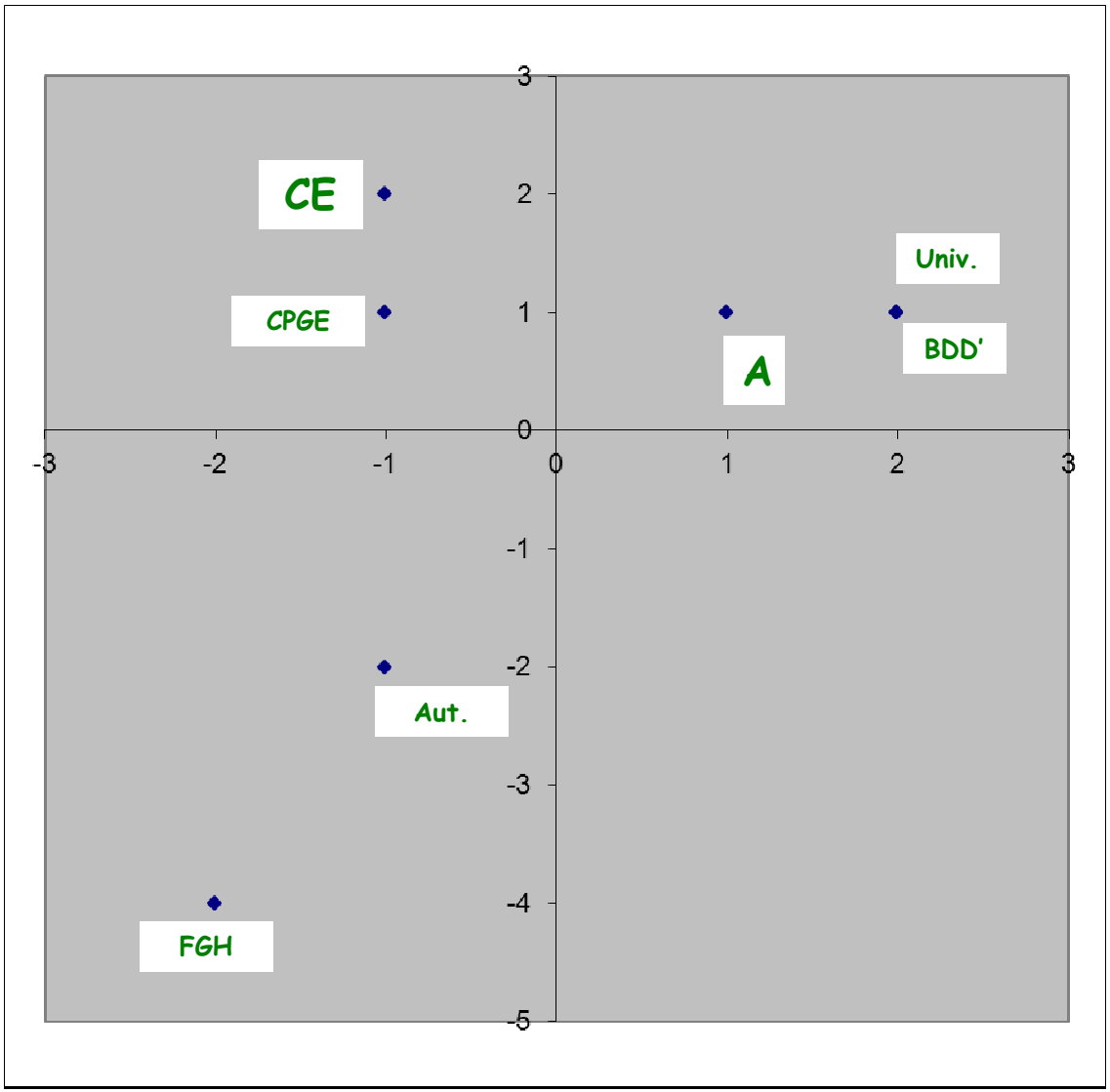
Représentation graphique de R

Un vecteur colonne (resp. ligne) correspond à une modalité des données en colonnes (resp. lignes)



Un axe unidimensionnel + un axe unidimensionnel = un repère

A	1	1
BDD'	2	1
CE	-1	2
FGH	-2	-4
Univ	2	1
CPGE	-1	1
Autres	-1	-2



Interprétation du graphique

1. Conjonction :

   Produit scalaire positif

   Les Bac CE ont une affinité pour la prépa

2. Opposition

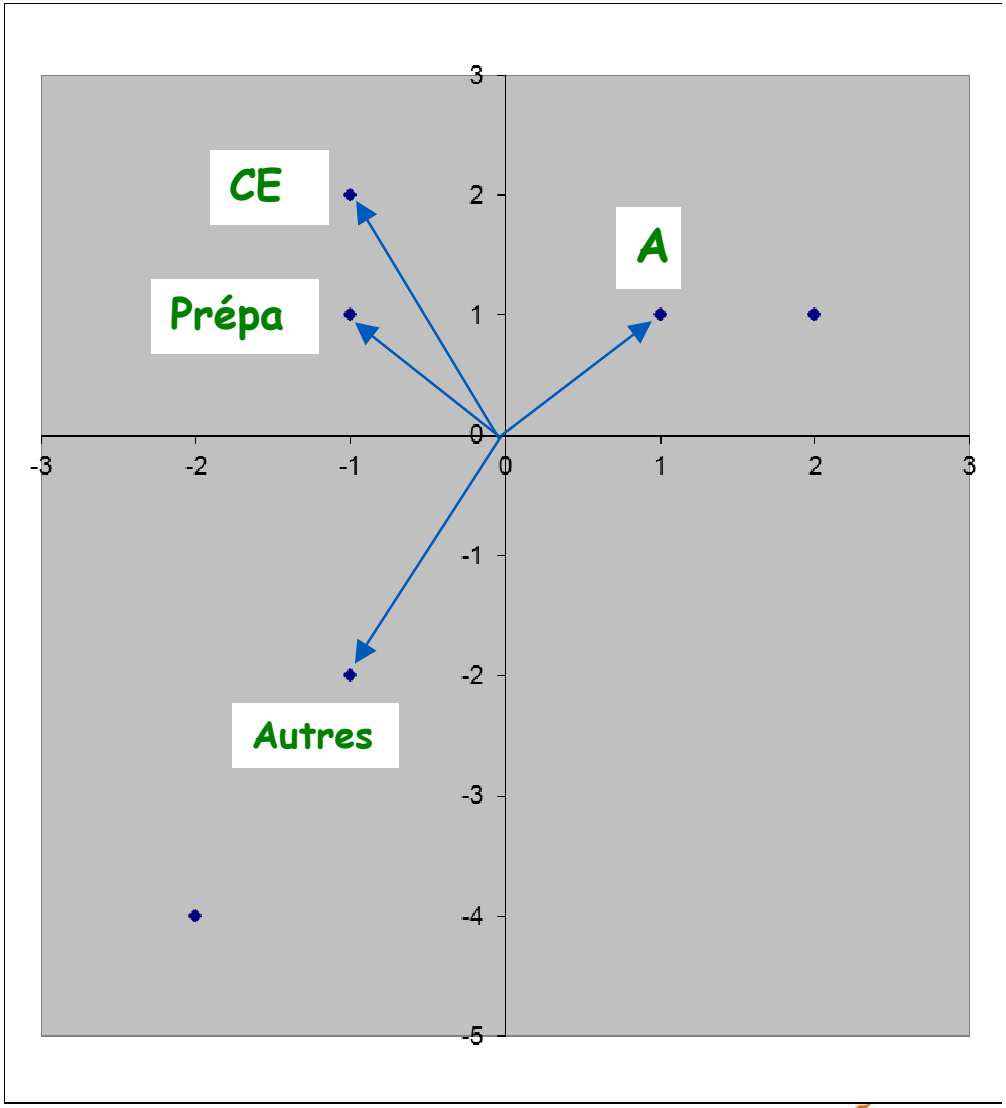
   Produit scalaire négatif

   Les Bacs A ne vont pas vers les « autres » (IUT, BTS)

3. Quadrature

   Produit scalaire nul

   Les bacs A ne vont ni plus ni moins vers les prépas que la moyenne des bacheliers



Question ouverte

$$R = T_1 + T_2 = C_1L_1 + C_2L_2$$

Quelle est la meilleure décomposition possible pour R ?

Quel est le critère permettant de définir les meilleurs T_1 et T_2 ?

L'idée sera de trouver séquentiellement le T_1 qui exprime le plus de sens, puis le T_2, …

Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. **Exemple**
3. Formalisation mathématique de l'AFC
4. Exemple

5. Analyse des correspondances multiples
6. Classification
7. Synthèse

8. Évaluation

Simulation d'accidents graves

Quantification de l'influence de certaines actions (appoints d'eau) lors d'un accident grave dans un réacteur nucléaire (fusion du cœur)

Scenario :

Dégradation coeur, transfert et interaction du corium (en cuve et hors cuve)

23 variables de sortie :

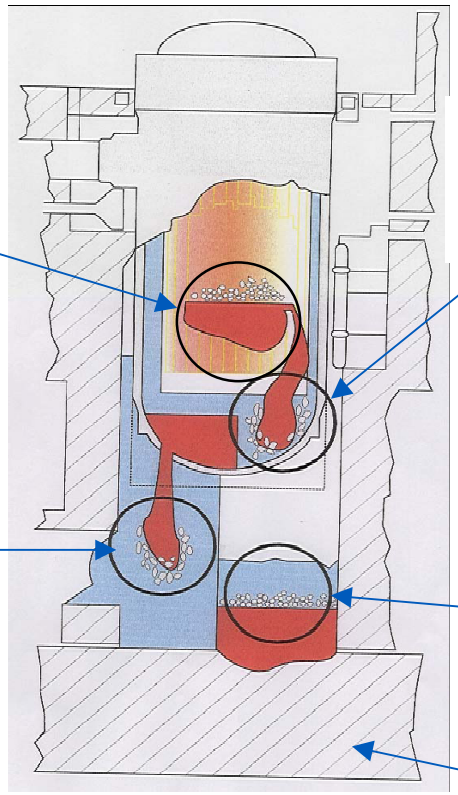
Masses de corium

Temps de percement de la cuve

Temps de percement du radier

32 variables d'entrée aléatoires (lois uniformes) :

Gestion de l'eau, propriétés physique, variables scénario, …



Analyse de 2 variables catégorielles

On analyse souvent des tableaux *individus x variables*

	Masse corium	Eau en cuve	Instant arrivée eau	Débit eau cuve	Percement cuve	Temps percement
1	11	oui	1000	1	oui	1500
2	15	non	NA	NA	oui	1000
3	15	oui	1000	5	non	NA
…						

Pour l'analyse bivariable de variables catégorielles, on utilise le

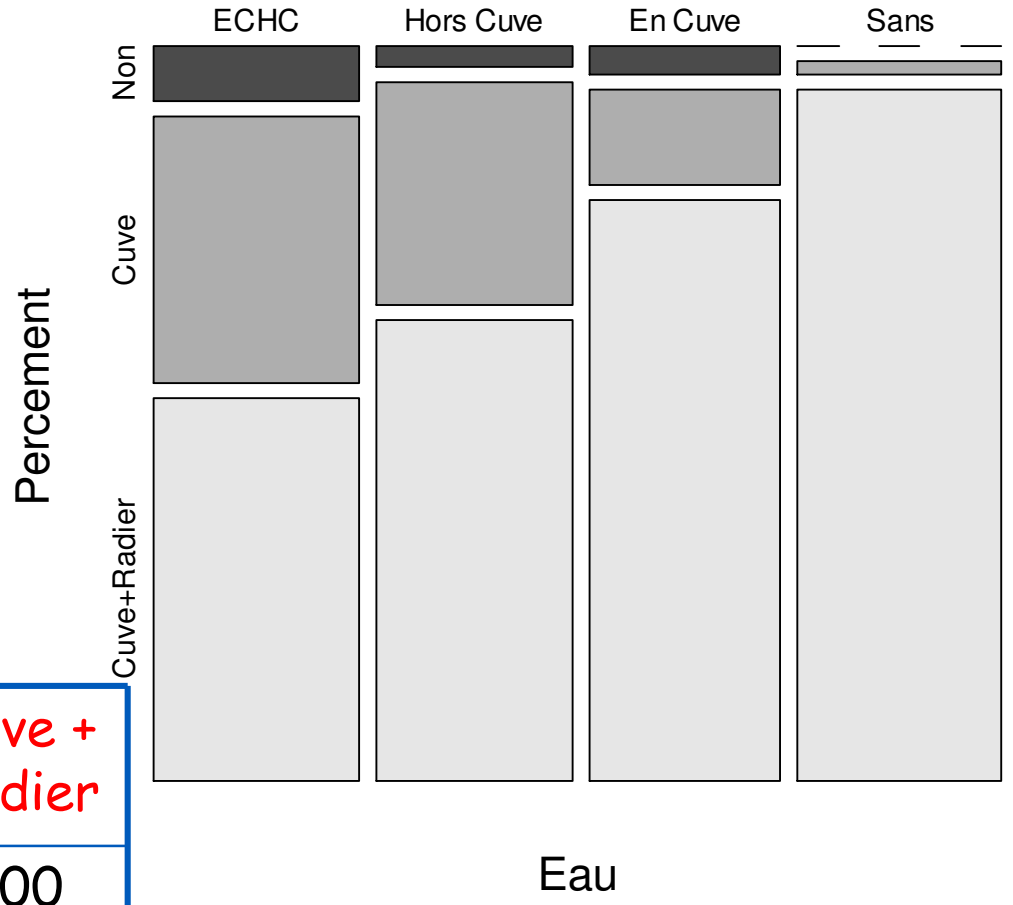
tableau de contingence

On regroupe les individus
Lignes = modalités 1ère variable
Colonnes = modalités 2ème variable

Percement / Eau	Non	Cuve	Cuve + Radier
Sans	0	2	100
En cuve	4	13	79
Hors cuve	3	31	64
En cuve + Hors cuve	8	39	56

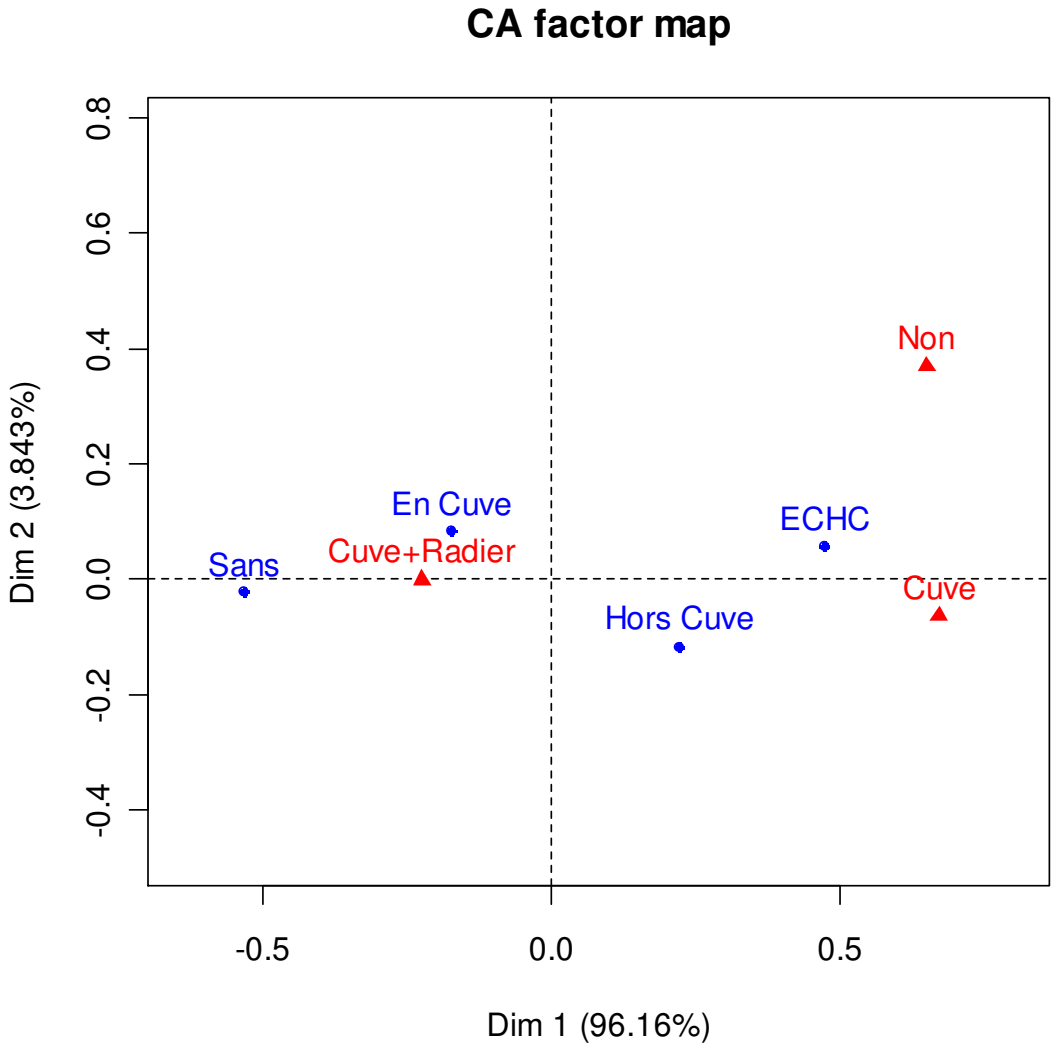
Analyse graphique

Mosaic plot



Percement / Eau	Non	Cuve	Cuve + Radier
Sans	0	2	100
En cuve	4	13	79
Hors cuve	3	31	64
En cuve + Hors cuve	8	39	56

Résultat d'une analyse des correspondances



The chi square of independance between the two variables is equal to 62.1863

The p-value associated to this chi square is equal to 1.6167e-11 .

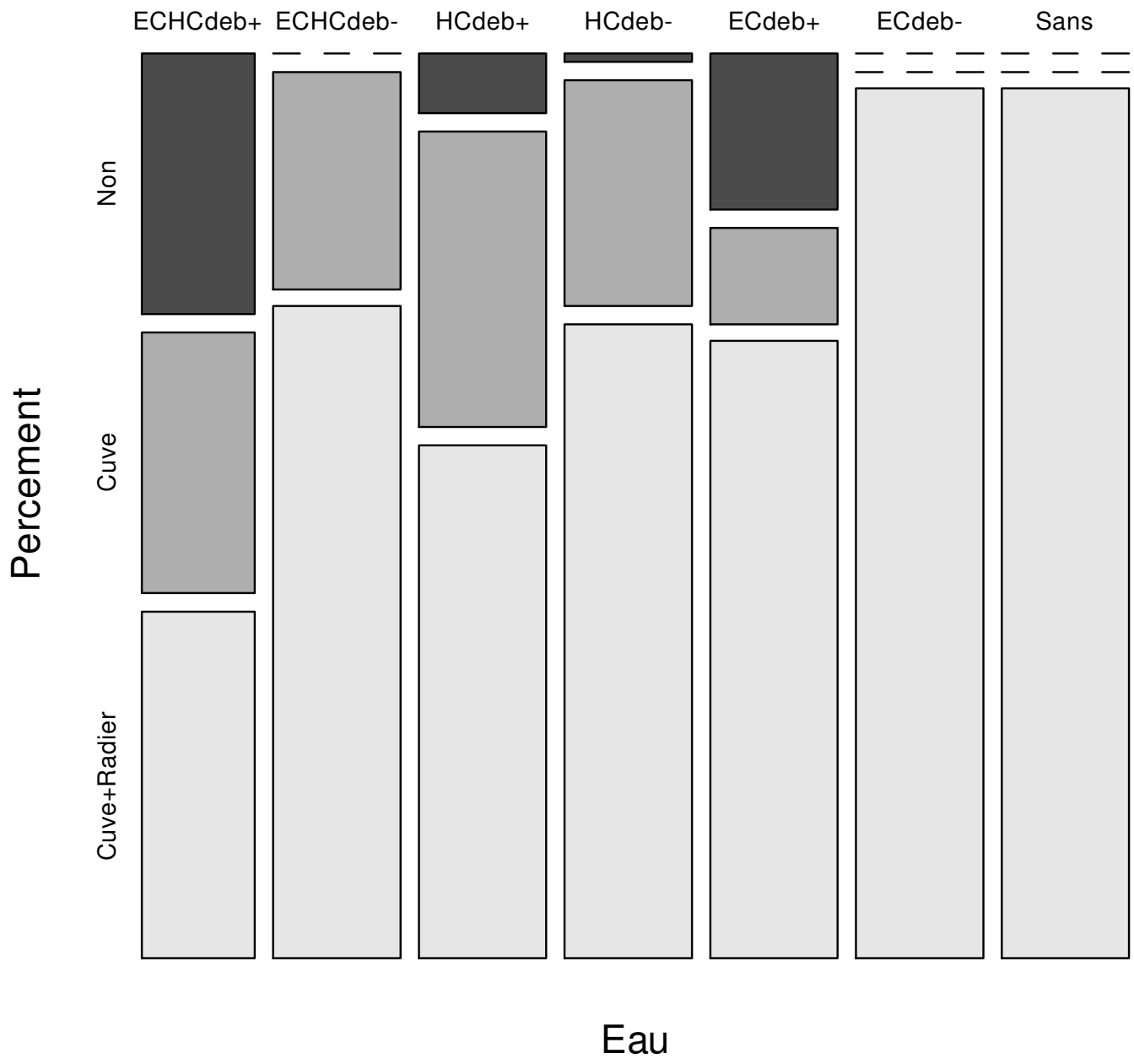
Conclusion : on rejette l'hyp. que les variables sont indépendantes avec un risque négligeable

- Analyse statistique quantitative du tableau de contingence
- Test statistique associé pour mesurer l'indépendance entre les 2 variables

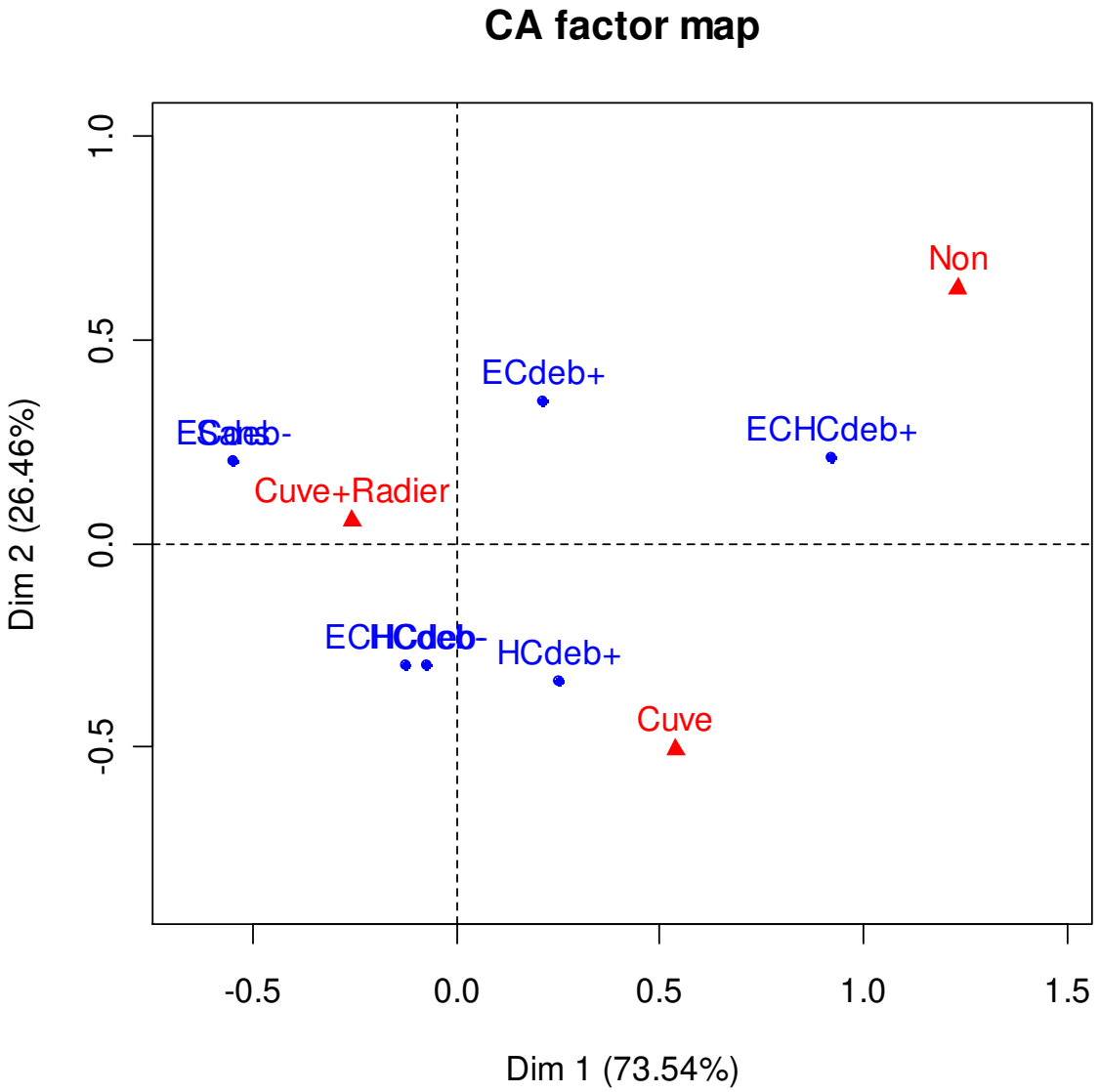
Exemple avec plus de modalités (1/3)

Percement / Eau	Non	Cuve	Cuve + Radier
Sans	0	0	100
En cuve petit débit	0	0	100
En cuve gros débit	18	11	71
Hors cuve petit débit	1	26	73
Hors cuve gros débit	7	34	59
En cuve + Hors cuve petits débits	0	25	75
En cuve + Hors cuve gros débits	27	27	36

Exemple avec plus de modalités (2/3)



Exemple avec plus de modalités (3/3)



The chi square of independance between the two variables is equal to 203.4634

The p-value associated to this chi square is equal to 6.284089e-37 .

Conclusion : on rejette l'hyp. que les variables sont indépendantes avec un risque négligeable

Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. Exemple
3. **Formalisation mathématique de l'AFC**
4. Exemple

5. Analyse des correspondances multiples
6. Classification
7. Synthèse

8. Évaluation

Tableau de contingence (ou tableau des effectifs)

Y	1	$\cdots$	j	$\cdots$	J	marges en ligne
X						
1	n_{11}				n_{1J}	
$\vdots$						
i			n_{ij}			$n_{i.} = \sum_j n_{ij}$
$\vdots$						
I	n_{I1}				n_{IJ}	
marges en colonne			$n_{.j} = \sum_i n_{ij}$			$n = \sum_{ij} n_{ij}$

Rôle symétrique des lignes et des colonnes  ACP

Tableau de contingence sur les fréquences

Y	1	$\cdots$	j	$\cdots$	J	marges
X						
1	f_{11}				f_{1J}	
$\vdots$						
i			$f_{ij} = \frac{n_{ij}}{n}$			$f_{i.} = \sum_j f_{ij}$
$\vdots$						
I	f_{I1}				f_{IJ}	
marges			$f_{.j} = \sum_i f_{ij}$			$f = 1$

• Question symétrique : l'appartenance à une modalité de X est elle liée à l'appartenance à une modalité de Y ?

• Question dissymétrique : quelles sont les modalités de X préférées d'une modalité de Y donnée (et réciproquement) ?

Intérêts et principes de l'AFC

- Problème : La lecture du tableau devient fastidieuse quand il y a beaucoup de modalités

- Outil AFC : visualisation en 2 dimensions des tableaux de contingence
Transformation de variables qualitatives en variables quantitatives

- Intérêts :
  - Etude des liens entre les modalités de chaque variable (typologie des lignes et typologie des colonnes)
  - Etude des corrélations entre les modalités des 2 variables

- AFC = ACP avec une métrique particulière (celle du χ^2 pondéré)

Comparaison de deux modalités de X en comparant leurs fréquences sur l'ensemble des modalités de Y

- Avantage par rapport à l'ACP : on superpose les modalités (les 2 dimensions de la matrice) dans un même plan

Test d'indépendance

Test :

H_0 : Les variables X et Y sont indépendantes

H_1 : Les variables X et Y sont liées entre elles

Statistique du χ^2 :

$$n_{ij} = \text{effectif observé}$$

$$\tilde{n}_{ij} = \frac{n_{i.} n_{.j}}{n} = \text{effectif attendu sous hypothèse d'indépendance}$$

$$\chi_{ij} = \frac{n_{ij} - \tilde{n}_{ij}}{\sqrt{\tilde{n}_{ij}}} = \text{résidu standardisé}$$

$$\chi^2 = \sum_{ij} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}} = n \sum_{ij} \frac{\left(f_{ij} - f_{i.} f_{.j}\right)^2}{f_{i.} f_{.j}} = n\phi^2$$

ϕ^2 mesure l'intensité de la relation entre les variables

Test d'indépendance

Si α est le risque associé au test, avec un niveau de confiance 1-α (par exemple 95 %), **on rejette ou on ne rejette pas l'hypothèse H_0**

Test du khi-deux (Pearson) : $D^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \sim \chi^2(k-1)$ où k est le nb de classes

Comparaison entre fréquences observées N_i et théoriques p_i

Remarque :
Test peu puissant et pas robuste pour de petits échantillons (< 50)

Pour les tableaux de contingence : $\chi^2_{obs} \sim \chi^2[(I-1)(J-1)]$

Décision :
On rejette H_0 au risque α de se tromper si $\chi^2_{obs} \geq \chi^2_{1-\alpha}[(I-1)(J-1)]$

Exemple – Test d'indépendance

Tableau de contingence

Percement / Eau	Non	Cuve	Radier	$n_{i.}$
Sans	0	2	100	102
En cuve	4	13	79	96
Hors cuve	3	31	64	98
ECHC	8	39	56	103
$n_{.j}$	15	85	299	n = 399

Résultats attendus sous hypothèse d'indépendance

Percement / Eau	Non	Cuve	Radier
Sans	3.8	21.7	76.4
En cuve	3.6	20.5	71.9
Hors cuve	3.7	20.9	73.4
ECHC	3.9	21.9	77.2

6 degrés de liberté , $\chi^2_{1-\alpha}(6) = 12.59$ pour $\alpha = 0.05$

$\chi^2_{obs} = 62.19$, p - value $= 1.62\text{E}-11$, Rejet de l'hyp. d'indépend.

Etude des relations entre modalités

$$i\begin{bmatrix} & j & \\ & p_j^i = \dfrac{f_{ij}}{f_{i.}} & \\ & & \end{bmatrix}\begin{matrix}\text{marges}\\ \vdots \\ 1 \\ \vdots\end{matrix}$$

$$p_{.j}$$

Tableau des profils lignes

(on normalise les marges des lignes pour pouvoir les comparer)

$$i\begin{bmatrix} & j & \\ & p_i^j = \dfrac{f_{ij}}{f_{.j}} & \\ & & \end{bmatrix} p_{i.}$$

$$\text{marges} \quad \cdots \quad 1 \quad \cdots$$

Tableau des profils colonnes

Distances du χ^2 entre deux modalités de *X* (profils lignes) et deux modalités de *Y* (profils colonnes) :

$$d_{\mathrm{PL}}^2(i,i') = \sum_{j=1}^{J} \frac{\left(p_j^i - p_j^{i'}\right)^2}{p_{.j}} \quad \text{et} \quad d_{\mathrm{PC}}^2(j,j') = \sum_{i=1}^{I} \frac{\left(p_i^j - p_i^{j'}\right)^2}{p_{i.}}$$

Cette métrique permet de donner autant d'importance à chacune des modalités

Inertie totale = dispersion des profils lignes (resp. colonnes) / barycentre = ϕ^2

Exemple - Profils

$n = 399$

Percement / Eau	Non	Cuve	Radier
Sans	0	2	100
En cuve	4	13	79
Hors cuve	3	31	64
ECHC	8	39	56

Profils lignes

Perct / Eau	Non	Cuve	Radier
Sans	0	.0196	.9804
En cuve	.0417	.1354	.8229
Hors cuve	.0306	.3163	.6531
ECHC	.0777	.3786	.5437

Profils colonnes

Perct / Eau	Non	Cuve	Radier
Sans	0	.0235	.3344
En cuve	.2667	.1529	.2642
Hors cuve	.2	.3647	.2140
ECHC	.5333	.4588	.1873

Exemple – Distances entre profils lignes

Eau	Sans	EC	HC	ECHC
Sans	0			
EC	.0356	0		
HC	.1455	.0489	0	
ECHC	.2554	.1042	.0233	0

Profils proches
Profils éloignés

Avec plus de modalités, l'interprétation est difficile

Il faut donc synthétiser l'information

Description d'un ensemble de profils par ACP

ACP → facteurs principaux (axes principaux + composantes principales)

- En ACP, on cherche des dimensions qui représentent au mieux la variabilité entre individus

En AFC, on cherche des dimensions qui représentent au mieux l'écart des données à l'indépendance

- ACP pondérée des profils lignes → F_1, F_2 , …
- ACP pondérée des profils colonnes → G_1, G_2 , …

Propriétés :

- Inertie = variance des facteurs
- Somme des inerties = ϕ^2
- Valeurs propres < 1
- Pourcentage de variance de chaque axe $\lambda_s/\Sigma\lambda_s$ où λ_s valeur propre de F_s
- Nb d'axes < min(I ,J) - 1

Dualité des 2 ACPs

1. Les 2 ACPs conduisent aux mêmes valeurs propres

2. Les facteurs principaux de l'une sont les composantes principales de l'autre

Formules de transitions :

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_j \frac{f_{ij}}{f_{i.}} G_s(j) \text{ pour } \quad i = 1 \ldots I$$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_i \frac{f_{ij}}{f_{.j}} F_s(i) \text{ pour } \quad j = 1 \ldots J$$

Conséquence : On ne réalise qu'une ACP (celle avec le moins de modalités)

Moyennes et variances des facteurs :

- Moyennes nulles
- Variances égales aux valeurs propres λs = Var(Fs) = Var(Gs)
- Covariances nulles

Interprétation

- Poids :

  chaque profil intervient d'autant plus dans l'analyse que son poids $p_{i.}$ *ou* $p_{.j}$ *est élevé*

- Contributions :

  mesurent l'influence des profils dans le calcul des axes principaux (inertie d'un point / inertie totale λ)

- Qualité de la représentation : $\cos^2$

  $\cos^2$ fort ⟶ point fortement expliqué et bien représenté par l'axe principal

Exemple : facteurs principaux

2 axes principaux

	eigenvalue	Percent. of var	Cumulative percent. of var
dim 1	1.498660e-01	9.615710e+01	96.1571
dim 2	5.989366e-03	3.842900e+00	100.0000

Profils lignes :

$contrib

	Dim 1	Dim 2
ECHC	38.712890	13.666212
Hors Cuve	8.180645	56.771183
En Cuve	4.668735	27.543608
Sans	48.437729	2.018997

$cos2

	Dim 1	Dim 2
ECHC	0.98608	0.0139118
Hors Cuve	0.78287	0.2171252
En Cuve	0.80920	0.1907917
Sans	0.99833	0.0016630

Profils colonnes :

$contrib

	Dim 1	Dim 2
Non	10.63364	85.60696
Cuve	64.30618	14.39056
Cuve+Radier	25.06018	0.00247

$cos2

	Dim 1	Dim 2
Non	0.75657	2.43421e-01
Cuve	0.99113	8.86412e-03
Cuve+Radier	0.99999	3.94618e-06

Représentation graphique

- **Interprétation des valeurs des modalités :**
  forte valeur absolue (influence de la modalité) / barycentre (non influence)

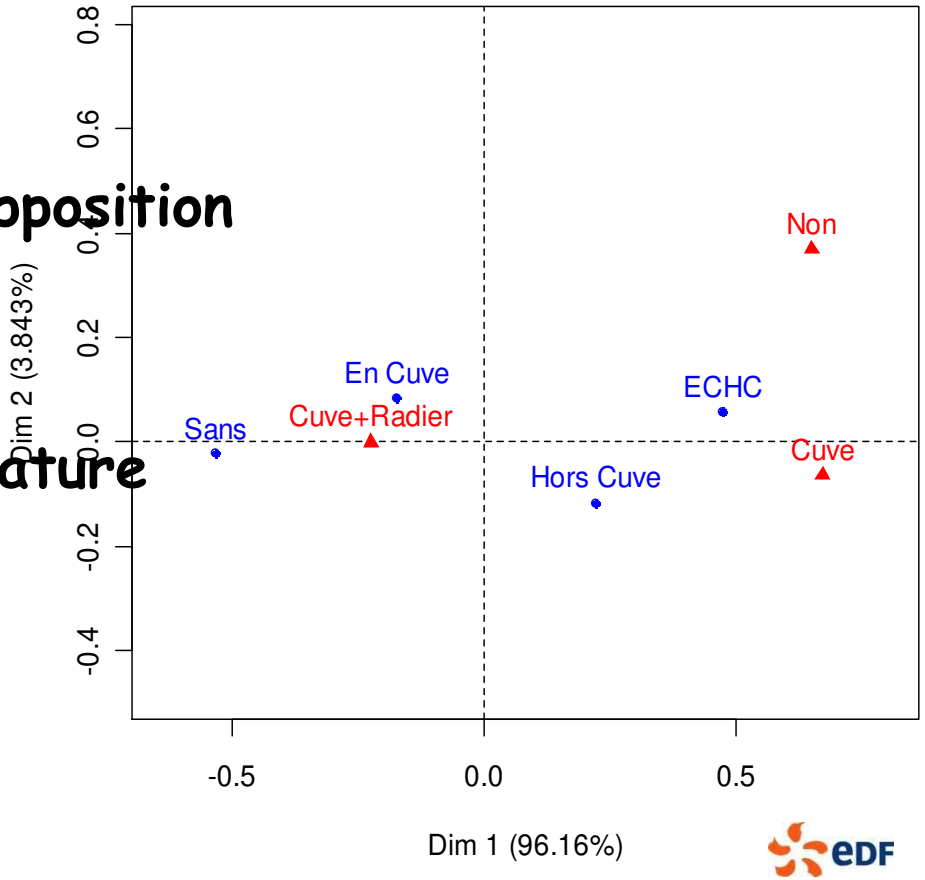
- **Produits scalaires positifs : Conjonction**
  - Les non percements sont liés à la présence d'eau hors cuve
  - Les modalités « Non percement » et « Percement Cuve » sont similaires vis-à-vis de l'« Eau »
  - …

- **Produits scalaires négatifs : Opposition**
  - Le non percement de la cuve

  est antinomique à l'absence d'eau

- **Produits scalaires nuls : Quadrature**
  - La présence d'eau en cuve

  n'entraîne pas plus de non percement que la moyenne des expériences



Aspects logiciels

- AFC se trouve dans la plupart des logiciels de statistiques

- Packages R avec AFC :
  MASS :: corresp
  Package FactoMineR (AgroCampus Rennes)

Liens entre AFC et ACP

	AFC	ACP
Données	Catégorielles	Métriques
Décomposition	T - T0 = T1 + T2	T = T1 + T2 + T3
Métrique	χ^2 pondéré	χ^2

Attention, le poids des cellules à faible effectif est renforcé

Rapports entre ACP et AFC

- Si on a des données permettant de faire une AFC, peut-on y appliquer une ACP ?
  - Non

- Si on a des données permettant de faire une ACP, peut-on y appliquer une AFC ?
  - Oui !

- .. Mais alors ?
  - .. Alors on traite les données numériques, les nombres comme des catégories
  - Si par exemple on travaille sur des notes, 18/20 n'est plus « supérieur à » 10/20, il n'est pas non plus « plus proche » de 16/20 que de 10/20.

Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. Exemple
3. Formalisation mathématique de l'AFC
4. **Exemple**

5. Analyse des correspondances multiples
6. Classification
7. Synthèse

8. Évaluation

Exemple 3 : enquête sur les séjours-vacances des français

Données publiées par l'INSEE en 2002 et étudiées dans *Saporta, 2006*

CSP

$n = 18532$

TABLEAU 6.3 Tableau de contingence

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	41	47	13	59	17	26	4	9	19
Artisans, commerçants, chefs d'entreprise	220	260	71	299	120	42	64	35	29
Cadres et professions intellectuelles supérieures	685	775	450	1242	706	139	122	100	130
Professions intermédiaires	485	639	292	1250	398	189	273	68	193
Employés	190	352	67	813	163	92	161	49	72
Ouvriers	224	591	147	1204	181	227	306	74	114
Retraités	754	393	692	1158	223	25	195	47	115
Autres inactifs	31	34	2	225	42	33	5	6	14

Mode d'hébergement

Rsec = rés. 2ndaire

Rppa = rés. principale parents amis

Rspa = rés. 2ndaire parents amis

Y-a-t-il un lien entre la CSP à 8 modalités et le mode de vacances à 9 modalités?

Tableaux des profils

Profils lignes

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV	Total
Agriculteurs	0.174	0.200	0.055	0.251	0.072	0.111	0.017	0.038	0.081	1
Artisans, commerçants, chefs d'entreprise	0.193	0.228	0.062	0.262	0.105	0.037	0.056	0.031	0.025	1
Cadres et professions intellectuelles supérieures	0.158	0.178	0.103	0.286	0.162	0.032	0.028	0.023	0.030	1
Professions intermédiaires	0.128	0.169	0.077	0.330	0.105	0.050	0.072	0.018	0.051	1
Employés	0.097	0.180	0.034	0.415	0.083	0.047	0.082	0.025	0.037	1
Ouvriers	0.073	0.193	0.048	0.392	0.059	0.074	0.100	0.024	0.037	1
Retraités	0.209	0.109	0.192	0.321	0.062	0.007	0.054	0.013	0.032	1
Autres inactifs	0.079	0.087	0.005	0.574	0.107	0.084	0.013	0.015	0.036	1

Profils colonnes

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	0.016	0.015	0.007	0.009	0.009	0.034	0.004	0.023	0.028
Artisans, commerçants, chefs d'entreprise	0.084	0.084	0.041	0.048	0.065	0.054	0.057	0.090	0.042
Cadres et professions intellectuelles supérieures	0.260	0.251	0.260	0.199	0.38`2	0.180	0.108	0.258	0.190
Professions intermédiaires	0.184	0.207	0.168	0.200	0.215	0.245	0.242	0.175	0.281
Employés	0.072	0.114	0.039	0.130	0.088	0.119	0.142	0.126	0.105
Ouvriers	0.085	0.191	0.085	0.193	0.098	0.294	0.271	0.191	0.166
Retraités	0.287	0.127	0.399	0.185	0.121	0.032	0.173	0.121	0.168
Autres inactifs	0.012	0.011	0.001	0.036	0.023	0.043	0.004	0.015	0.020
Total	1	1	1	1	1	1	1	1	1

Effectifs théoriques et contributions au χ^2

	Hotel	Location	Rsec	Rppa	Rspa	Tente	Caravane	AJ	VillageV
Agriculteurs	33.35 1.75	39.2 1.55	21.99 3.67	79.25 5.18	23.46 1.78	9.8 26.77	14.33 7.45	4.92 3.38	8.7 12.2
Artisans, commerçants, chefs d'entreprise	161.79 20.95	190.14 25.66	10.67 11.93	384.47 19	113.8 0.34	47.55 0.65	69.51 0.44	23.87 5.19	42.2 4.13
Cadres et professions intellectuelles supérieures	617.2 7.45	725.8 3.39	406.93 4.56	1466.72 34.43	434.15 170.22	181.4 9.91	265.18 77.31	91.05 0.88	160.99 5.96
Professions intermédiaires	537.44 5.12	631.64 0.09	354.34 10.97	1277.18 0.58	378.05 1.05	157.96 6.10	230.91 7.67	79.29 1.61	140.18 19.9
Employés	278.01 27.86	326.75 1.95	183.3 73.79	660.68 35.12	195.56 5.42	81.71 1.3	119.45 14.45	41.02 1.55	75.52 0.00
Ouvriers	435.4 102.64	511.72 12.28	287.07 68.34	1034.7 27.70	306.7 51.24	127.97 76.63	187.07 75.6	64.23 1.48	113.57 0.00
Retraités	511.18 115.34	600.79 71.86	337.03 373.86	1214.79 2.65	359.68 51.88	150.25 104.41	219.63 2.76	75.41 10.71	133.34 2.52
Autres inactifs	55.63 10.91	65.38 15.06	36.68 32.79	132.2 65.14	39.13 0.21	16.35 16.95	23.9 14.95	8.21 0.59	14.51 0.02

Écart à l'indépendance

$\chi^2 = 27.6$

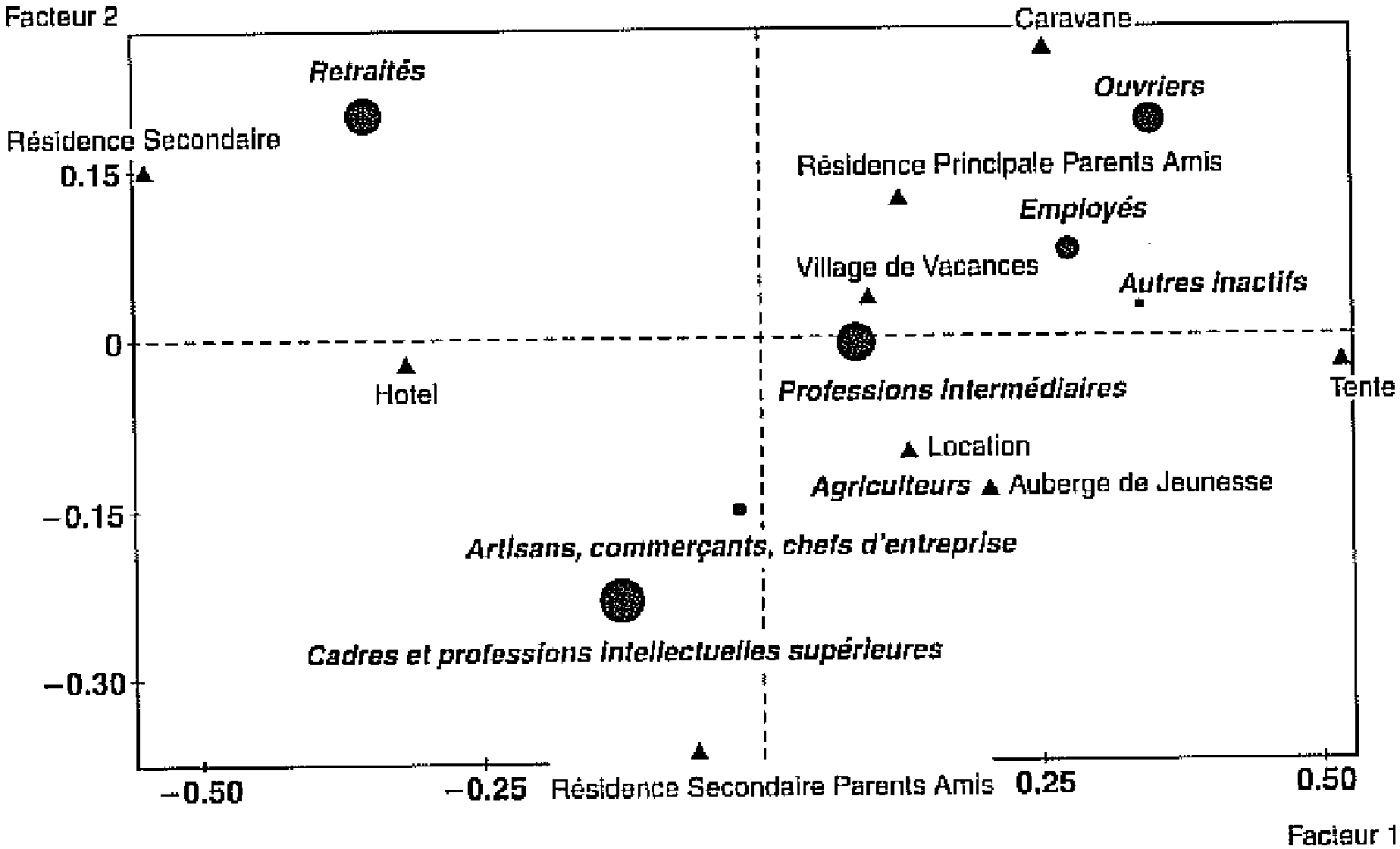
Contribution au χ^2 > 60

Application de l'AFC

Valeurs propres :

1. 0.0657 – contribution à la variance = 61.24 %
2. 0.0254 – contribution à la variance = 23.70 %
3. 0.0081 – contribution à la variance = 7.55 %
4. 0.0037 – contribution à la variance = 3.46 %
5. 0.0028 – contribution à la variance = 2.60 %
6. 0.0014 – contribution à la variance = 1.29 %
7. 0.0012 – contribution à la variance = 0.16 %

Graphe de l'analyse des correspondances



Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. Exemple
3. Formalisation mathématique de l'AFC
4. Exemple

5. Analyse des correspondances multiples

6. Classification
7. Synthèse

8. Évaluation

ACM – Analyse des correspondances multiples

Objectif : Même objectif que l'ACP ou AFC ! Visualiser, résumer l'information contenue dans des masses volumineuses de données

Les données type :

- N individus décrit par q variables nominales ou ordinales X_1, …, X_q

Exemple INSEE : réponse à une enquête basée sur des questions fermées à choix multiples (catégories d'âge, de revenus, activités sportives, …)

L'ACM vise à mettre en évidence :

- **Les relations entre les modalités des différentes variables**
- **Les relations entre les individus statistiques**
- **Les relations entre les variables telles qu'elles apparaissent à partir des relations entre modalités**

ACM - exemple

Enquête auprès de 383 étudiants

5 questions en rapport avec le logement étudiant.

Modalité représentant un faible nombre d'étudiants.

Question	N°	Réponses possibles	Poids (%)	Abréviation
Habitez-vous (variable "mode d'occupation")	1	seul	48,30%	Seul
	2	colocataires	13,84%	Coloc
	3	en couple	13,05%	Couple
	4	avec les parents	23,50%	Parents
	5	non réponse	1,31%	NR1
Quel type d'habitation occupez-vous ? (variable "type d'habitation")	6	cité universitaire	10,70%	Cité
	7	studio	28,20%	Studio
	8	appartement	30,29%	Appart
	9	chambre chez un particulier	5,22%	Chambre
	10	autre	19,84%	Autre
	11	non réponse	5,74%	NR2
Si vous vivez en dehors du foyer familial, depuis combien de temps ? (variable "ancienneté")	12	moins de 1 an	20,89%	< 1 an
	13	1 à 3 ans	24,80%	1-3 ans
	14	plus de 3 ans	28,72%	> 3 ans
	15	non applicable	24,80%	NA
	16	non réponse	0,78%	NR3
A quelle distance approximative de la Fac vivez-vous ? (variable "éloignement")	17	moins de 1 km	26,89%	< 1 km
	18	1 à 5 km	49,87%	1 à 5 km
	19	plus de 5 km	20,89%	> 5 km
	20	non réponse	2,35%	NR4
Quelle est la superficie de votre logement ? (variable "superficie")	21	moins de 10 m^2	9,14%	< 10 m^2
	22	10 à 20 m^2	17,75%	10 à 20 m^2
	23	20 à 30 m^2	24,80%	20 à 30 m^2
	24	plus de 30 m^2	39,16%	> 30 m^2
	25	non réponse	9,14%	NR5

ACM – Démarche pour l'analyse

Données qualitatives et quantitatives

Données qualitatives

Construction du tableau disjonctif complet

- comporte une colonne pour chaque modalité des variables étudiées, et une ligne pour chaque individu statistique. Les cellules du tableau contiennent 1 ou 0 selon que l'individu considéré présente la modalité correspondante ou non

Construction du tableau de BURT

- chaque cellule du tableau indique le nombre d'individus statistiques qui possèdent en même temps la modalité ligne et la modalité colonne correspondante

Le travail se fait ensuite à partir de l'un des 2 tableaux

Les 2 analyses conduisent à des résultats analogues mais pas identiques

On recherche une représentation graphique en plus faible dimension des lignes et des colonnes de la matrice de Burt

ACM - Tableau disjonctif complet

Z = tableau disjonctif complet ; p = nombre total de modalités
La somme des éléments d'une même ligne est constante et vaut q (=nb de variables)
La somme de tous les éléments du tableau vaut nq
La somme des éléments d'une même colonne n'est pas constante mais égale à l'effectif possédant la modalité j_k de la variable k considérée

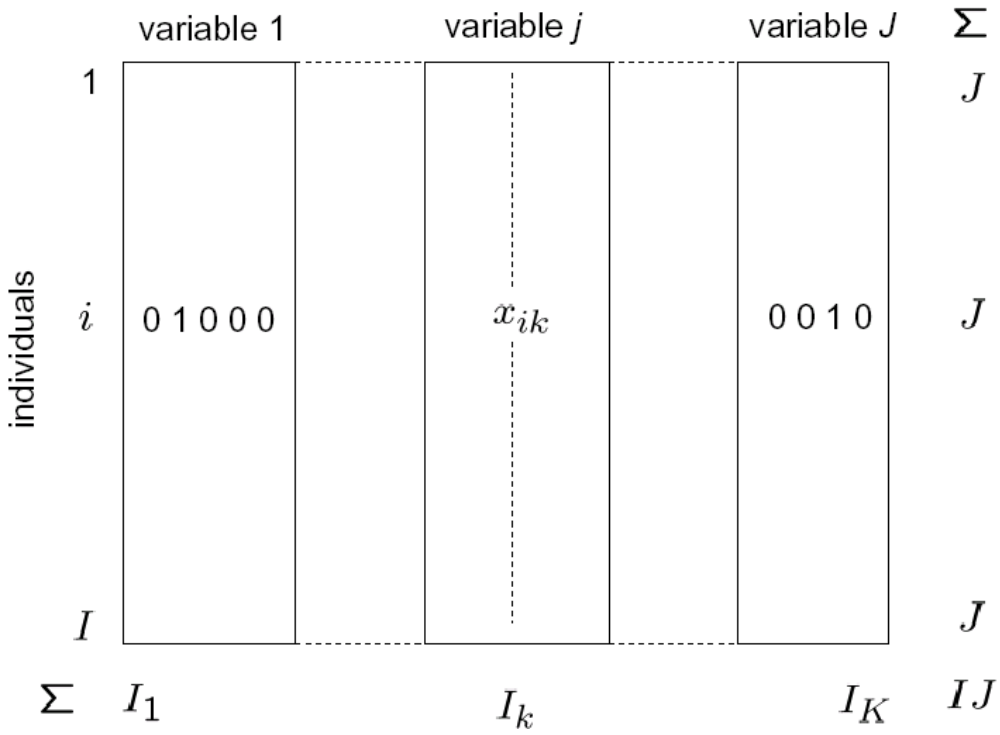
	1…	Variable k	…q	
	1…	Modalités …	…p	
1	…		…	
…				
i	…	z_{ij}	…	$z_{i.} = q$
…				
n	…		…	
		$z_{.j} = n_j$		$z.. = nq$

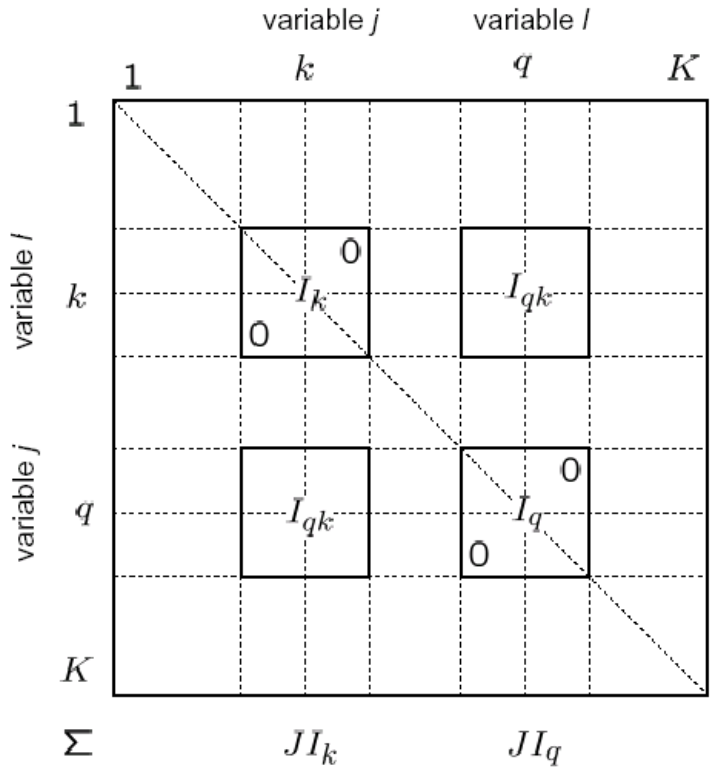
ACM - Tableau de BURT

On appelle tableau de Burt, noté B, le produit du transposé de Z par Z soit

$$B = {}^{t}ZZ$$

Le tableau de Burt est carré et sa taille est égale au nombre de modalités K possédées par les J variables





ACM - Tableau de BURT

Tableau qui présente de nombreuses propriétés remarquables :

- **Symétrie : $n_{ij} = n_{ji}$**
- Les encadrés situés le long de la diagonale principale (haut à gauche vers le bas à droite) donnent les effectifs correspondant à chaque modalité
- La ligne j du tableau est la somme des lignes du tableau disjonctif complet correspondant aux individus qui possèdent la modalité j

Ce tableau peut être vue comme une juxtaposition de tableaux de contingence.

Il est obtenu facilement à partir du tableau disjonctif complet

En revanche il n'existe pas de moyen simple permettant de recomposer le tableau disjonctif complet à partir du tableau de Burt

ACM – Les tableaux de données

Tableau brut de données : pour chaque étudiant, j'ai la réponse aux 5 questions

```
      Mode      Type      Ancien       Eloign         Superf
[1,] "Par_NR" "Autre"  "NA_NR"      "plus_5km_NR" "plus_30m"
[2,] "Par_NR" "Autre"  "NA_NR"      "de_1_5km"    "NR5"
[3,] "Par_NR" "Autre"  "NA_NR"      "plus_5km_NR" "NR5"
[4,] "Par_NR" "Autre"  "NA_NR"      "plus_5km_NR" "NR5"
[5,] "Par_NR" "Autre"  "NA_NR"      "plus_5km_NR" "NR5"
[6,] "Par_NR" "Autre"  "NA_NR"      "plus_5km_NR" "NR5"
                      ... ... ... ...
[380,] "Couple" "Autre"  "de_1_3ans" "de_1_5km"    "plus_30m"
[381,] "Seul"   "NR2"    "plus_3ans" "plus_5km_NR" "moins_10m"
[382,] "Coloc"  "NR2"    "moins_1an" "de_1_5km"    "plus_30m"
[383,] "Par_NR" "Autre"  "de_1_3ans" "plus_5km_NR" "NR5"
```

Tableau disjonctif complet

```
library(ade4)

etudiants.disj <- acm.disjonctif(etudiant)
```

```
> etudiants.disj
    .Coloc .Couple .Par_NR .Seul .Appart .Autre .Chamb .Cite .NR2 .Studio .de_1_3ans .moins_1an .NA_NR
1        0       0       1     0       0      1      0     0     0       0          0          0      1
2        0       0       1     0       0      1      0     0     0       0          0          0      1
3        0       0       1     0       0      1      0     0     0       0          0          0      1
4        0       0       1     0       0      1      0     0     0       0          0          0      1
5        0       0       1     0       0      1      0     0     0       0          0          0      1
6        0       0       1     0       0      1      0     0     0       0          0          0      1
                                          ... ... ... ...
380      0       1       0     0       0      1      0     0     0       0          1          0      0
381      0       0       0     1       0      0      0     0     1       0          0          0      0
382      1       0       0     0       0      0      0     0     1       0          0          1      0
383      0       0       1     0       0      1      0     0     0       0          1          0      0
```

Le tableau représente en colonne chaque modalité des variables
1=modalité répondue par l'étudiant sinon 0

ACM - Démarche

Le traitement opéré sur les données du tableau de Burt est identique à celui opéré sur un tableau de contingence lors d'une AFC

Comme pour une AFC, on obtiendra :

- Des axes factoriels associés à des valeurs propres
- Pour chaque ligne ou colonne du tableau de Burt, des coordonnées, des contributions à la formation des axes et des qualités de représentation

ACM – Inerties - Définitions

L'inertie totale du nuage des modalités est déterminée uniquement par le nombre total de modalités p et le nombre de variables q

- **Inertie totale du nuage des modalités** : $I = \sum_{j=1}^{p} I(X_j) = \frac{p-q}{q}$

q étant le nombre total de variables et p le nombre total de modalités

- **Inertie d'une variable** $I(X_j) = \frac{p_j - 1}{q}$

p_j est le nombre de modalités de la variable X_j

- **Inertie d'une modalité** $I(X_j = a_k) = \frac{n - n_k}{nq}$

n_k étant l'effectif de la modalité a_k de la variable X_j

A noter que plus une modalité est rare et plus son influence globale est élevée

ACM - Démarche

Sélection des axes

- règle courante : garder les axes tels que $\mu_k > 1/p$ (la moyenne des valeurs propres est 1/p)

- les axes intéressants sont ceux que l'on peut interpréter, en regardant les contributions des variables actives et les valeurs-tests associées aux variables supplémentaires

- En pratique on se contente souvent d'interpréter le premier plan principal

L'inertie expliquée est moins intéressante qu'en ACP

ACM – Inertie – Application à l'exemple

Le nombre total de modalités p=25 et le nombre de variables q=5

- **Inertie totale du nuage des modalités** $I = \sum_{j=1}^{p} I(X_j) = \frac{25-5}{5} = 4$
- **Ex : inertie de la première variable** $I(X_j) = \frac{5-1}{5} = 0.8$
- **Inertie d'une modalité** $I(X_j = a_k) = \frac{n - n_k}{nq}$

n_k étant l'effectif de la modalité a_k de la variable X_j

ACM – Inertie – Application à l'exemple

L'analyse du tableau de Burt produit au plus p-q valeurs propres non nulles

La décroissance de ces valeurs propres est beaucoup moins rapides que dans le cas de l'AFC (difficile d'énoncer un critère relatif au nombre d'axes factoriels à conserver)

De ce fait, il arrive souvent que l'on se limite au premier plan factoriel

Nombre de Dims.	Valeurs Propres et Inertie de toutes les Dimensions (Etudiants-ville.sta)				
	Inertie Totale = 4,0000				
	ValSing.	ValProp.	%age Inertie	%age Cumulé	Chi^2
1	0,8505	0,7233	18,08	18,08	2287,23
2	0,6567	0,4313	10,78	28,86	1363,94
3	0,6213	0,3860	9,65	38,51	1220,72
4	0,5713	0,3264	8,16	46,67	1032,18
5	0,5023	0,2523	6,31	52,98	797,87
6	0,4659	0,2171	5,43	58,41	686,46
7	0,4473	0,2001	5,00	63,41	632,69
8	0,4442	0,1974	4,93	68,34	624,11
9	0,4293	0,1843	4,61	72,95	582,70
10	0,4111	0,1690	4,23	77,18	534,56
11	0,4086	0,1670	4,17	81,35	528,09
12	0,3799	0,1443	3,61	84,96	456,46
13	0,3519	0,1238	3,10	88,06	391,59
14	0,3293	0,1084	2,71	90,77	342,94
15	0,3237	0,1048	2,62	93,39	331,34
16	0,2926	0,0856	2,14	95,53	270,79
17	0,2719	0,0739	1,85	97,38	233,74
18	0,2328	0,0542	1,35	98,73	171,32
19	0,1948	0,0380	0,95	99,68	120,03
20	0,1134	0,0129	0,32	100,00	40,69

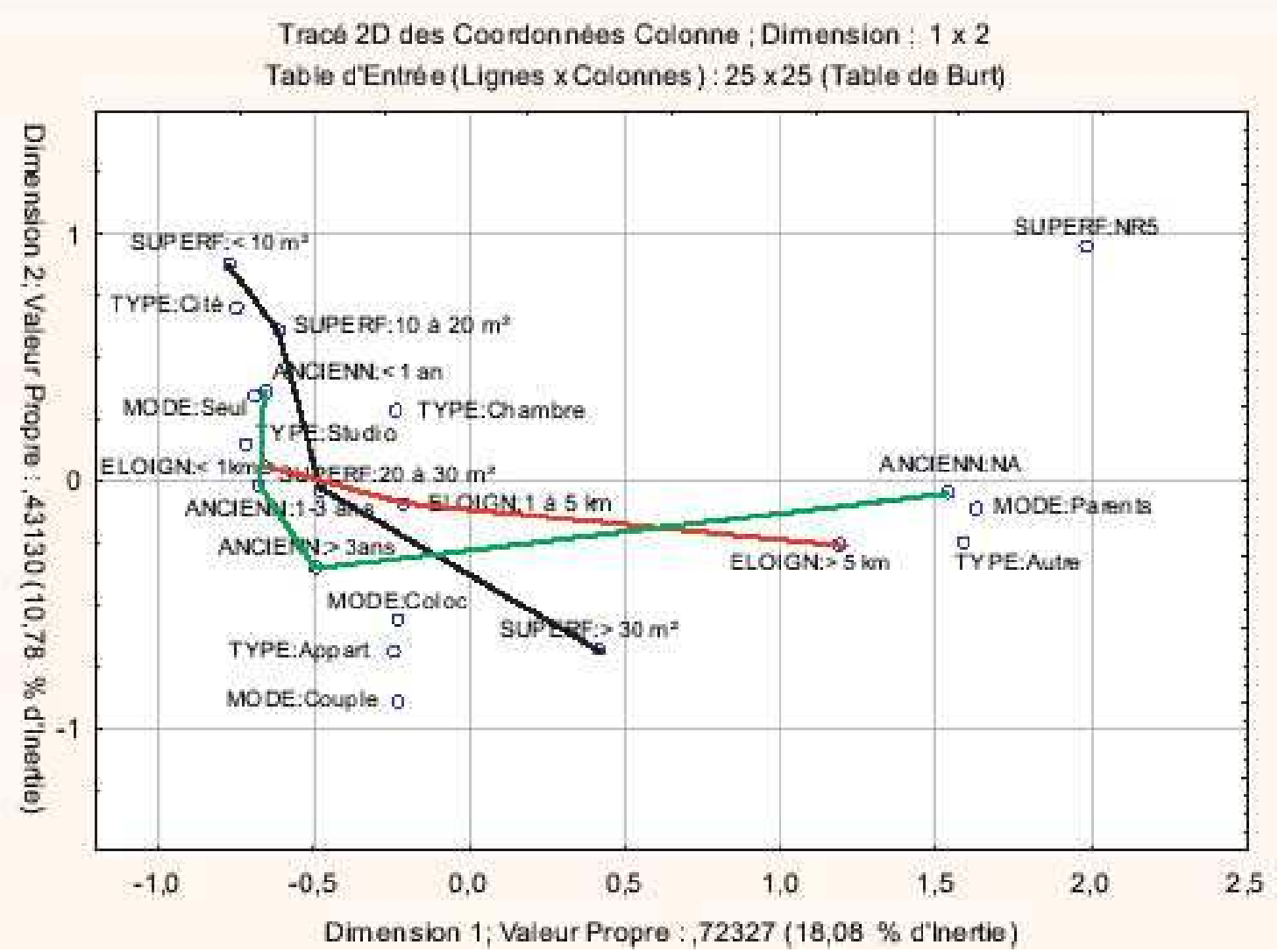
ACM – les modalités – Application à l'exemple

NomLigne	Coordonnées Colonne et Contributions à l'Inertie (Etudiants-ville.sta)									
	Inertie Totale = 4,0000									
	Ligne Numéro	Coord. Dim.1	Coord. Dim.2	Masse	Qualité	Inertie Relative	Inertie Dim.1	Cosinus2 Dim.1	Inertie Dim.2	Cosinus2 Dim.2
MODE:Seul	1	-0,6890	0,3383	0,0966	**0,5505**	0,0258	**0,0634**	**0,4436**	0,0256	0,1069
MODE:Coloc	2	-0,2283	-0,5691	0,0277	0,0604	0,0431	0,0020	0,0084	0,0208	0,0520
MODE:Couple	3	-0,2247	-0,9016	0,0261	0,1296	0,0435	0,0018	0,0076	**0,0492**	0,1221
MODE:Parents	4	1,6354	-0,1130	0,0470	**0,8255**	0,0383	**0,1738**	**0,8215**	0,0014	0,0039
MODE:NR1	5	0,7226	4,5647	0,0026	**0,2825**	0,0493	0,0019	0,0069	**0,1261**	**0,2756**
TYPE:Cité	6	-0,7458	0,6987	0,0214	0,1252	0,0446	0,0165	0,0667	0,0242	0,0585
TYPE:Studio	7	-0,7181	0,1428	0,0564	0,2105	0,0359	**0,0402**	0,2025	0,0027	0,0080
TYPE:Appart	8	-0,2417	-0,6967	0,0606	0,2362	0,0349	0,0049	0,0254	**0,0682**	0,2109
TYPE:Chambre	9	-0,2307	0,2778	0,0104	0,0072	0,0474	0,0008	0,0029	0,0019	0,0043
TYPE:Autre	10	1,5929	-0,2506	0,0397	**0,6437**	0,0401	**0,1392**	**0,6281**	0,0058	0,0156
TYPE:NR2	11	0,8965	2,2834	0,0115	**0,3667**	0,0471	0,0128	0,0490	**0,1389**	**0,3177**
ANCIENN:< 1 an	12	-0,6530	0,3547	0,0418	0,1458	0,0396	0,0246	0,1126	0,0122	0,0332
ANCIENN:1-3 ans	13	-0,4761	-0,0368	0,0496	0,0752	0,0376	0,0155	0,0748	0,0002	0,0004
ANCIENN:> 3ans	14	-0,4892	-0,3545	0,0574	0,1471	0,0356	0,0190	0,0964	0,0167	0,0506
ANCIENN:NA	15	1,5386	-0,0532	0,0496	**0,7818**	0,0376	**0,1624**	**0,7809**	0,0003	0,0009
ANCIENN:NR3	16	1,7019	6,3875	0,0016	**0,3450**	0,0496	0,0063	0,0229	**0,1482**	**0,3221**
ELOIGN:< 1km	17	-0,6525	0,0493	0,0538	0,1575	0,0366	0,0317	0,1566	0,0003	0,0009
ELOIGN:1 à 5 km	18	-0,2123	-0,0983	0,0997	0,0545	0,0251	0,0062	0,0448	0,0022	0,0096
ELOIGN:> 5 km	19	1,1946	-0,2628	0,0418	**0,3950**	0,0396	**0,0824**	**0,3768**	0,0067	0,0182
ELOIGN:NR4	20	1,3552	3,8589	0,0047	**0,4025**	0,0488	0,0119	0,0442	**0,1623**	**0,3583**
SUPERF:< 10 m2	21	-0,7705	0,8750	0,0183	0,1367	0,0454	0,0150	0,0597	0,0324	0,0770
SUPERF:10 à 20 m2	22	-0,6083	0,6080	0,0355	0,1597	0,0411	0,0182	0,0799	0,0304	0,0798
SUPERF:20 à 30 m2	23	-0,6704	-0,0253	0,0496	0,1485	0,0376	0,0308	0,1483	0,0001	0,0002
SUPERF:> 30 m2	24	0,4160	-0,6849	0,0783	**0,4134**	0,0304	0,0187	0,1114	**0,0852**	**0,3019**
SUPERF:NR5	25	1,9890	0,9474	0,0183	**0,4881**	0,0454	**0,1000**	**0,3979**	0,0380	0,0903

On retrouve l'inertie relative de chaque variable comme somme des inerties relatives des modalités qui la compose

$$\frac{I(X_1)}{I} = \frac{0.8}{4} = 0.2 = 0.0258 + 0.0431 + 0.0435 + 0.0383 + 0.0493$$

ACM – les modalités – Application à l'exemple



1er axe : oppositions 'avec les parents' 'plus de 5km' 'autre' vs 'seul' 'moins de 1 km'

2ème axe : opposition 'cité universitaire' 'moins de 10m²' vs 'appartement' 'en couple'

ACM – Proximités entre les modalités

- Si deux modalités d'une même variable sont proches, cela signifie que les individus qui possèdent l'une des modalités et ceux qui possèdent l'autre sont globalement similaires du point de vue des autres variables

- Si deux modalités de deux variables différentes sont proches, cela peut signifier que ce sont globalement les mêmes individus qui possèdent l'une et l'autre

Remarque : Pour l'ACM, il est quasiment **indispensable de regrouper les modalités dont la fréquence est trop faible** (inférieure à 5% par exemple) avec d'autres modalités

Comparaison ACM /AFC – Les points communs

Cas p=2	Les coordonnées des modalités sont les mêmes pour les deux analyses
Représentation	Toutes les modalités peuvent être représentées sur le même diagramme
Contribution d'une modalité à un axe	$\text{poids} \times \frac{(\text{coordonnee})^2}{\text{valeur\_propre}}$
Qualité de la représentation d'1 modalité par un sous espace	$\cos^2\theta = \frac{\sum_{\text{axes\_sous\_esp}} (\text{coord\_axe})^2}{\sum_{\text{tous\_axes}} (\text{coord\_axe})^2}$

Comparaison ACM - AFC– Les points communs

	AFC	ACM
Individus	non	oui
Données	Tableau de contingence profils ligne/colonne	Tableau disjonctif Tableau de BURT
Poids d'une modalité	$\frac{n_{j.}}{n}$ $\frac{n_{.j}}{n}$	$\frac{n_j}{n}$
Nb de valeur propres	$\min(m_1 - 1, m_2 - 1)$	$\sum_{i=1}^{p} m_i - p$
Axes à conserver	Pas de régle Kaiser Part d'inertie	$\mu > \frac{1}{p}$
Variables supplémentaires	Pas vraiment de sens	Qualitatives et quantitatives

ACM – logiciels

SAS

- proc CORRESP, option mca

R

- package MASS : mva
- package ade4 : dudi.acm
- package FactoMineR (cf. http://factominer.free.fr) : MCA

Les résultats les plus complets semblent être ceux fournis par la procédure MCA

Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. Exemple
3. Formalisation mathématique de l'AFC
4. Exemple

5. Analyse des correspondances multiples
6. **Classification**
7. Synthèse

8. Évaluation

Les méthodes de classification – ça sert à quoi?

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère.

Les différentes techniques visent toutes à répartir n individus caractérisés par un ensemble de variables en un certain nombre de sous-groupes aussi homogènes que possible.

Ces méthodes viennent compléter les méthodes factorielles qui fournissent une interprétation de ces derniers.

Les applications

- Marketing : découper le marché en groupe d'individus ou de ménages ayant un comportement homogène en terme de consommation de différents produits, marques, …
- Finance
- Fiabilité : regrouper des populations de matériels
- Biostatistique : identifier différents types de population face à la réaction à un médicament, …

Classification – les 2 grandes familles de méthodes

Les 2 grandes familles de méthodes de classification :

Hiérarchique : aboutit à un arbre de classification (dendrogramme) par agrégations successives d'objets

- Classification ascendante hiérarchique
- Classification descendante hiérarchique (procédé inverse)

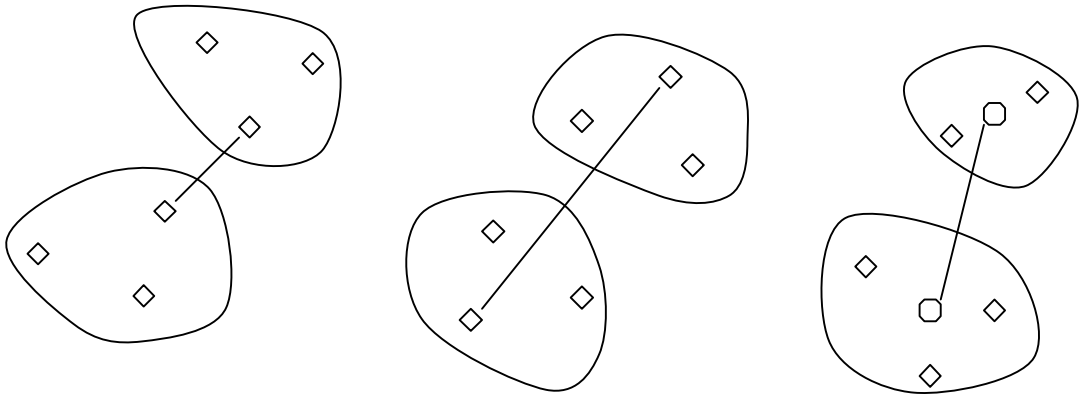
Non hiérarchique : aboutit à un nombre de classes choisi à l'avance

- Agrégation autour de centre mobiles
- Nuées dynamiques

Classification - Principe générale des méthodes de classification hiérarchique

Démarche

- n éléments à classer.
- cherche les deux éléments les plus proches au sens de la distance d de notre espace E
- Nécessité de choisir une distance d



Ce choix de d (qu'on appelle encore stratégie d'agrégation) va différencier les méthodes de classification

Classification – choix d'un indice de dissimilarité ou distance entre individus

Choix d'une distance entre individus

-Distance euclidienne $$d(I_i, I_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$

-Distance euclidienne au carré $$d(I_i, I_j) = \sum_k (x_{ik} - x_{jk})^2$$

-Distance du City-Block $$d(I_i, I_j) = \sum_k \left| x_{ik} - x_{jk} \right|$$

-Distance de Tchebychev $$d(I_i, I_j) = \max \left| x_{ik} - x_{jk} \right|$$

-Distance à la puissance $$d(I_i, I_j) = \left(\sum_k \left| x_{ik} - x_{jk} \right|^p \right)^{\frac{1}{p}}$$

Classification – choix d'un indice d'agrégation

Choix de la distance entre classes (Ex 2 classes, A et B)

-Saut minimum ou "single linkage" (distance minimum)

$$D(A,B) = \min_{I \in A} \min_{J \in B} d(I,J)$$

-Diamètre ou "complete linkage" (distance maximum)

$$D(A,B) = \max_{I \in A} \max_{J \in B} d(I,J)$$

-Moyenne non pondérée des groupes associés

$$D(A,B) = \frac{1}{n_A n_B} \sum_{I \in A, J \in B} d(I,J)$$

-Centroïde non pondéré des groupes associés

- *Le centroïde d'une classe est le point moyen d'un espace multidimensionnel, défini par les dimensions. Dans cette méthode, la distance entre deux classes est déterminée par la distance entre les centroïdes respectifs*

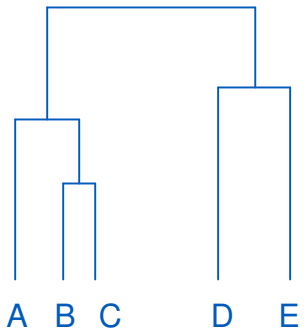
-Méthode de Ward (décrite après)

Classification ascendante hiérarchique (CAH)

Algorithme :

- A la première étape, on regroupe les deux éléments les plus proches (choix d'une métrique).
- On itère le procédé, en regroupant à une étape donnée les deux groupes les plus proches, leur distance étant le niveau de cette agrégation (choix d'un critère d'agrégation)
- L'algorithme est terminé quand on est parvenu à une seule classe qui regroupe la totalité des objets.

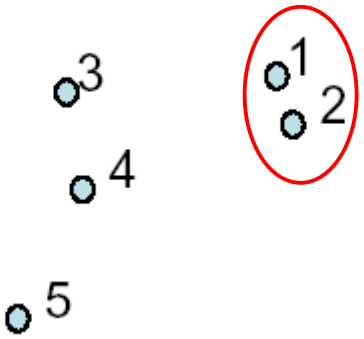
La suite de partitions "emboîtées" résultant d'une CAH est une **hiérarchie de parties**, il est commode de la représenter par un **arbre indicé (dendrogramme)**, chaque **noeud** portant l'indication de son niveau d'agrégation.

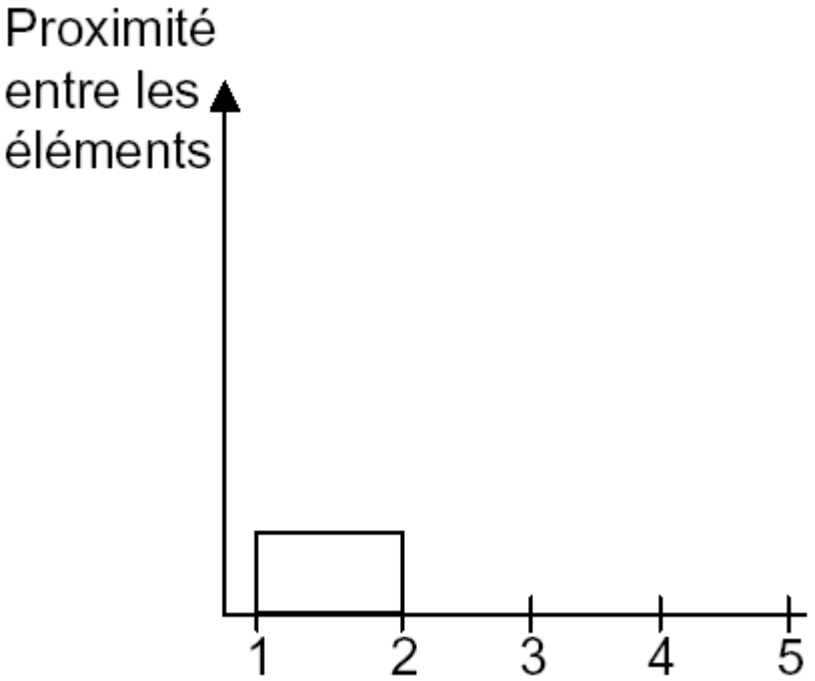


Si on souhaite une partition en un nombre donné de classes, il suffit de tronquer l'arbre au niveau convenable.

Classification

Début : n individus / n classes

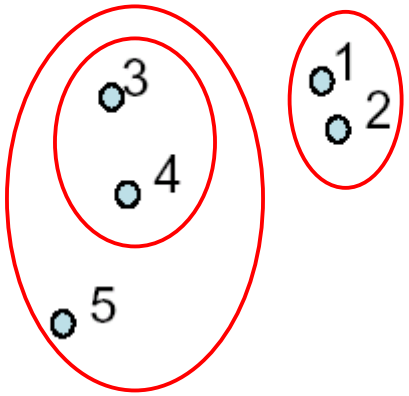


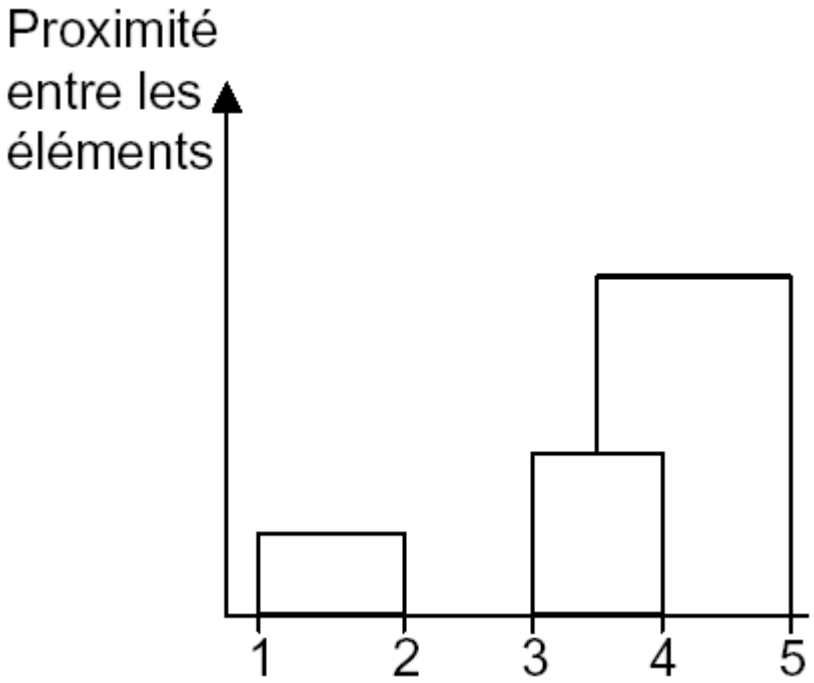


1) Choix d'une métrique entre les classes : distance euclidienne
2) On construit la matrice des distances et on regroupe les 2 éléments les plus proches
3) n-1 classes, comment mesurer la distance entre une classe et un individu ? On utilise le critère d'agrégation (ici distance minimum)

Classification

Suite de la démarche

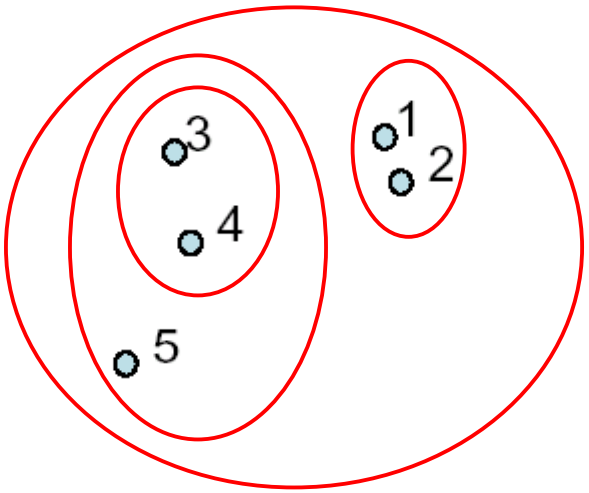




4) On utilise toujours le même critère pour mesurer la distance entre les individus et les classes déjà formées

Classification

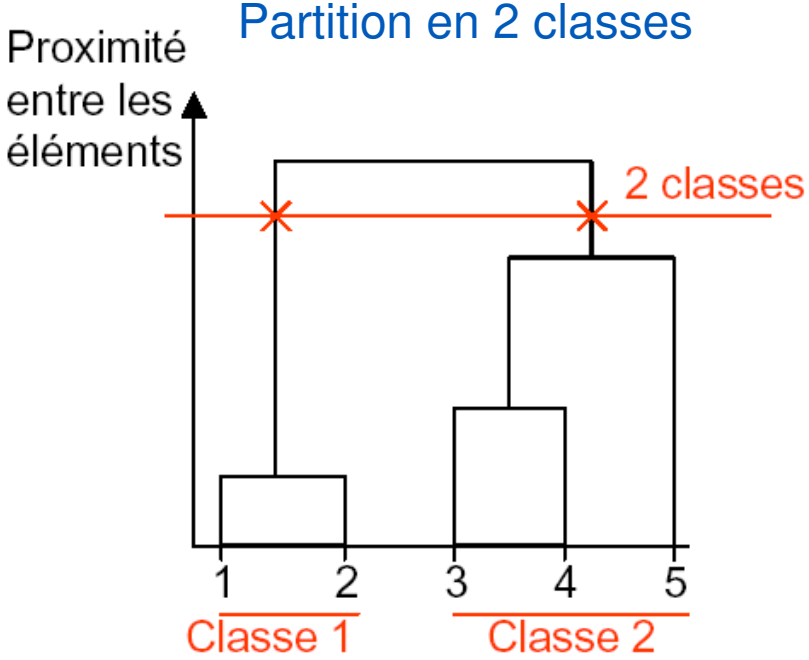
Fin de la démarche : obtention d'1 classe



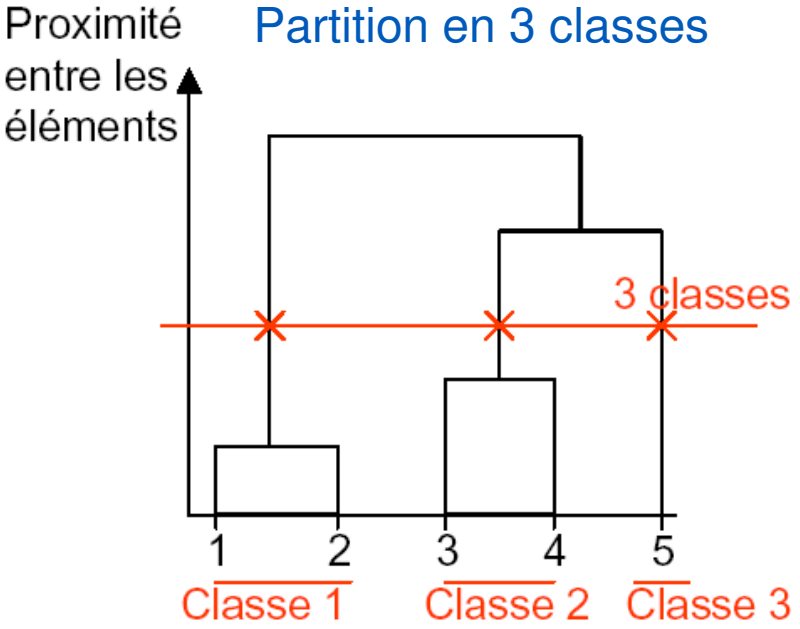
L'historique de la classification est représentée sur l'arbre (aussi appelé dendrogramme)

Le choix du nombre de classes est déterminé a posteriori.

Partition en 2 classes



Partition en 3 classes



EDF

Classification

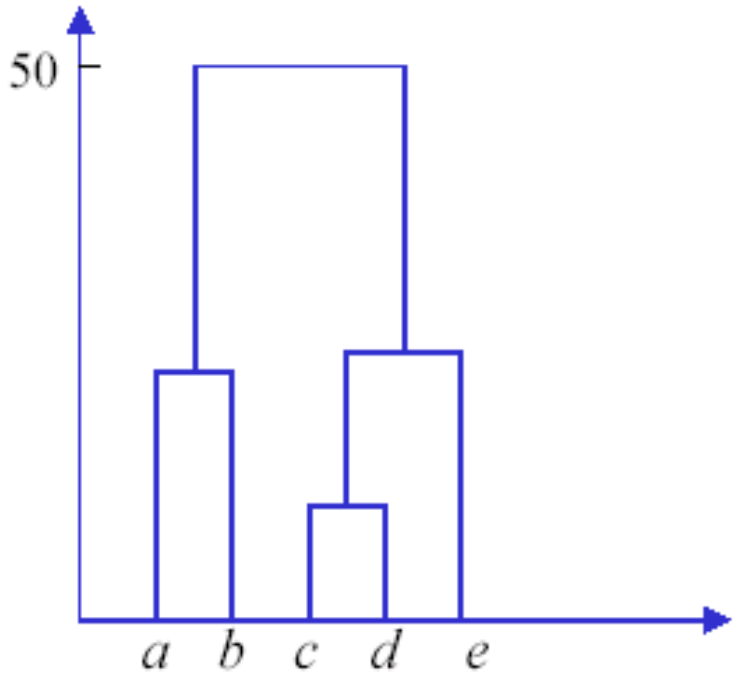
Exemple en utilisant une agrégation selon la distance maximum

	a	b	c	d	e
a	0	23	35	43	50
b	23	0	21	32	45
c	35	21	0	11	25
d	43	32	11	0	17
e	50	45	25	17	0

	a	b	e	cd
a	0	23	50	43
b	23	0	45	32
e	50	45	0	25
cd	43	32	25	0

	e	cd	ab
e	0	25	50
cd	25	0	43
ab	50	43	0

	ab	cde
ab	0	50
cde	50	0



Classification dans un espace euclidien – Inerties interclasse et intraclasse

On considère :

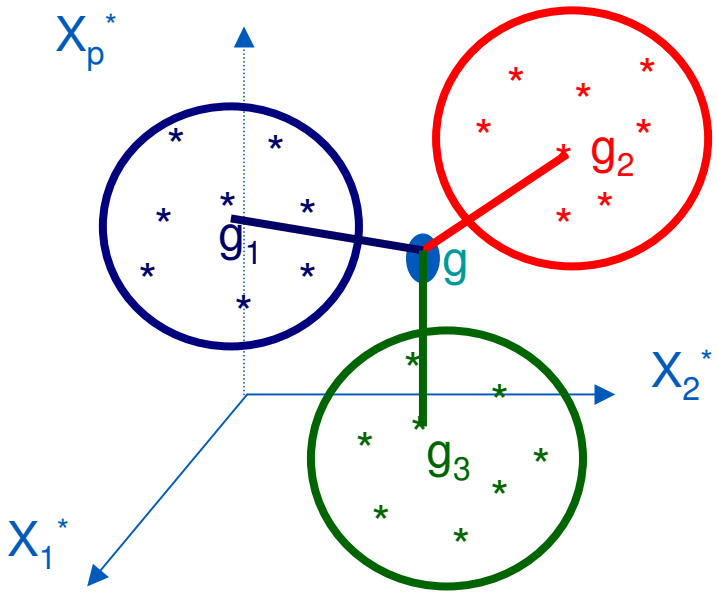
une classification en k groupes $G_1,\ldots,G_k$ d'effectifs $n_1, \ldots, n_k$
les individus sont des points e_i (i =1…n) d'un espace euclidien
on note $g_1,\ldots g_k$ les centres de gravité des groupes
on note g le centre de gravité du nuage de points

- Inertie totale : $$I_{tot} = \frac{1}{n}\sum_{i=1}^{n} d^2(e_i, g)$$

- Inertie interclasse : $$I_{\text{interclasse}} = \frac{1}{n}\sum_{i=1}^{k} n_i d^2(g_i, g)$$

- Inertie intraclasse : $$I_{\text{intraclasse}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{e\in G_i} d^2(e, g_i)$$

Classification – Théorème de Huygens



Au fur et à mesure que les regroupements sont effectués, l'inertie intra-classe augmente et l'inertie interclasse diminue, car leur somme est une constante liée aux données analysées

$$\sum_{i=1}^{n} d^2(x_i^*, g) = \sum_{k=1}^{K} n_k d^2(g_k, g) + \sum_{k=1}^{K} \sum_{i \in G_k} d^2(x_i^*, g_k)$$

Somme des carrés totale = (n-1)*p

=

Somme des carrés inter-classes

+

Somme des carrés intra-classes

Classification – Méthode de Ward

Celle-ci correspond à une **stratégie d'agrégation selon la variance**

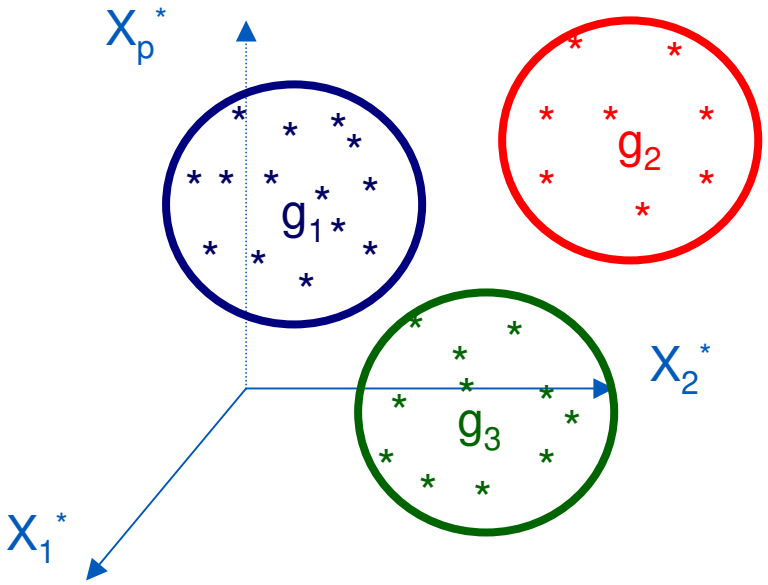
Le principe général de cette méthode est de **rechercher à chaque étape les deux éléments dont l'agrégation entraîne la plus faible perte d'inertie inter ou la plus faible augmentation d'inertie intra**

- Au départ, car chaque classe est réduite à un élément
- La croissance de l'inertie intra au cours du processus indique l'hétérogénéité croissante des classes

Classification – Méthode de Ward

La Méthode de Ward

Elle consiste à choisir à chaque étape le regroupement de classes tel que l'augmentation de l'inertie intraclasse, utilisée comme indice de niveau, soit minimum



Distance de Ward : $D(G_i, G_j) = \dfrac{n_i n_j}{(n_i + n_j)} d^2(g_i, g_j)$

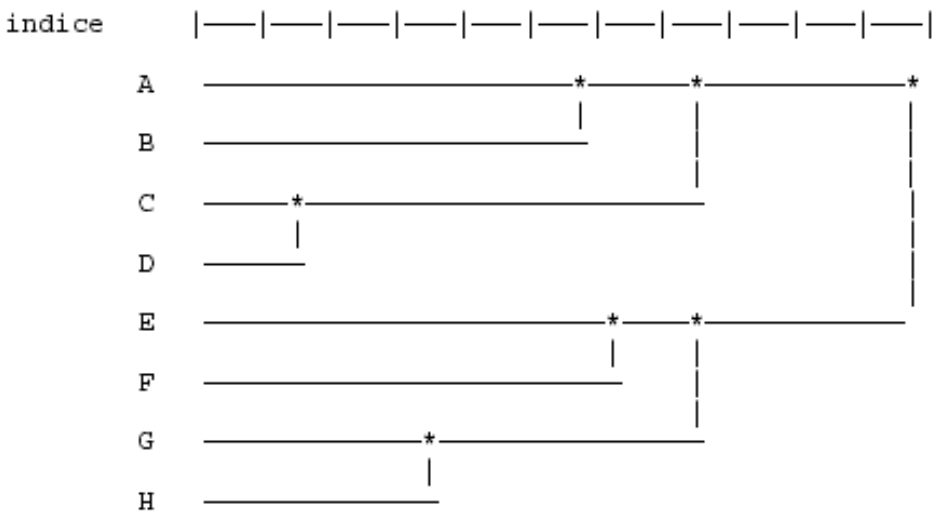
n_i = effectif de la classe G_i

Classification – Arbre de classification

L'arbre de classification ou dendrogramme nous indique l'ordre dans lequel les agrégations successives ont été opérées, ainsi que **la valeur de l'indice d'agrégation** (la longueur des branches de l'arbre représente l'augmentation de l'inertie inter-classes résultant de l'agrégation de 2 classes)

Une typologie est d'autant meilleure que les classes sont ramassées (faible variance intraclasse) et qu'elles sont nettement séparées (forte variance interclasse)

Il est pertinent de **couper l'arbre au niveau d'un saut important de cet indice** → partition de bonne qualité car **les individus regroupés en dessous de la coupure sont proches et ceux regroupés après la coupure sont éloignés**.



Aide à l'interprétation d'une partition

- **Caractérisation de la classe par les individus**
  - Son effectif
  - Son diamètre
  - La séparation : distance entre la classe considérée et la classe la plus proche
  - Les identités des individus les plus proches du centre de gravité de la classe ou « parangons »
  - Les identités des individus les plus éloignés du centre de gravité
- **Caractérisation de la classe par les valeurs continues :**
  - On compare la moyenne et l'écart-type d'une variable X dans la classe k à la moyenne générale et à l'écart-type général
- **Caractérisation de la classe par les valeur nominales :**
  - Pourcentage dans la classe vs global

	Classe k	Autres classes	Population
Modalité j	n_{kj}	*	n_j
Autres modalités	*	*	*
Population	n_k	*	n

Pourcentage global $\Rightarrow$ n_j / n

Pourcentage « mod/clas » $\Rightarrow$ n_{kj} / n_k

Pourcentage « cla / mod » $\Rightarrow$ n_{kj} / n_j

Aide à l'interprétation d'une partition

Valeur-test : permet de sélectionner les variables continues ou les modalités des variables nominales les plus caractéristiques de chaque classe

Variables continues

La valeur-test est égale à l'écart entre la moyenne dans la classe et la moyenne générale exprimée en nb d'écarts-types :

$$v\_test = \frac{\overline{x_k} - \overline{x}}{s_k(X)}$$ où $$s_k^2(X) = \frac{n - n_k}{n-1}\frac{s^2(X)}{n_k}$$

Variables nominales

$$v\_test = \frac{n_{jk} - n_k \frac{n_j}{n}}{\sqrt{n_k \frac{n - n_k}{n-1}\frac{n_j}{n}\left(1 - \frac{n_j}{n}\right)}}$$

Idée : comparer, pour chaque variable, la moyenne observée dans la classe à celle que l'on obtiendrait en effectuant un tirage au hasard d'éléments parmi n.

Classification – Méthode des centres mobiles

Méthode bien adaptée aux très grands tableaux de données

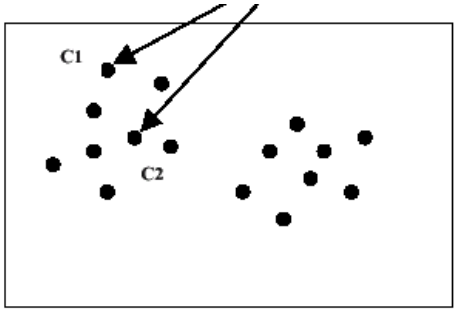
Démarche :

- On choisi une métrique pour calculer la distance entre individus
- On définit un nombre de classes k
- On choisi de façon arbitraire k centres de classes -> le plus souvent k individus tirés au hasard
- Les individus sont affectés au centre de classe le plus proche

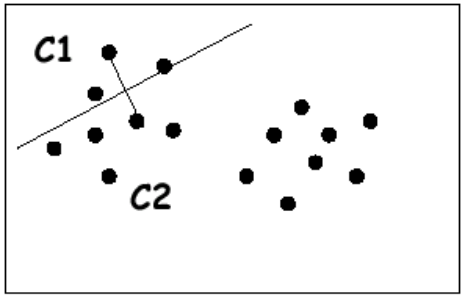
Classification - Méthode des centres mobiles

Les différentes étapes – Cas d'un regroupement en 2 classes – k=2

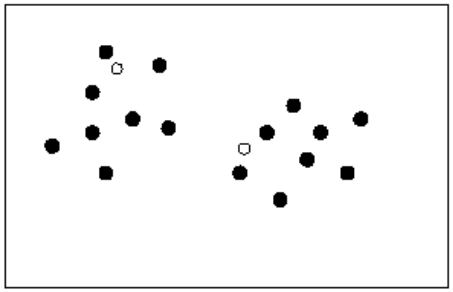
1) **Choix au hasard de 2 centres de classe**



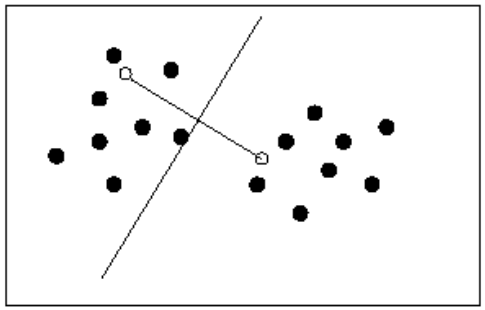
2) **Constitution des 2 premières classes**



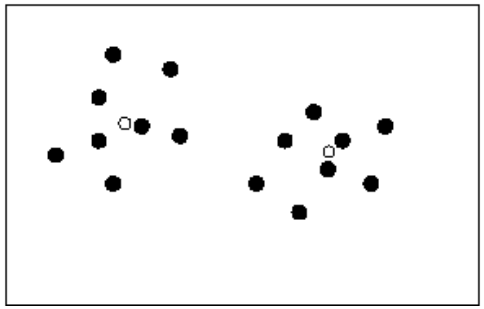
3) **Calcul des nouveaux centres de classes**



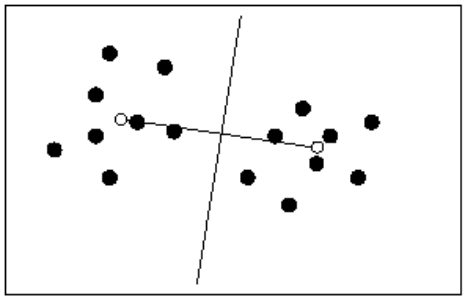
4) **Affection des individus aux nouveaux centres de classes**



5) **Calcul des nouveaux centres de classes**



6) **Affection des individus aux nouveaux centres de classes**



L'algorithme s'arrête lorsque : la variance intra classe cesse de décroître ou la variance interclasse cesse d'augmenter ou l'affectation des individus aux classes ne changent plus ou le nombre d'itérations maximales est atteint

Classification - Méthode des Nuées dynamiques

Généralisation de la méthode des centres mobiles

Le centre de chaque classe n'est plus défini par un point, mais par un noyau d'individus

Si ces noyaux sont bien choisis, ils seront plus représentatifs du centre des classes qu'un unique individu

Conclusion concernant ces 2 méthodes non hiérarchiques

-Si le nombre de classes est mal choisi, on obtiendra une partition…mais pas forcément la meilleure ni la plus facile à interpréter

-Si on n'a aucune idée du nombre de classes, on mettra en œuvre de préférence une méthode hiérarchique

La classification - Synthèse

- Choisir une distance entre individu
- Choisir un indice d'agrégation
- Couper l'arbre pour définir les partitions : On coupe l'arbre hiérarchique de façon à avoir des classes les plus homogènes possibles tout en étant bien séparées entre elles en se référant à l'histogramme des indices de niveau
- Caractériser les partitions en utilisant les valeurs test.

Dans la pratique, on réalise une analyse factorielle et ensuite une classification

cela permet d'éliminer les fluctuations aléatoires et d'obtenir des classes plus stables, les axes factoriels étant très stables relativement à l'échantillonnage

- Réaliser une ACP, AFC ou ACM à partir de nos données et l'interpréter
- Lancer une CAH sur les coordonnées factorielles
- Ensuite c'est la même démarche, on projette les classes sur les axes factoriels

La classification - Synthèse

Quelques remarques

- Plusieurs partitions sont parfois possible, il faut veiller à prendre la plus logique par rapport à l'analyse qui est en train d'être réalisée
- Méthodes plus simples encore dans leurs principes que les méthodes factorielles
- Elles demandent toutefois beaucoup de calculs dès que les données atteignent une certaine taille.

La classification - Exemple

Tableau de données « olympic » du package ade4 de R

Individus : 33 athlètes

Variables : résultats aux épreuves du décathlon (10 variables numériques)

```
> olympic$tab
      100 long  poid haut   400   110  disq perc  jave   1500
1   11.25 7.43 15.48 2.27 48.90 15.13 49.28  4.7 61.32 268.95
2   10.87 7.45 14.97 1.97 47.71 14.46 44.36  5.1 61.76 273.02
3   11.18 7.44 14.20 1.97 48.29 14.81 43.66  5.2 64.16 263.20
4   10.62 7.38 15.02 2.03 49.06 14.72 44.80  4.9 64.04 285.11
5   11.02 7.43 12.92 1.97 47.44 14.40 41.20  5.2 57.46 256.64
6   10.83 7.72 13.58 2.12 48.34 14.18 43.06  4.9 52.18 274.07
7   11.18 7.05 14.12 2.06 49.34 14.39 41.68  5.7 61.60 291.20
8   11.05 6.95 15.34 2.00 48.21 14.36 41.32  4.8 63.00 265.86
9   11.15 7.12 14.52 2.03 49.15 14.66 42.36  4.9 66.46 269.62
10 11.23 7.28 15.25 1.97 48.60 14.76 48.02  5.2 59.48 292.24
11 10.94 7.45 15.34 1.97 49.94 14.25 41.86  4.8 66.64 295.89
12 11.18 7.34 14.48 1.94 49.02 15.11 42.76  4.7 65.84 256.74
13 11.02 7.29 12.92 2.06 48.23 14.94 39.54  5.0 56.80 257.85
14 10.99 7.37 13.61 1.97 47.83 14.70 43.88  4.3 66.54 268.97
15 11.03 7.45 14.20 1.97 48.94 15.44 41.66  4.7 64.00 267.48
```

La classification - Exemple

Construction des données centrées réduites

```
> olympic_cr<-scale(olympic$tab,center=TRUE,scale=TRUE)
> olympic_cr
            100        long         poid       haut          400         110
1   0.22043458  0.97478654  1.128864125  3.0566196 -0.352136754  0.16025591
2  -1.34128835  1.04050249  0.745978530 -0.1354199 -1.464639598 -1.16185535
3  -0.06725123  1.00764452  0.167896357 -0.1354199 -0.922411321 -0.47120021
4  -2.36873765  0.81049668  0.783516334  0.5029880 -0.202556540 -0.64879725
5  -0.72481878  0.97478654 -0.793071411 -0.1354199 -1.717056209 -1.28025337
6  -1.50568024  1.92766777 -0.297572406  1.4605999 -0.875667504 -1.71437946
7  -0.06725123 -0.27381644  0.107835872  0.8221920  0.059208835 -1.29998638
8  -0.60152486 -0.60239618  1.023758276  0.1837841 -0.997201428 -1.35918539
9  -0.19054514 -0.04381063  0.408138299  0.5029880 -0.118417670 -0.76719527
10  0.13823863  0.48191694  0.956190230 -0.1354199 -0.632599656 -0.56986523
11 -1.05360255  1.04050249  1.023758276 -0.1354199  0.620134638 -1.57624843
12 -0.06725123  0.67906478  0.378108057 -0.4546238 -0.239951594  0.12078990
13 -0.72481878  0.51477492 -0.793071411  0.8221920 -0.978503901 -0.21467116
14 -0.84811269  0.77763870 -0.275049723 -0.1354199 -1.352454437 -0.68826326
15 -0.68372080  1.04050249  0.167896357 -0.1354199 -0.314741701  0.77197903
16 -0.43713297 -0.17524252  0.400630739  0.5029880  0.573390821 -0.53039923
17  1.08349199 -1.25955565  1.571810206  0.1837841  1.872868932  1.99542527
```

La classification - Exemple

Calcul du tableau des distances

```
> tab_dist<-dist(olympic_cr,method="euclidean",diag=FALSE,upper=FALSE)
> tab_dist
          1        2        3        4        5        6        7        8
2  4.363550
3  4.108952 1.887325
4  4.183513 2.168189 3.185110
5  5.193806 2.385945 2.190279 3.979429
6  4.280036 2.937114 3.666441 3.346085 2.968765
7  5.074714 3.539372 3.339522 3.787598 4.012323 4.347930
8  4.355115 2.341259 2.532821 3.066357 3.055066 3.909132 3.676821
9  3.722694 2.535716 1.908689 2.747200 3.237166 3.923577 3.100064 1.650532
10 4.055957 2.607391 2.766697 3.088161 4.002633 4.049507 2.925362 3.376789
11 4.968535 3.086611 3.592354 2.312123 4.634123 4.148434 3.612632 3.311037
12 4.198315 2.953623 1.918693 3.474213 3.297635 4.435820 4.550511 2.543238
13 4.395999 2.982101 2.473135 3.738605 1.840845 2.936181 3.944493 2.912951
14 4.416138 2.859251 2.950102 3.170025 3.480274 3.977034 5.145748 2.739774
15 4.093557 2.849020 2.217997 2.906354 3.377025 4.010082 4.428299 2.948360
16 3.706542 2.937969 2.612385 2.946285 3.301819 3.385726 3.237312 2.314765
17 5.740263 6.727530 6.165516 5.901689 8.018739 8.104503 6.062115 6.291773
18 4.478967 4.346522 3.875494 4.550456 5.396229 5.663395 4.133221 3.795724
19 5.225693 3.667333 3.609871 4.372737 3.028644 3.851285 5.083747 3.060084
20 5.697279 3.598784 3.692932 3.453045 4.798658 5.452768 4.789771 3.726547
21 5.112275 4.848836 3.551860 4.988473 4.878775 5.795367 5.174627 4.192207
22 4.601990 4.819984 3.666543 4.873570 4.572135 5.192766 4.426400 3.643515
23 3.854752 4.326088 3.835488 4.257103 4.683658 4.525109 4.776138 3.598080
24 4.273328 5.026616 4.462374 4.729985 4.806756 4.508894 4.472678 4.123669
```

La classification - Exemple

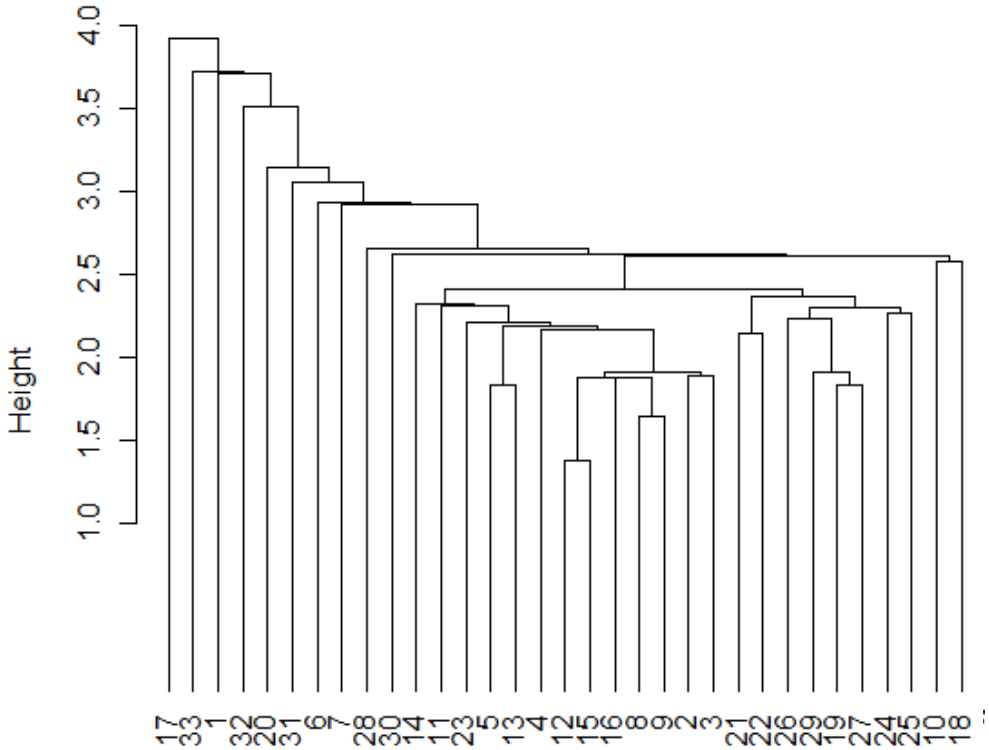
Choix du critère d'agrégation pour la classification

Par ex fonction hclust de R : plusieurs choix possibles : "ward", "single", "complete", "average", "mcquitty", "median" or "centroid".

1er essai : "single" (dist. minimale)

```
> classif_hier<-hclust(tab_dist,"single
> plot(classif_hier,hang=-1)
```

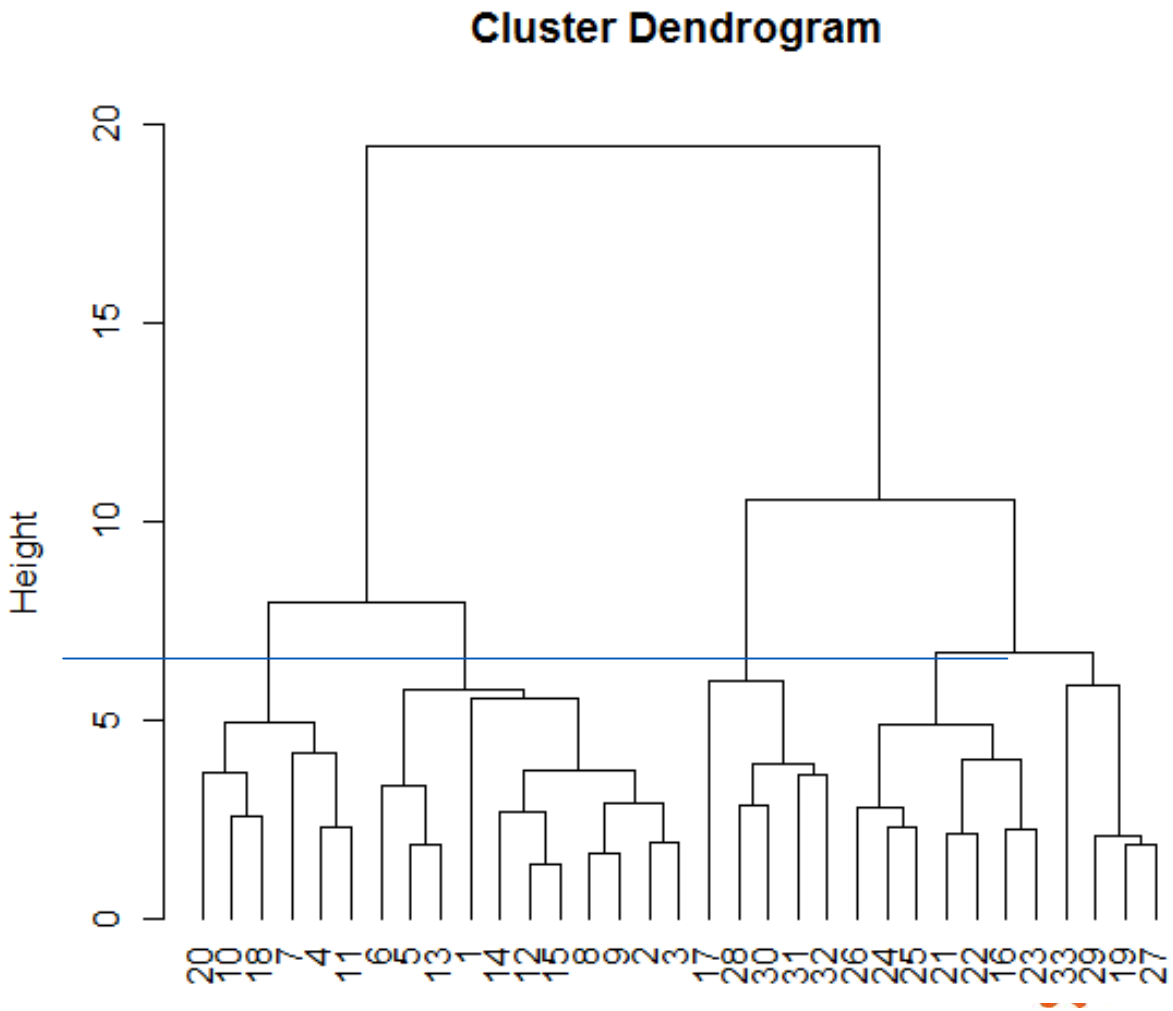
Pas très lisible !



La classification - Exemple

2ème essai : “ward”

```
> classif_hier<-hclust(tab_di
> plot(classif hier,hang=-1)
```



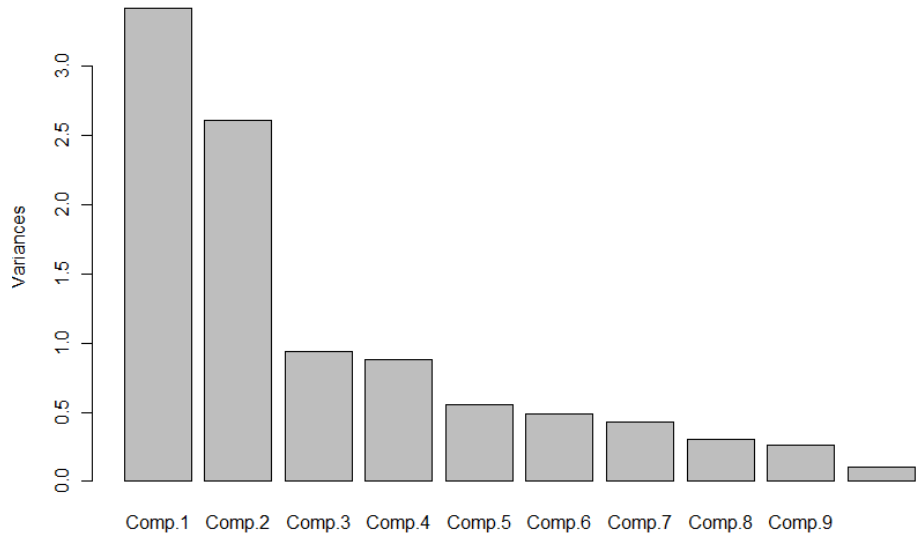
La classification - Exemple

Choix du nombre de classes : par rapport au dendrogramme, ici 4

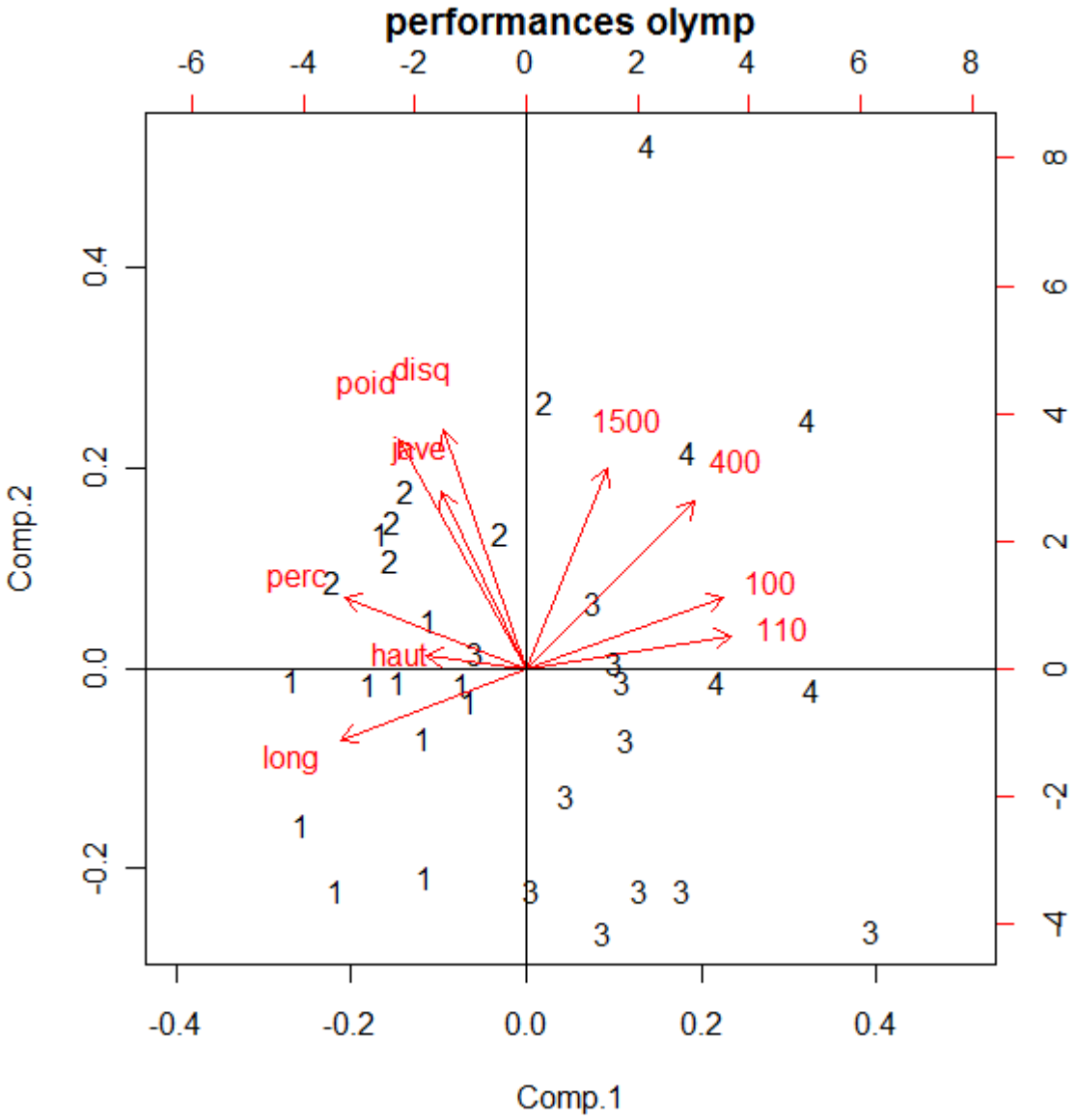
```
> cl4<-cutree(classif_hier,4)
> cl4
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  1  1  2  1  1  2  1  1  2  2  1  1  1  1  3  4  2  3  2  3  3  3  3  3  3
27 28 29 30 31 32 33
 3  4  3  4  4  4  3
```

Représentation des données sur un plan d'ACP

```
> z<-princomp(as.matrix(olympic_cr),cor=TRUE)
> plot(z,main="les valeurs propres")
```



La classification - Exemple



Visualisation des classes sur les 2 premiers axes

La classification - Exemple

Caractérisation des moyennes des classes

```
> olympic_classe<-merge(olympic$tab,cl4,by="row.names")
> moyennes<-aggregate(olympic_classe[,-1],list(olympic_classe$y),mean)
> moyennes[2:11]<-round(moyennes[2:11],2)
> moyennes[,-12]
  Group.1   100 long  poid haut   400   110  disq perc  jave   1500
1       1 11.05 7.36 14.20 2.02 48.37 14.74 43.01 4.86 61.77 265.67
2       2 11.07 7.21 15.36 1.96 49.41 14.75 44.72 5.07 62.07 288.05
3       3 11.28 7.06 13.03 1.98 49.49 15.29 39.32 4.55 55.64 270.04
4       4 11.47 6.70 13.90 1.92 50.65 15.54 44.74 4.50 59.50 297.63
```

Classe 1 : + fort en saut en longueur et hauteur

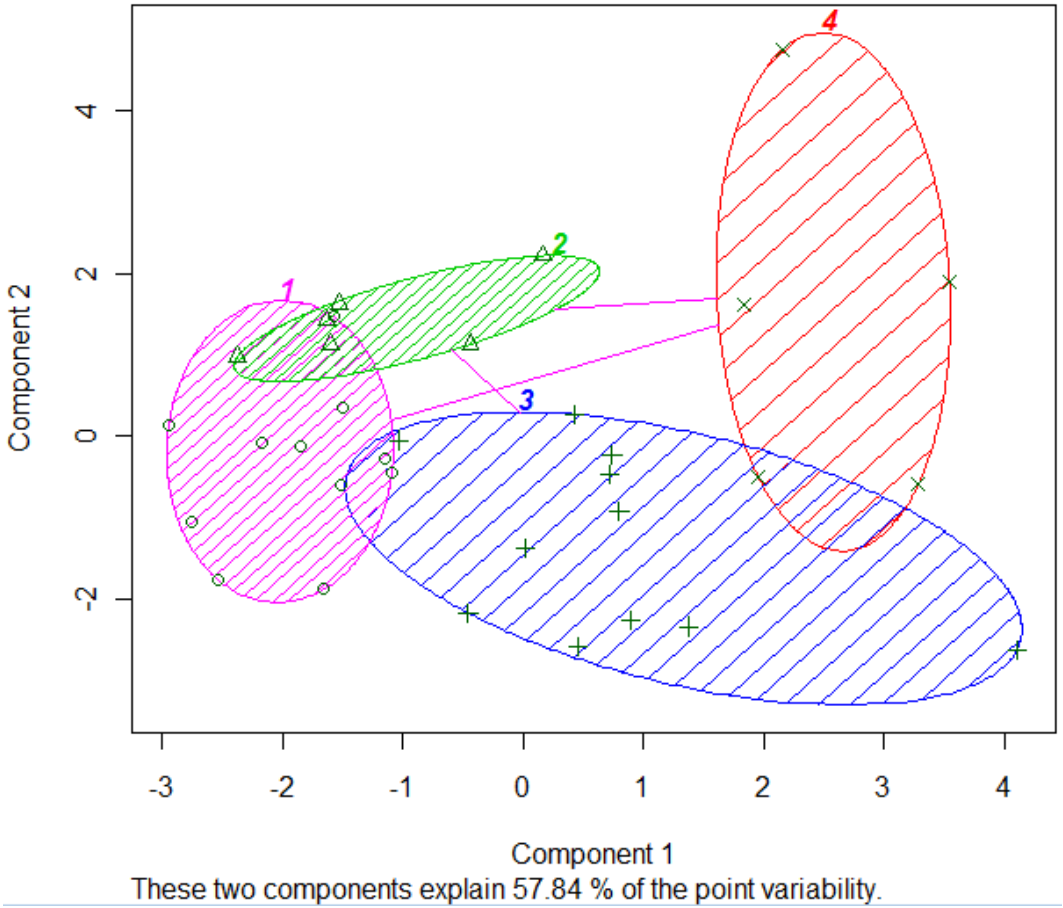
Classe 2 : + javelot, poids, assez fort 1500 m

Classe 3 : Assez fort 100, 110m, 400, en dessous dans les autres épreuves

Classe 4 : Meilleur 100, 110, 400 et 1500m

La classification - Exemple

Représentation des classes sur le plan de l'ACP à l'aide de la fonction « clusplot » du package cluster de R.



Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. Exemple
3. Formalisation mathématique de l'AFC
4. Exemple

5. Analyse des correspondances multiples
6. Classification
7. **Synthèse**

8. Évaluation

Les méthodes d'analyse factorielles

Visualiser, résumer l'information contenue dans des masses volumineuses de données

ACP	Variables quantitatives et les relations sont linéaires
AFC	Tableau de données croisant 2 variables qualitatives (tableau de contingence)
ACM	Lorsque les données sont un mélange de variables qualitatives et quantitatives contenant des relations non linéaires ou présentant des effets de seuil

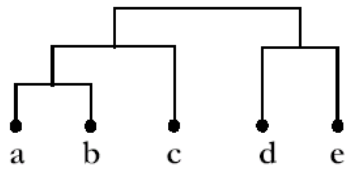
Les méthodes de classification

Opérer des regroupements en classes homogènes d'un ensemble d'individus.

Les données se présentent en général sous la forme d'un tableau individus × variables.

- Ayant défini un critère de distance ou dissimilarité entre les individus, on procède au regroupement des individus.
- Ce regroupement nécessite une stratégie de classification : critère de classification.

Méthodes hiérarchiques



OU

a, b, c, d, e
ab, c, d, e
abc, de
abcde

- Suite de partition emboitées
- Avantages : La lecture de l'arbre permet de déterminer le nombre optimal de classes
- Inconvénients : Coûteux en temps de calcul.

Méthodes non hiérarchiques

- Centres mobiles et Nuées dynamiques
- Avantages : Permettent la classification d'ensembles volumineux.
- Inconvénients : On impose au départ le nombre de classes.

Références

- Littérature :
  - G. Govaert (ed), *Analyse des données*, Lavoisier, 2003
  - G. Saporta, *Probabilités, analyse des données et statistique*, Editions Technip, 2006
  - *Statistique explicative appliquée*, J-P Nakache et Josiane Confais, Editions TECHNIP
  - *Etude de Cas en statistique décisionnelle*, Stephane Tuffery, Editions TECHNIP
  - *Statistique : Méthodes pour décrire, expliquer et prévoir*, M. Tenenhaus, Editions Dunod,

- Présentations et cours sur le web :
  - M. Tenenhaus, Analyse factorielle des correspondances
  - R. Bachelet, L'AFC pour les nuls
  - J. Josse, F. Husson, S. Lê, Correspondance analysis, *useR-2008*
  - La classification automatique, Université Paris Dauphine, M. Gettler-Summa, C. Pardoux
  - Introduction aux analyses multidimensionnelles, F.-G. Carpentier

EDF

Plan du cours

1. Introduction à l'Analyse Factorielle des Correspondances
2. Exemple EDF
3. Formalisation mathématique de l'AFC
4. Exemple et démo en R
5. TD

6. Analyse des correspondances multiples
7. Classification
8. Synthèse

9. Évaluation