

Analyse de données

L. BELLANGER


Master 1 Ingénierie Statistique
Dpt de Mathématiques - Université de Nantes

Plan

- O. Introduction
- I. Outils de représentation d'un échantillon
- II. Analyse en Composantes Principales (ACP)
- III. Analyse Factorielle des Correspondances (AFC)
- IV. Classification et classement
- V. Conclusion

2

Plan ch. III

1. Notions de base
2. Principe de l'AFC
3. Eléments principaux de l'AFC
4. Exemple sous 
5. AFCM

3

Brève introduction ...

- **But de l'AFC**
 - Etudier les tableaux de contingence (croisement de 2 variables qualitatives),
 - ensuite étendue au cas d'un nb qcque de variables qualitatives (analyse factorielle des correspondances multiples (AFCM)).
 - **Principe:**
 - AFC est équivalente à une ACP utilisant une métrique spéciale : la **métrique du Chi-deux**.
 - **Intérêt**
 - Etudier l'ensemble des liaisons entre les modalités des 2 variables
 - Décrire synthétiquement le tableau N à travers des représentations graphiques
 -
- Origine : (Guttman, 1940 - Benzécri, 1964)

4

Brève introduction ...

A l'origine :

Conçue pour étudier des
tableaux de contingence
= tableaux d'effectifs (comptages)
croisant les modalités de 2 variables

- 1940 - Guttman
Fisher
- 1964 - Benzécri

Application en écologie : tableaux espèces x échantillons

- 1971 - Hatheway
- 1973 - Hill

Généralisée à d'autres types de données (condition : valeurs positives)

- 1970-80
Benzécri et coll.

5

Brève introduction ...

Pbs relevant d'une AFC

• Analyse d'un tableau de contingence :

- Répartition des baccalauréats délivrés une année donnée, selon l'académie et la série.

Quels sont les liens existants entre ces 2 variables ?

- Etude de l'efficacité de différentes politiques d'insertion sur des chômeurs de longue durée

Efficacité des politiques d'insertion sur chaque personne ?

- A partir d'un inventaire communal, enregistrement des présences-absences d'un certain nombre d'équipements dans plusieurs communes d'une certaine région. L'ensemble I est celui des équipements, J des communes. Le tableau N est logique (ou disjonctif complet) est composé de 0 (absence de l'équipement dans i dans la commune j).

Description des équipements suivants les communes.

6

Brève introduction ...

Domaine d'application

Tableau de contingence

Analyse les liens entre
variables qualitatives

		Variable 2				
		Modalité 1	Modalité 2	Modalité 3	...	Modalité p
Variable 1	Modalité 1	N_{11}	N_{12}	N_{13}		N_{1p}
	Modalité 2	N_{21}	N_{22}			
	Modalité 3	N_{31}	N_{32}			
			
	Modalité n	N_{n1}	N_{n2}			N_{np}

Exemple : couleur des yeux, profession, classe d'âge, ...

7

Brève introduction ...

Types de tableaux traités par l'AFC

• Tableau de contingence N à I lignes et J colonnes

- n_{ij} représente le nombre d'individus possédant à la fois le caractère i et le caractère j
- Les individus ne sont présents que par leurs effectifs : les lignes et les colonnes jouent le même rôle.

• Tableau logique ou « disjonctif complet »

- Composé de données binaires.
- Le terme situé ligne i et colonne j vaut 1 ou 0 selon que le caractère i est présent ou non dans le caractère j .

• Certains tableaux de mesures

- Si toutes les mesures contenues dans le tableau sont positives, alors celui-ci peut être analysé par une AFC.

8

Brève introduction ...

2 différences entre l'ACP et l'AFC

- La **métrique** utilisée en **AFC** pour définir la proximité entre 2 lignes ou 2 colonnes est la **métrique du Chi-deux** alors que l'on utilise la distance euclidienne en **ACP**.
- L'**AFC** autorise une **représentation superposée des lignes et des colonnes (à utiliser avec précaution)**
 - 2 graphes indépendants en **ACP**!

9

1. Notions de base

- Tableau de contingence** $N_{I \times J} = [n_{ij}] ; i = 1, \dots, I ; j = 1, \dots, J$

	Variable 2				
	Modalité 1	Modalité 2	Modalité j	Modalité J	Effectif marginal
Variable 1	Modalité 1	n_{11}	n_{12}	n_{1j}	n_{1+}
	Modalité 2	n_{21}	n_{22}	n_{2j}	n_{2+}
	\vdots				
	Modalité i	n_{i1}	n_{i2}	n_{ij}	n_{i+}
	\vdots				
	Modalité I	n_{I1}	n_{I2}	n_{IJ}	n_{I+}
	Effectif marginal	n_{+1}	n_{+2}	n_{+j}	$n_{++} = n$

- I : nb de lignes, J : nb de colonnes ;
- n_{ij} : **nb d'indiv.** possédant à la fois la modalité i de la 1ère variable et la modalité j de la 2ème variable ;
- les individus ne sont présents que par leurs effectifs :
 - les lignes et les colonnes jouent un rôle symétrique.
- On verra par la suite que l'on peut adapter l'**ACP** pour l'analyser, en utilisant une métrique spéciale appelée **métrique du χ^2** pour définir la proximité entre 2 lignes ou 2 colonnes.

10

1. Notions de base

- Tableau de contingence**

On appelle :

- Tableau des fréquences associé à N** : $P = N/n_{++}$ où $n_{++} = \sum_{i,j} n_{ij}$, la matrice $I \times J$, d'elt courant $f_{ij} = \frac{n_{ij}}{n_{++}}$.
 ➤ On a : $\sum_{i,j} f_{ij} = 1$.
- Les **effectifs marginaux** (resp. fréquences marginales) du tableau N (resp. P), les vecteurs :
 - colonne dont l'élément courant i est $n_{i+} = \sum_{j=1}^J n_{ij}$ (resp. $f_{i+} = n_{i+}/n_{++}$) aussi appelé **profil-colonne moyen** ; (marge en lignes de N)
 - ligne dont l'élément courant j est $n_{+j} = \sum_{i=1}^I n_{ij}$ (resp. $f_{+j} = n_{+j}/n_{++}$) aussi appelé **profil-ligne moyen** ; (marge en colonnes de N)
- $D_1 = \text{diag}\{n_{1+}, n_{2+}, \dots, n_{I+}\}$ et $D_J = \text{diag}\{n_{+1}, n_{+2}, \dots, n_{+J}\}$, les deux matrices diagonales dont les éléments sont les effectifs marginaux des 2 variables du tableau N.
- $\tilde{D}_1 = \text{diag}\{f_{1+}, f_{2+}, \dots, f_{I+}\}$ et $\tilde{D}_J = \text{diag}\{f_{+1}, f_{+2}, \dots, f_{+J}\}$, les deux matrices diagonales dont les éléments sont les fréquences marginales des 2 variables du tableau P.

1. Notions de base

- Tableau de contingence**

⇒ **Analyse du tableau de contingence** :

- Ce ne sont pas les effectifs bruts qui nous intéressent ; mais les répartitions en % à l'intérieur d'une ligne ou d'une colonne.
- On parle de **profils-lignes** et de **profils-colonnes** (freq. Cond.) :
 - Un profil ligne (resp. colonne) est déterminé en divisant la ligne (colonne) par le total de la ligne (resp. colonne) :

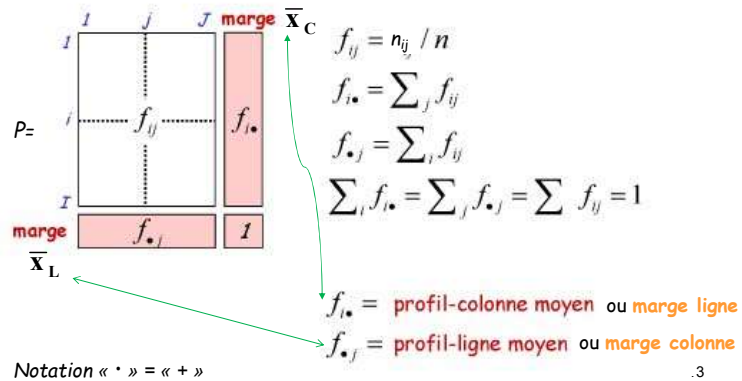
$$\text{profil ligne : } f_{j|i} = \frac{n_{ij}}{n_{i+}} = \frac{f_{ij}}{f_{i+}} \quad \forall j = 1, \dots, J \text{ et } i \text{ fixé}$$

$$\text{profil colonne : } f_{i|j} = \frac{n_{ij}}{n_{+j}} = \frac{f_{ij}}{f_{+j}} \quad \forall i = 1, \dots, I \text{ et } j \text{ fixé}$$

12

1. Notions de bases

Transformation → Tableau des fréquences relatives définit une mesure de probabilité



3

1. Notions de base

Propriétés et écriture matricielle

- M1** : $1 = n^{-1}(\mathbf{1}_I)^T D_I \mathbf{1}_I = n^{-1}(\mathbf{1}_J)^T D_J \mathbf{1}_J = n^{-1}(\mathbf{1}_I)^T P \mathbf{1}_I$ ou encore

$$1 = \sum_{i=1}^I f_{i+} = \sum_{j=1}^J f_{+j} = \sum_{i=1}^I \sum_{j=1}^J f_{ij}$$

- M2** : le **profil-ligne moyen** (ou marge colonne) s'écrit :

$$\bar{x}_L = (\mathbf{1}_I)^T P = \begin{bmatrix} f_{1+} \\ \vdots \\ f_{I+} \end{bmatrix} \quad \text{vecteur ligne}$$

et le **profil-colonne moyen** (ou marge ligne) :

$$\bar{x}_C = P \mathbf{1}_J = \begin{bmatrix} f_{+1} \\ \vdots \\ f_{+J} \end{bmatrix} \quad \text{vecteur colonne}$$

Notation : $n = n_{++}$

14

1. Notions de base

- Indépendance de 2 variables qualitatives**

Il y a **indépendance** entre les deux variables si

$$f_{ij} = f_{i\bullet} f_{\bullet j}$$

$$\Leftrightarrow \begin{cases} \text{toutes les lignes sont proportionnelles} & \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \\ \text{toutes les colonnes sont proportionnelles} & \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet} \end{cases}$$

Il y a **liaison** entre les deux variables lorsque certaines cases f_{ij} diffèrent du produit $f_{i\bullet} f_{\bullet j}$

$f_{ij} > f_{i\bullet} f_{\bullet j}$: modalités i et j s'associent plus qu'elles ne le feraient sous l'hypothèse d'indépendance (H_0)
 i et j s'attirent

$f_{ij} < f_{i\bullet} f_{\bullet j}$: modalités i et j s'associent moins que sous H_0
répulsion entre les deux modalités

5

1. Notions de base

- Le test d'indépendance du Chi2**

⇒ Construction d'une mesure de l'écart à l'indépendance

Notant t_{ij} l'effectif théorique $t_{ij} = \frac{n_{i+} \times n_{+j}}{n_{++}}$, on définit :

$$D_{\chi^2} = \sum_{i,j} \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = n_{++} \sum_{i=1}^I \sum_{j=1}^J \frac{1}{f_{i+} f_{+j}} (f_{ij} - f_{i+} f_{+j})^2$$

Pearson a montré que sous l'hypothèse nulle H_0 « les 2 variables sont indépendantes », D_{χ^2} suit une loi du $\chi^2((I-1)(J-1))$

Pour un risque d'erreur α , le plus souvent 5%, on a la zone de rejet :

$$W_\alpha = \{ D_{\chi^2} > \chi^2_{1-\alpha; (I-1)(J-1)} \}$$

16

1. Notions de base

• Le test d'indépendance du Chi2

L'analyse de l'**écart à l'indépendance** va permettre d'analyser la structure des nuages.

- Notons qu'à cause des relations entre profils ligne et colonne, les dim des espaces de représentation sont au plus $(I - 1)$ et $(J - 1)$ resp.

- Exemple** : Une population de 200 femmes a été interrogée sur le **nombre et le sexe des enfants** qu'elles ont eus.

Soit X la variable nb de garçons et Y la variable nb de filles. Le tableau suivant présente la distribution obtenue en croisant les variables X et Y .

17

1. Notions de base

• Le test d'indépendance du Chi2

Tableau des effectifs observés

X/Y	0	1	2	n_{i+}
0	16	24	20	60
1	22	24	34	80
2	22	32	6	60
n_{+j}	60	80	60	200

- L'hypothèse H_0 est que ces variables sont indépendantes.
- Considérons un risque standard $\alpha = 5\%$.
- Le nombre de degrés de liberté est $(3 - 1) \times (3 - 1) = 4$.
- La valeur critique se lit dans la table du Chi2 (ligne 4 et colonne 0.05) : 9.488.
- Donc, si la distance calculée d est supérieure à 9.488, l'hypothèse H_0 d'indépendance est rejetée.

18

1. Notions de base

• Le test d'indépendance du Chi2

Tableau des effectifs théoriques sous hyp. d'indépendance.

X/Y	0	1	2	n_{i+}
0	18	24	18	60
1	24	32	24	80
2	18	24	18	60
n_{+j}	60	80	60	200

Exemple de calcul : $t_{11} = 18 = (60 \times 60)/200$

19

1. Notions de base

• Le test d'indépendance du Chi2

Tableau des contributions au Chi2 et calcul du Chi2 par sommation

X/Y	0	1	2	Tot
0	0.22	0.00	0.22	0.44
1	0.17	2.00	4.17	6.33
2	0.89	2.67	8.00	11.56
Tot	1.28	4.67	12.39	18.33

i.e. $d_{ij} = (t_{ij} - n_{ij})^2 / t_{ij}$ - Exemple de calcul : $0.22 = (16 - 18)^2 / 18$

- Ce tableau comporte $(3 - 1)(3 - 1) = 4$ degrés de liberté. La table donne pour le seuil de 5 % le nombre 9.488.
- Décision** : le Chi2 calculé s'élevant à $d = 18.33 (> 9.488)$, on rejette donc l'hypothèse d'indépendance.

Remarques :

- Envisager d'effectuer une AFC \Rightarrow supposer \exists une liaison entre les 2 variables étudiées.
- Sous R : `chisq.test (stats)`

A faire ex. Td page 2

20

1. Notions de base

Objectifs

L'AFC cherche à obtenir une typologie des lignes, une typologie des colonnes, et relier ces deux typologies entre elles

Originalité

La notion de ressemblance entre 2 lignes ou entre 2 colonnes est différente de celle de l'ACP :

Les lignes et les colonnes jouent un rôle absolument symétrique

Objectif fondamental

Etudier la liaison entre 2 variables

= étudier la proximité entre chaque profil et son profil moyen

= étudier l'écart du tableau à l'hypothèse d'indépendance

Analyse factorielle : réduire la dimension des données en conservant le plus d'information possible

2. Principe de l'AFC

Analyse d'un tableau de contingence : pondération

On appelle :

- **Profil ligne** i le vecteur ligne des fréquences conditionnelles à $i \in \{1, \dots, I\}$ fixé :

$$L_i = x_i = [n_{i1}/n_{i+}, \dots, n_{ij}/n_{i+}] = [f_{i1}/f_{i+}, \dots, f_{ij}/f_{i+}] ; i = 1, \dots, I$$

- **Tableau $I \times J$ des profils-lignes, le tableau des fréquences conditionnelles** f_{ij}/f_{i+} par $L = D_I^{-1}N \in \mathcal{M}_{I \times J}$:

➤ les profils-lignes forment un nuage de I points dans \mathbb{R}^J ;

➤ chaque profil est affecté d'un poids égal à sa fréquence marginale (**matrice des poids** $D_I/n_{++} = \bar{D}_I$) ;

➤ le **centre de gravité (profil-ligne moyen)** de ce nuage est :

$$\bar{x}_L = \frac{1}{n_{++}} (1_I)^T D_I (D_I^{-1} N) = [f_{+1}, \dots, f_{+J}] \in \mathbb{R}^J$$

- C'est le profil marginal des lignes.

- De manière plus générale, chacun des I profils est représenté dans un espace à $J - 1$ dimensions.

22

2. Principe de l'AFC

Analyse d'un tableau de contingence : pondération

On appelle :

- **Profil colonne** j le vecteur colonne des fréquences conditionnelles à $j \in \{1, \dots, J\}$ fixé

$$C^j = x^j = [n_{1j}/n_{+j}, \dots, n_{ij}/n_{+j}]^T = [f_{1j}/f_{+j}, \dots, f_{ij}/f_{+j}]^T ; j = 1, \dots, J.$$

- **Tableau $I \times J$ des profils-colonnes, le tableau des fréquences conditionnelles** f_{ij}/f_{+j} par $C = N D_J^{-1} \in \mathcal{M}_{I \times J}$:

➤ les profils-colonnes forment un nuage de J points dans \mathbb{R}^I ;

➤ chaque profil est affecté d'un poids égal à sa fréquence marginale (**matrice des poids** $D_J/n_{++} = \bar{D}_J$) ;

➤ le **centre de gravité (profil-colonne moyen)** de ce nuage est :

$$\bar{x}_C = \frac{1}{n_{++}} (N D_J^{-1}) D_J 1_J = [f_{1+}, \dots, f_{I+}]^T \in \mathbb{R}^I.$$

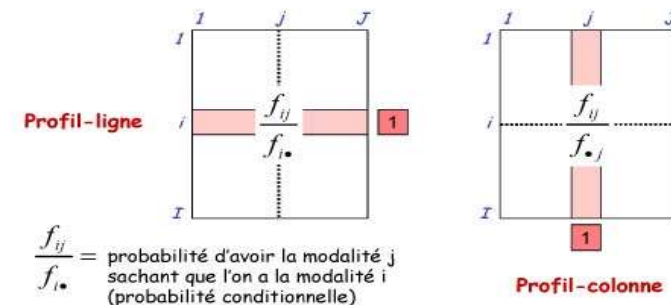
- C'est le profil marginal des colonnes.

23

2. Principe de l'AFC

Principe

Transformations des données en profils



➡ Selon que l'on s'intéresse aux lignes ou aux colonnes, on ne considère pas le même tableau transformé

24

2. Principe de l'AFC

• Modèle d'indépendance et AFC

- L'analyse d'un tableau de contingence s'effectue en référence à la situation d'indépendance ;

- C'est ce que fait l'AFC en écrivant le **modèle d'indépendance** sous la forme :

$$L_i = x_i = [f_{i1}/f_{i+}, \dots, f_{ij}/f_{i+}] = [f_{+1}, \dots, f_{+j}] = \bar{x}_L ; \forall i$$

et

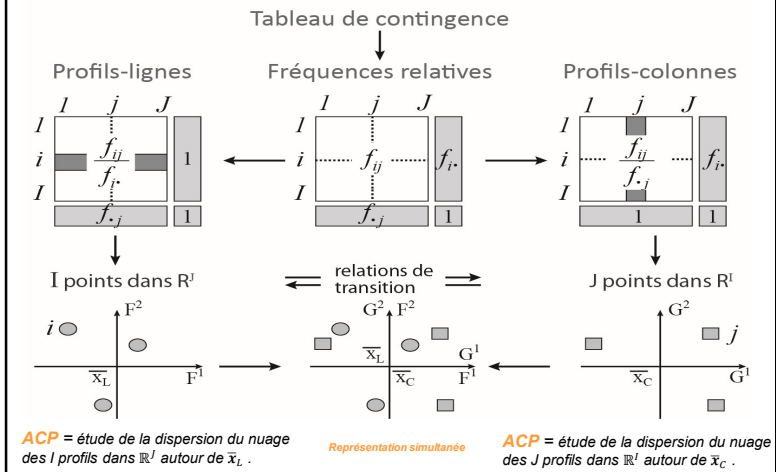
$$C^j = x^j = [f_{1j}/f_{+j}, \dots, f_{ij}/f_{+j}]^T = [f_{1+}, \dots, f_{i+}]^T = \bar{x}_C ; \forall j$$

- Le modèle d'indépendance stipule donc que : les profils-lignes (resp. profils-colonnes) sont égaux au profil-ligne (resp. profil-colonne) moyen.

25

2. Principe de l'AFC

Schéma général



2. Principe de l'AFC

• Analyse d'un tableau de contingence : pondération

- Dans le cas de l'**indépendance statistique** entre les 2 caractères, on a pour tout (i, j) :

$$\left\{ \begin{array}{l} \text{toutes les lignes sont proportionnelles : } \frac{n_{ij}}{n_{i+}} = \frac{n_{+j}}{n_{++}} \\ \text{toutes les colonnes sont proportionnelles : } \frac{n_{ij}}{n_{+j}} = \frac{n_{i+}}{n_{++}} \end{array} \right. \Leftrightarrow f_{ij} = f_{i+}f_{+j}$$

Il y a donc une liaison entre les 2 variables lque certaines cases f_{ij} diffèrent du produit $f_{i+}f_{+j}$.

⇒ Dans le cas de l'indépendance, chaque nuage est alors réduit à un seul point, son point moyen.

L'analyse de l'écart à l'**indépendance** va permettre d'étudier la structure des nuages.

27

2. Principe de l'AFC

• Caractéristiques des nuages

- Nuage des I profils-lignes N_J : $(L = \tilde{D}_I^{-1}P, Q_{J+J} = \tilde{D}_J^{-1}, D_{I+I} = \tilde{D}_I)$

Un profil-ligne est associé à un point dans un e.v. de dim J auquel on attribue la distance suivante entre profils-lignes :

➤ Distance entre 2 profils-lignes :

$$d_{\chi^2}^2(L_i, L_{i'}) = \sum_{j=1}^J \frac{n_{++}}{n_{+j}} \left(\frac{n_{ij}}{n_{i+}} - \frac{n_{i'j}}{n_{i'+}} \right)^2 = \sum_{j=1}^J \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$

- Utiliser cette distance revient à utiliser la **métrique diagonale** $Q = n_{++}D_J^{-1} = \tilde{D}_J^{-1}$ sur les profils-lignes
 - espace engendré par N_J est **e.v. euclidien** ($L = \tilde{D}_I^{-1}P, Q_{J+J} = \tilde{D}_J^{-1}$)
 - La **pondération** par $[1/f_{+j}]$ de chaque carré de différence revient à donner des **importances comparables** aux diverses « variables » considérées.

28

2. Principe de l'AFC

• Caractéristiques des nuages

Nuage des profils-lignes : $(L = \tilde{D}_I^{-1}P, Q_{j \cdot j} = \tilde{D}_J^{-1}, D = \tilde{D}_I)$

➤ **Pondération :** poids de chaque profil-ligne : $n_{i+}/n_{++} = f_{i+}$

La **matrice des poids « D »** est \tilde{D}_I

- On définit le **nuage pondéré** des profils-lignes N_j
Nécessaire pour le calcul du centre de gravité du nuage \bar{X}_I et de l'inertie $I(N_j)$

➤ **En résumé :** en reprenant les notations du chapitre ACP, l'étude des **Profils-lignes** N_j (**I lignes dans \mathbb{R}^J**), on va étudier le triplet :

$$(L = \tilde{D}_I^{-1}P, Q = \tilde{D}_J^{-1}, D = \tilde{D}_I)$$

29

2. Principe de l'AFC

• Caractéristiques des nuages

▪ **Nuage des profils-colonnes N_j :** $(C^T = \tilde{D}_J^{-1}P^T, Q_{i+} = n_{i+}, D_1^{-1} = \tilde{D}_I^{-1}, D_{j \cdot j} = \tilde{D}_J)$

Un profil-colonne est associé à un point dans un e.v. de dim I auquel on attribue la distance suivante entre profils-lignes :

➤ **Distance entre 2 profils-colonnes**

$$d_{\chi^2}^2(C^j, C^{j'}) = \sum_{i=1}^{I} \frac{n_{++}}{n_{i+}} \left(\frac{n_{ij}}{n_{+j}} - \frac{n_{ij'}}{n_{+j'}} \right)^2$$

- Utiliser cette distance revient à utiliser la **métrique diagonale** $Q = n_{++}D_1^{-1} = \tilde{D}_I^{-1}$ sur les profils-lignes
 - espace engendré par N_j est **e.v. euclidien** ($C^T = \tilde{D}_J^{-1}P^T, \tilde{D}_I^{-1}$)
 - La **pondération** par $[1/f_{i+}]$ de chaque carré de différence revient à donner des **importances comparables** aux diverses « variables » considérées.

30

2. Principe de l'AFC

• Caractéristiques des nuages

Nuage des profils-colonnes : $(C^T = \tilde{D}_J^{-1}P^T, Q_{i+} = \tilde{D}_I^{-1}, Q_{j \cdot j} = \tilde{D}_J)$

➤ **Pondération :** poids de chaque profil-colonne : $n_{+j}/n_{++} = f_{+j}$

La **matrice des poids « D »** est donc \tilde{D}_J

- On définit le **nuage pondéré** des profils-colonnes N_j
Nécessaire pour le calcul du centre de gravité du nuage \bar{X}_C et de l'inertie $I(N_j)$

➤ **En résumé :** en reprenant les notations du chapitre ACP, l'étude des **Profils-colonnes** N_j (**J lignes dans \mathbb{R}^I**), revient à étudier le triplet :

$$(C^T = \tilde{D}_J^{-1}P^T, Q = \tilde{D}_I^{-1}, D = \tilde{D}_J)$$

31

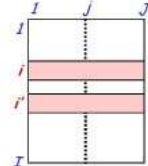
2. Principe de l'AFC

• En résumé :

Principe

Ressemblance entre profils

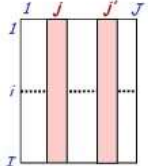
La ressemblance entre deux lignes ou entre deux colonnes est définie par une **distance** entre leurs **profils** : la **distance du χ^2**



$Q = \tilde{D}_J^{-1}$

distance entre 2 profils-lignes

$$d^2(i, i') = \sum_j \frac{1}{f_{+j}} \left(\frac{f_{ij}}{f_{i+}} - \frac{f_{i'j}}{f_{i'+}} \right)^2$$



$Q = \tilde{D}_I^{-1}$

distance entre 2 profils-colonnes

$$d^2(j, j') = \sum_i \frac{1}{f_{i+}} \left(\frac{f_{ij}}{f_{+j}} - \frac{f_{i'j'}}{f_{+j'}} \right)^2$$

2. Principe de l'AFC

Principe

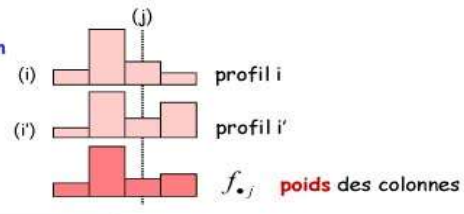
Ressemblance entre profils

La distance du χ^2 est une distance pondérée

La pondération $\frac{1}{f_{\bullet j}}$ équilibre l'influence des colonnes sur la distance entre les lignes

$\frac{1}{f_{i\bullet}}$ équilibre l'influence des lignes sur la distance entre les colonnes

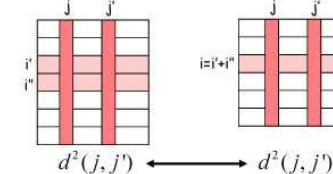
Ex de pondération



2. Principe de l'AFC

Interprétation de la métrique

- Choix des 2 métriques largement justifié par la propriété dite d'équivalence distributionnelle qui en découle :
 - si 2 lignes L_i et $L_{i'}$ de N ont le même profil, les regrouper en une seule d'effectif $n_{ij} + n_{i'j}$ ne modifie pas les distances entre profils-colonnes.
 - Cette propriété, qui n'est pas vérifiée par la métrique euclidienne, va assurer la stabilité des résultats de l'analyse.



Intérêt : assure la robustesse des résultats vis à vis de l'arbitraire du découpage en modalités des variables qualitatives

34

2. Principe de l'AFC

Caractéristiques des nuages : Centre de gravité

- Le nuage N_J (nuage profils-lignes) des I profils-lignes dans \mathbb{R}^J a pour centre de gravité la moyenne pondérée des L_i , soit \bar{x}_L .
- Le nuage N_I (nuage profils-colonnes) des J profils-colonnes dans \mathbb{R}^I a pour centre de gravité la moyenne pondérée des C_j , soit \bar{x}_C .

35

2. Principe de l'AFC

Caractéristiques des nuages : Notion de dualité

- Le terme de métrique du χ^2 vient de ce que les 2 nuages ont la même inertie totale égale à la quantité mesurant leur écart à l'indépendance statistique :

$$I_g = \frac{1}{n_{++}} \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij} - \frac{n_{i+} n_{+j}}{n_{++}}}{\frac{n_{i+} n_{+j}}{n_{++}}} \right)^2 = \frac{1}{n_{++}} D_{\chi^2}$$

$$I(N_J) = \sum_{i=1}^I f_{i+} \|L_i - \bar{x}_L\|_{D_J^{-1}}^2 = \frac{1}{n_{++}} D_{\chi^2} = I(N_I)$$

- Comme en ACP, en AFC on va rechercher une suite d'axes orthogonaux d'inertie maximum sur lesquels projeter le nuage des profils-lignes (resp. profils-colonnes).
- Le cœur de l'AFC est la diagonalisation de matrice dont les valeurs propres sont les inerties projetées triées en ordre décroissant.

36

2. Principe de l'AFC

L'AFC en tant qu'ACP des 2 nuages de profils

- Lorsque l'on dispose d'un tableau de contingence, 2 ACP possibles, sur chacun des nuages ;
- les 3 matrices nécessaires dans les 2 cas sont données ci-dessous :

Profils	Tableau de données : X	Q : métrique	D : pondération
lignes	$D_I^{-1}N = L$	$n_{++}D_I^{-1} = \tilde{D}_I^{-1}$	$D_I / n_{++} = \tilde{D}_I$
colonnes	$D_J^{-1}N^T = C^T$	$n_{++}D_J^{-1} = \tilde{D}_J^{-1}$	$D_J / n_{++} = \tilde{D}_J$

- Conséquences** : on peut aussi voir l'AFC comme une DVS du triplet (X, Q, D) et $I = \text{tr}(X^TDXQ) = \text{tr}(S_I)$

37

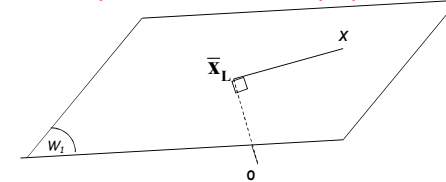
2. Principe de l'AFC

ACP non centrée et facteur trivial

La matrice de var-cov du nuage, après centrage, est :

- pour les lignes $S_L = X^TDXQ - \bar{x}_L\bar{x}_L^T$ et
- pour les colonnes $S_C = X^TDXQ - \bar{x}_C\bar{x}_C^T$

- En fait ce centrage n'est pas utile, car \exists une valeur propre toujours nulle, due à l'orthogonalité des vecteurs moyenne aux sous-espaces des deux nuages resp.
- On peut effectuer une ACP non centrée sur chaque nuage, et ne pas tenir compte de la 1^{ère} valeur propre $\lambda_0 = 1$.



38

2. Principe de l'AFC

ACP non centrée et facteur trivial

Centrage ou non centrage du nuage ?

DVS de $(L_Q = D_I^{-1}P - I_I'F_J, Q = D_J^{-1}, D = D_J)$
avec $I_I = [1 \dots 1]$ et $F_J = [p_1 \dots p_J]$ (centre de gravité du nuage)

Matrice d'inertie : $(P D_I^{-1} - F_J' I_I) D_I (D_I^{-1} P - I_I' F_J) D_J^{-1}$
 $= P D_I^{-1} P D_J^{-1} - F_J' I_I D_I D_I^{-1} P D_J^{-1} - P D_I^{-1} D_I I_I' F_J D_J^{-1} + F_J' I_I D_I I_I' F_J D_J^{-1}$
 $= P D_I^{-1} P D_J^{-1} - F_J' I_I P D_J^{-1} - P I_I' F_J D_J^{-1} + F_J' I_I D_I I_I' F_J D_J^{-1}$
 $= P D_I^{-1} P D_J^{-1} - F_J' I_I P D_J^{-1} - P I_I' F_J D_J^{-1} + F_J' I_I D_I I_I' F_J D_J^{-1}$
 $= P D_I^{-1} P D_J^{-1} - F_J' I_I$
 $= P D_I^{-1} P D_J^{-1} - F_J' I_I$

DVS : $(P D_I^{-1} P D_J^{-1} - F_J' I_I) U_{LD} = U_{LD} \Lambda_{LD}$ avec $I_{LD} = D_I^{-1} U_{LD} U_{LD}^T = I$

On remarque que F_J est vecteur propre de $(P D_I^{-1} P D_J^{-1} - F_J' I_I)$ associé à la valeur propre 0 :
 $P D_I^{-1} P D_J^{-1} F_J - F_J' I_I F_J = F_J \cdot F_J$

On remarque que pour tout vecteur propre u de $P D_I^{-1} P D_J^{-1}$, u est vecteur propre de $P D_I^{-1} P D_J^{-1} - F_J' I_I$ associé à la même valeur propre car $F_J \cdot D_J^{-1} u = 0$ (U_L matrice de vecteurs propres D_J^{-1} orthonormée) donc $F_J' I_I u = \{0\}_{1 \times J}$

Conclusion : La DVS de $(L_Q = D_I^{-1}P - I_I'F_J, Q = D_J^{-1}, D = D_J)$ conduit aux mêmes valeurs propres à l'exception de 1 et aux mêmes vecteurs propres U . Dans la pratique on effectue la DVS sur L et C (ou X) et on élimine la valeur propre 1 et le vecteur propre associé des résultats.

39

2. Principe de l'AFC

ACP non centrée des nuages de profils

- Les facteurs principaux et les composantes principales sont les solutions des diagonalisations suivantes :

ACP	Facteurs principaux	Composantes principales	
	Vect. propres	Vect. propres	normalisation
Lignes	$D_J^{-1}N^T D_I^{-1}N$	F_L de $D_I^{-1}N D_J^{-1}N^T = L C^T$	$F_L^T \frac{D_I}{n_{++}} F_L = \lambda$
Colonnes	$D_I^{-1}N D_J^{-1}N^T$	F_C de $D_J^{-1}N^T D_I^{-1}N = C^T L$	$F_C^T \frac{D_J}{n_{++}} F_C = \lambda$

- Ces analyses conduisent aux mêmes valeurs propres ; simplement échange entre facteurs principaux et composantes principales.
- Les valeurs propres λ sont positives et inférieures à 1.
- Les coordonnées des lignes et des colonnes se déduisent des vecteurs propres associés à ces vp (cf ch. 1).

40

2. Principe de l'AFC

• Compléments sur l'ajustement des nuages en AFC

COMPLÉMENTS SUR L'	AJUSTEMENT DES NUAGES EN AFC
<p>DVS de $(L=D_1^{-1}P, Q=D_1^{-1}, D=D_1)$.</p> <p>Matrice d'inertie</p> $P D_1^{-1} P D_1^{-1} U_L = U_L A_L \text{ avec } U_L D_1^{-1} U_L = I$ $D_1^{-1} P D_1^{-1} P D_1^{-1} D_1 = D_1^{-1} P D_1^{-1} P$ <p>DVS :</p> $L = V_L A_L^{1/2} U_L \Rightarrow V = X \cup X^{1/2}$ <p>Transition</p> $V_L = D_1^{-1} P D_1^{-1} U_L A_L^{-1/2} \Rightarrow V = X \cup X^{1/2}$ <p>Composantes principales</p> $F_L = L D_1^{-1} U_L = D_1^{-1} P D_1^{-1} U_L$	<p>DVS de $(C=D_1^{-1}P, Q=D_1^{-1}, D=D_1)$.</p> <p>Matrice d'inertie</p> $P D_1^{-1} P D_1^{-1} U_C = U_C A_C \text{ avec } U_C D_1^{-1} U_C = I$ $D_1^{-1} P D_1^{-1} P D_1^{-1} D_1 = D_1^{-1} P D_1^{-1} P$ <p>DVS :</p> $C = V_C A_C^{1/2} U_C$ <p>Transition</p> $V_C = D_1^{-1} P D_1^{-1} U_C A_C^{-1/2}$ <p>Composante principale</p> $F_C = D_1^{-1} P D_1^{-1} U_C$

41

2. Principe de l'AFC

• Compléments sur l'ajustement des nuages en AFC

Lien entre les deux DVS

$$P D_1^{-1} P D_1^{-1} U_C = U_C A_C \text{ avec } U_C D_1^{-1} U_C = I \Rightarrow D_1^{-1} P D_1^{-1} P D_1^{-1} U_C = D_1^{-1} U_C A_C \text{ avec } U_C D_1^{-1} D_1^{-1} U_C = I$$

$$\Rightarrow V_L = D_1^{-1} U_C \text{ et } V_C = D_1^{-1} U_L \text{ et } A_C = A_L$$

On en déduit les formules de transition

$$U_L = D_1^{-1} V_C = D_1^{-1} P D_1^{-1} U_C A_C^{1/2} = P D_1^{-1} U_C A_C^{1/2}$$

$$F_L = D_1^{-1} P D_1^{-1} U_L = D_1^{-1} P D_1^{-1} P D_1^{-1} U_C A_C^{1/2} \text{ avec } F_C = D_1^{-1} P D_1^{-1} U_C \Rightarrow F_L = D_1^{-1} P F_C A_C^{1/2} \text{ et par symétrie } F_C = D_1^{-1} P F_L A_C^{1/2}$$

Conclusion : L'ajustement d'un nuage se déduit de celui de l'autre nuage par les formules de transition. L'AFC correspond à l'ajustement des deux nuages et l'étude de leur correspondance.

42

2. Principe de l'AFC

• L'AFC d'ordre k du tableau N correspond à la double ACP sur le triplet :

$$(L = \tilde{D}_1^{-1} P, \tilde{D}_1^{-1}, \tilde{D}_1) \text{ et } (C^T = \tilde{D}_1^{-1} P^T, \tilde{D}_1^{-1}, \tilde{D}_1)$$

qui se résume à l'ACP généralisée d'ordre k du triplet :

$$\left(X = \tilde{D}_1^{-1} P \tilde{D}_1^{-1} = \left[\frac{f_{ij}}{f_{i+} f_{+j}} \right], \tilde{D}_1, \tilde{D}_1 \right)$$

↔ Au sens où les DVS ont les mêmes valeurs singulières - Voir ch ACP p. 28 lien DVS

$$\left(Z = \left[\frac{f_{ij}}{\sqrt{f_{i+}} \sqrt{f_{+j}}} \right], I_J, I_I \right)$$

On a alors :

- Coordonnées des profils-lignes sur l'axe k : $F_X^k = L \tilde{D}_1^{-1/2} (U_Z)_k = F_L^k$ (cf DVS classique U_Z matrice des vecteurs propres de $Z^T Z$) puis en utilisant la formule de transition
- Coordonnées des profils-colonnes sur l'axe k : $G_X^k = \sqrt{\lambda_k} \tilde{D}_1^{-1/2} (U_Z)_k = F_C^{k43}$

2. Principe de l'AFC

Explication solution alternative DVS de $(\tilde{D}_1^{-1} P \tilde{D}_1^{-1} = \left[\frac{f_{ij}}{f_{i+} f_{+j}} \right], \tilde{D}_1, \tilde{D}_1)$

Solution alternative : DVS de $(X = D_1^{-1} P D_1^{-1}, Q = D_1, D = D_1)$

Le nuage défini par X possède les mêmes distances entre points que L compte tenu des métriques. Les distances entre deux lignes de X sont les mêmes qu'entre deux lignes de L .

Le nuage défini par X possède les mêmes distances entre points que C compte tenu des métriques.

Les distances entre deux lignes de X sont les mêmes qu'entre deux lignes de C . $\rightarrow (X', Q = D_1, D = D_1)$

Matrice d'inertie

$$P D_1^{-1} P D_1^{-1} D_1 = D_1^{-1} P D_1^{-1} P$$

DVS :

$$D_1^{-1} P D_1^{-1} P U_X = U_X A_X \text{ avec } U_X D_1^{-1} U_X = I \Rightarrow U_X = V_C \quad A = A_X$$

$$D_1^{-1} P D_1^{-1} P V_X = V_X A_X \text{ avec } V_X D_1^{-1} V_X = I \Rightarrow V_X = V_L$$

Transition

$$X = V_X A_X^{1/2} U_X \text{ et } X = U_X A_X^{1/2} V_X$$

Transition

$$V_X = D_1^{-1} P D_1^{-1} U_X A_X^{-1/2} = D_1^{-1} P U_X A_X^{-1/2} \text{ et } U_X = D_1^{-1} P D_1^{-1} V_X A_X^{1/2} = D_1^{-1} P V_X A_X^{1/2}$$

Composantes principales

$$F_X = D_1^{-1} P D_1^{-1} U_X A_X = D_1^{-1} P U_X A_X = D_1^{-1} P V_C = D_1^{-1} P D_1^{-1} U_L \Rightarrow F_L = F_X$$

$$G_X = D_1^{-1} P D_1^{-1} V_X A_X = D_1^{-1} P V_X A_X = D_1^{-1} P V_L = D_1^{-1} P D_1^{-1} U_C \Rightarrow F_C = G_X$$

Propriété : L'ACP de L et l'ACP de C revient à faire celle de X . L'AFC revient donc à faire l'ACP de X .

3. Eléments principaux de l'AFC

- **Projections suivant un plan factoriel** : les coord. d'un profil-ligne i suivant un plan (U_α, U_α') sont $(F^\alpha(i), F^{\alpha'}(i))$.
- **Dimension de représentation** est égale à $K = \min(I - 1, J - 1)$:
on fera donc les calculs de diagonalisation sur la matrice de plus petite dim. Comme en ACP, l'**inertie** est décomposée :
$$I_g = \lambda_1 + \lambda_2 + \dots + \lambda_K$$
 - Chaque composante (exprimable en % d'inertie), traduit une part de $(n_{ij} - n_{i+}n_{+j}/n_{++})$ écart à l'indépendance des obs.
- **Représentation simultanée**
 - La parfaite symétrie entre ACP des profils-lignes et ACP des profils-colonnes conduit à **superposer** les plans principaux des 2 ACP :
➢ Obtention possible d'une **représentation simultanée** des catégories des 2 variables croisées dans la matrice N.

45

3. Eléments principaux de l'AFC

• Formules de transition

- Passage des coord. d'un ens. à celles de l'autre par des **formules** dites **de transition** :
➢ Intérêt : éviter de réaliser 2 diagonalisations : diagonaliser la matrice la + petite !
- Pour chaque axe noté k ($k = 1, \dots, K$), connaissant les K vecteurs coord. des pts-lignes F_X^k , on en déduit les K vecteurs coord. des pts-colonnes (et G_X^k réciproquement) :

$$\begin{cases} G_X^k = D_J^{-1} N^T F_X^k / \sqrt{\lambda_k} \text{ soit } G_X^k(j) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^I \frac{n_{ij}}{n_{+j}} F_X^k(i) \\ F_X^k = D_I^{-1} N G_X^k / \sqrt{\lambda_k} \text{ soit } F_X^k(i) = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^J \frac{n_{ij}}{n_{i+}} G_X^k(j) \end{cases}$$

Remarque : à noter conséquence sur l'interprétation de la représentation simultanée.

46

3. Eléments principaux de l'AFC

• Relations quasi-barcentriques

Principe

Représentation simultanée ligne-colonnes

Au coefficient $\frac{1}{\sqrt{\lambda_k}}$ près, les projections des points d'un nuage sur un axe sont les barycentres des projections des points de l'autre nuage
= **propriété barycentrique**

la projection de la modalité i sur un axe est le barycentre des modalités j de l'autre variable pondérées par les fréquences conditionnelles du profil de i

Les éléments « lourds » attirant le barycentre, une colonne j attire d'autant plus une ligne i que la valeur de f_{ij} est élevée

Les points éloignées de l'origine sont les profils les plus différents du profil moyen

47

3. Eléments principaux de l'AFC

• Reconstitution des données

- Comme en ACP, on peut reconstituer les valeurs du tableau analysé par :

$$n_{ij} = \frac{n_{i+}n_{+j}}{n_{++}} \left(1 + \sum_{k=1}^K F_X^k(i) G_X^k(j) / \sqrt{\lambda_k} \right)$$

Autre façon de voir l'écart à l'indépendance des observations

48

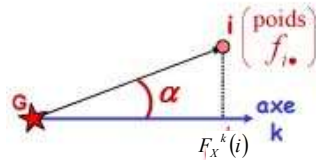
3. Eléments principaux de l'AFC

• Interprétation des axes : \cos^2 et ctr

- Les indices d'aides à l'interprétation définis en ACP sont valables pour un nuage quelconque et s'appliquent donc en AFC

- Qualité de représentation (contribution relative: qlt)

$$qlt_k(i) = \cos^2 \alpha = \frac{\text{inertie de } i \text{ projetée sur l'axe } k}{\text{inertie totale de } i}$$



Même principe pour j .

49

3. Eléments principaux de l'AFC

• Interprétation des axes

- Utilisation des contributions des modalités aux inerties des axes factoriels (i.e. aux valeurs propres) :

➤ Contributions absolues : ctr

- Contribution du profil-ligne i à l'inertie de l'axe k :

$$ctr_k(i) = \frac{\text{inertie de } i \text{ projetée sur l'axe } k}{\text{inertie de } N_i \text{ projetée sur l'axe } k} = \frac{\frac{n_{i+}}{n_{++}} (F_X^k(i))^2}{\lambda_k}$$

- Contribution du profil-colonne j à l'inertie de l'axe k :

$$ctr_k(j) = \frac{\text{inertie de } j \text{ projetée sur l'axe } k}{\text{inertie de } N_j \text{ projetée sur l'axe } k} = \frac{\frac{n_{+j}}{n_{++}} (G_X^k(j))^2}{\lambda_k}$$

50

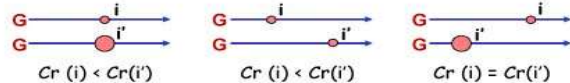
3. Eléments principaux de l'AFC

• Interprétation des axes

➤ Contributions absolues

Remarque :

En ACP, les poids de tous les éléments sont en général égaux
En AFC, ce n'est pas le cas et les poids interviennent dans la contribution d'un point à l'inertie d'un axe.



- Critère simple consiste à retenir les lignes i tq $ctr_k(i) > \frac{n_{i+}}{n_{++}}$
(resp. les colonnes j tq $ctr_k(j) > \frac{n_{+j}}{n_{++}}$)

51

3. Eléments principaux de l'AFC

Conclusion

- Méthode particulièrement bien adaptée à l'étude d'un tableau de contingence (historiquement imaginée pour ce type de tableau)
- Propriétés remarquables : représentation simultanée
- Méthode couramment appliquée à d'autres tableaux (ex : espèces x échantillons en écologie)
 - Condition d'application : valeurs positives
 - On ne raisonne plus en terme de liaison entre deux variables qualitatives
 - Typologie des lignes et des colonnes à travers leurs profils

⇒ Faire Exercice « à la main » dans Td AFC II Partie A.

52

4. Exemple sous

tiré de © 2006, André Bouchier (20 Janvier 2006).

1. Rappels

- L'ACP nous permet de projeter sur un plan un résumé pertinent d'un tableau de données quantitatives.
- Ici, nous avons 2 variables **qualitatives** dont le croisement des modalités donne le tableau de fréquences. Nous travaillons ici sur des effectifs.
- L'ACP simples nous permettra de projeter sur un plan les structures d'un tableau de contingence.
- **Attention** : l'ACP étudie la structure des écarts à l'indépendance, pas leur intensité.

53

4. Exemple sous

2. Les données d'exemple : une table de contingence

- Le tableau des données d'exemple : `housetasks(ade4)`. Il contient 13 « **tâches ménagères** » et leur répartition dans le couple.

Il est rebaptisé ici `TacheMenage` : c'est une table de contingence avec des effectifs :

- les lignes sont 13 tâches ménagères et
- les colonnes indiquent si elles sont réalisées par la femme, alternativement, par l'homme ou de concert
- Chaque valeur numérique du tableau de données est donc un effectif.

54

4. Exemple sous

3. Lecture des données :

Le tableau de données est fourni

```
> library(ade4)
> TacheMenage <- read.table("TacheMenage.txt", h=T, row.names=1)
> TacheMenage
```

	Femme	Alternativement	Homme	Ensemble
Lessive	156	14	2	4
Repas	124	20	5	4
Diner	77	11	7	13
Déjeuner	82	36	15	7
Nettoyage	53	11	1	57
Vaisselle	32	24	4	53
Achats	33	23	9	55
Officiel	12	46	23	15
Conduite	10	51	75	3
Finances	13	13	21	66
Assurances	8	1	53	77
Réparations	0	3	160	2
Vacances	0	1	6	153

55

4. Exemple sous

3. Lecture des données :

Relation entre les lignes et les colonnes par un test du χ^2

```
# Résultats du test du Chi-deux d'indépendance
> chisq.test(TacheMenage)
Pearson's Chi-squared test
data: TacheMenage
X-squared = 1944.456, df = 36, p-value < 2.2e-16

=> Rejet de l'indépendance ; nous pouvons donc effectuer une AFC
```

56

4. Exemple sous R

4. L'AFC - fonction `dudi.coa()` de la bibliothèque `ade4`:

- Les résultats de l'AFC sont stockés dans la variable `z`

```
> z<-dudi.coa(df = TacheMenage, scannf = F, nf = 3)
```

- L'éboulis des valeurs propres

```
> inertie<-z$eig/sum(z$eig)*100
```

```
> barplot(inertie,ylab="% d'inertie",names=round(inertie,2))
```

```
> title("Eboulis des valeurs propres en %")
```

- Les valeurs propres

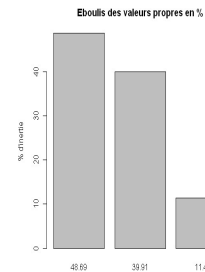
```
> round(z$eig,2)
```

```
[1] 0.54 0.45 0.13
```

- Les valeurs propres en %

```
>round(z$eig/sum(z$eig)*100,2)
```

```
[1] 48.69 39.91 11.40
```



Ici si nous conservons trois axes, nous gardons toute l'information => la somme des vp est égale à 100 % car nb d'axes=min(4-1,12-1)=3,

57

4. Exemple sous R

5. Interprétation des facteurs : les contributions

- Contributions absolues** (ctr) des colonnes à la construction des axes :

```
> inertia.dudi(z,col.inertia = T)$col.abs/100
```

	Comp1	Comp2	Comp3
Femme	44.46	10.31	10.82
Alternativement	0.10	2.78	82.55
Homme	54.23	17.79	6.13
Ensemble	1.20	69.12	0.50
Somme	100	100	100

- Les q1t ne sont pas indiquées ; mais il faut les regarder avec attention avant de commenter !

```
> inertia.dudi(z,col.inertia = T)$col.rel/100
```

58

4. Exemple sous R

6. Interprétation des facteurs : les contributions

- Contributions absolues** des lignes à la construction des axes :

```
> inertia.dudi(z,row.inertia = T)$row.abs/100
```

	Axis1	Axis2	Axis3
Lessive	18.29	5.56	7.97
Repas	12.39	4.74	1.86
Diner	5.47	1.32	2.10
Déjeuner	3.82	3.70	3.07
Nettoyage	2.00	2.97	0.49
Vaisselle	0.43	2.84	3.63
Achats	0.18	2.52	2.22
Officiel	0.52	0.80	36.94
Conduite	8.08	7.65	18.60
Finances	0.88	5.56	0.06
Assurances	6.15	4.02	5.25
Réparations	40.73	15.88	16.60
Vacances	1.08	42.45	1.21
Somme	100	100	100

- Les q1t ne sont pas indiquées ; mais il faut les regarder avec attention avant de commenter !

```
> inertia.dudi(z,row.inertia = T)$col.rel/100
```

59

4. Exemple sous R

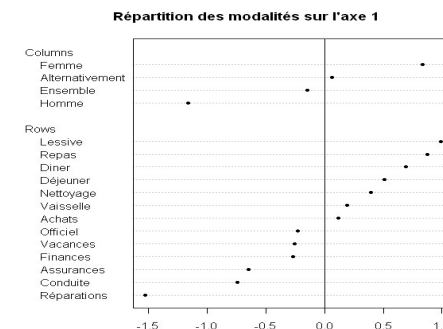
7. AFC simple, une aide à l'interprétation - axe 1:

```
# Aide graphique à l'interprétation des axes
```

```
> score.coa (TacheMenage.afc,xax = 1,dotchart = TRUE) >
```

```
title("Répartition des modalités sur l'axe 1")
```

```
> abline(v=0)
```

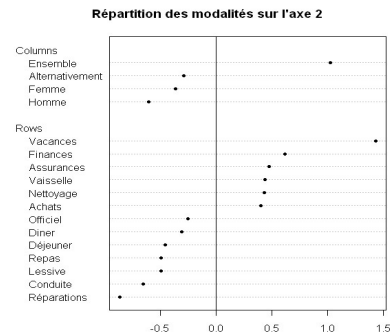


60

4. Exemple sous R

8. AFC simple, une aide à l'interprétation - axe 2 :

```
> score.coa (TacheMenage.afc,xax = 2,dotchart = TRUE)
> title("Répartition des modalités sur l'axe 2")
> abline(v=0)
```

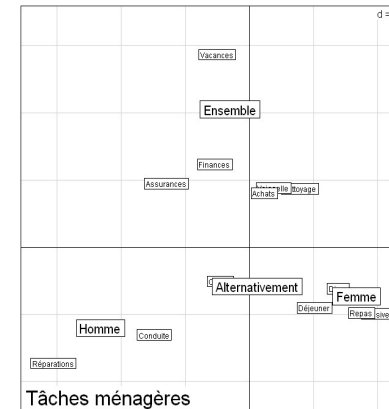


61

4. Exemple sous R

9. Représentation du plan factoriel

```
> scatter.coa(TacheMenage.afc, method=1,sub="Tâches ménagères",posieg="none")
```



le premier plan factoriel représente près de 88.6 % de l'information du tableau de contingence :

la dépendance entre tâches ménagères et sexe n'est pas qu'une idée de militantes féministes !

62

4. Exemple sous R

10. Les données supplémentaires :

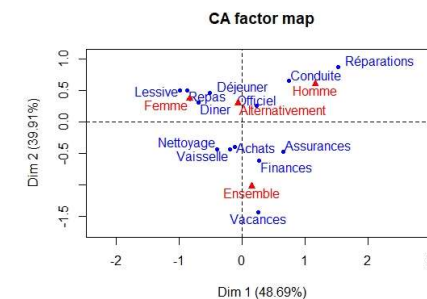
- La bibliothèque `ade4` propose les fonctions `supcol()` et `suprow()` pour calculer les coordonnées des variables et individus supplémentaires. Ces fonctions s'utilisent après le calcul de l'AFC.

63

4. Exemple sous R

11. Avec CA (FactoMineR)

```
library(FactoMineR)
res.ca<-CA(TacheMenage, ncp=2, graph=T)
```



64

4. Exemple sous R

11. Avec CA (FactoMineR)

summary(res.ca)

Call:

CA(X = TacheMenage, **kcp = 2**, graph = T)

The **chi square of independence** between the two variables is equal to 1944.456 (p-value = 0).

Eigenvalues
Variance 0.543 0.445 0.127
% of var. 48.692 39.913 11.395
Cumulative % of var. 48.692 88.605 100.000

Rows (the 10 first)

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Lessive	134.160	-0.992	18.287	0.740	0.495	5.564	0.185
Repas	90.692	-0.876	12.389	0.742	0.490	4.736	0.232
Diner	38.246	-0.693	5.471	0.777	0.308	1.321	0.154
Déjeuner	41.124	-0.509	3.825	0.505	0.453	3.699	0.400
Nettoyage	24.667	-0.394	1.998	0.440	-0.434	2.966	0.535
Vaisselle	19.587	-0.189	0.426	0.118	-0.442	2.844	0.646
Achats	14.970	-0.118	0.176	0.064	-0.403	2.515	0.748
Officiel	53.300	0.227	0.521	0.053	0.254	0.796	0.066
Conduite	101.509	0.742	8.078	0.432	0.653	7.647	0.335
Finances	29.564	0.271	0.875	0.161	-0.618	5.559	0.837

Columns

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Femme	301.019	-0.838	44.462	0.802	0.365	10.312	0.152
Alternativement	117.824	-0.062	0.104	0.005	0.292	2.783	0.105
Homme	381.373	1.161	54.234	0.772	0.602	17.787	0.208
Ensemble	314.725	0.149	1.200	0.021	-1.027	69.118	0.977

Attention, 10 modalités max, selon % d'inertie totale expliquée par la modalité et non ctr par axe !!!!! D'où absence de réparation

65

4. Exemple sous R

12. Avec le package Factoshiny

Factoshiny permet d'utiliser une interface graphique interactive et conviviale pour réaliser les analyses précédentes via la fonction **CAshiny** !

library(Factoshiny)

CAshiny(TacheMenage)

66

5. l'Analyse Factorielle des Correspondances Multiples (AFCM)

• Introduction

- 1941 - Guttman
- 1950 - Burt
- 1956 - Hayashi

- Homogeneity analysis
- Dual scaling

Extension du domaine d'application de l'AFC
Procédures de calcul et règles d'interprétation spécifiques

67

5. L'AFCM

Objectifs

Les objectifs de l'ACM font intervenir **trois familles d'objets**:

- **Typologie des individus**
Basée sur une notion de ressemblance : 2 individus sont proches s'ils possèdent un grand nombre de modalités en commun
- **Liaisons entre variables**
Implique de se situer au niveau des modalités
Cherche à résumer l'ensemble des variables par un petit nombre de variables synthétiques
- **Typologie des modalités**
Deux modalités se ressemblent si :
 - elles sont présentes ou absentes chez un grand nombre d'individus
 - elles s'associent beaucoup ou peu aux mêmes autres modalités

Problématique riche et complexe qui s'articule autour de la typologie des modalités

5. L'AFCM

Domaine d'application Tableau **individus** x **variables qualitatives**

Exemple : enquêtes socio-économiques

		Variable 1	...	Variable j	...	Variable G
Individuals	1	mod ₂		mod ₁		mod ₂
	2	mod ₃		mod ₃		mod ₃
	3	mod ₁		mod ₃		mod ₁
	...	mod ₂	...	mod ₁	...	mod ₁
	...	mod ₁		mod ₂		mod ₃
	h	mod ₂		mod ₁		mod ₂

mod_k : modalité k de la variable j
 J_j : nombre de modalités de la variable j

Sous cette forme, le tableau n'est pas exploitable
→ **Recodage des variables**

69

5. L'AFCM : Notions de base

- Codage disjonctif

- C variables mesurées sur n individus ;
- la $j^{\text{ème}}$ variable a J_j modalités, $j = 1, \dots, C$;
- Construction à partir du tableau de données, du tableau Z à n lignes et $J = \sum_{j=1}^C J_j$ colonnes décrivant les C réponses des n individus par un codage binaire :

$$\mathbf{Z}_{n \times J} = [\mathbf{Z}_{1(n \times J_1)} \vdots \mathbf{Z}_{2(n \times J_2)} \vdots \cdots \vdots \mathbf{Z}_{C(n \times J_C)}]$$

- Le sous-tableau Z_j est tq sa $i^{\text{ème}}$ ligne contient $J_j - 1$ fois la valeur 0 et une fois la valeur 1 (modalité choisie par l'individu i).

70

5. L'AFCM : Notions de base

- **Codage disjonctif** (d'après Lebart et al. (2006))

Diagram illustrating the transformation of a matrix R into a matrix Z .

Matrix R is of size (n, C) with $C=3$. Matrix Z is of size (n, J) with $J=9$.

The transformation is shown as:

$$R = \begin{bmatrix} 2 & 2 & 4 \\ 2 & 1 & 3 \\ 3 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 1 & 2 \\ 2 & 2 & 3 \\ 3 & 2 & 2 \\ 1 & 1 & 4 \end{bmatrix} \rightarrow Z = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

71

5. L'AFCM : Notions de base

- **codage disjonctif** (d'après Lebart et al. (2006))

- Le tableau Z est appelé **tableau disjonctif complet**, de terme général :

- Selon que le sujet i a choisi la modalité j : $z_{ij} = 1$ ou $z_{ij} = 0$

	<i>Variable 1</i> 1 ... J ₁				<i>Variable k</i> 1 ... k ... J _k				<i>Variable C</i> 1 ... J _C		<i>Marges en ligne</i>		
<i>Individus</i>	1					.					C		
	i	0	1	0	...		z _{ik}		...		0	1	C
						.							C
						.							C
	n					.							C
<i>Marges en colonnes</i>	z ₊₁		z _{+J₁}		z _{+J_C}		nC	<i>Effectif total du tableau Z</i>

72

5. L'AFCM : Notions de base

• Codage disjonctif (d'après Lebart et al. (2006))

▪ Le tableau Z est donc tq :

➤ ses **marges en ligne** sont constantes et égales au nombre de variables C :

$$z_{i+} = \sum_{j=1}^J z_{ij} = C$$

➤ ses **marges colonnes** correspondent au nb de sujets ayant choisi la modalité j :

$$z_{+j} = \sum_{i=1}^n z_{ij}$$

➤ on vérifie que pour chaque ss-tableau Z_j , l'effectif total est bien n ;

➤ la somme des marges donne l'**effectif total** de Z : nC

73

5. L'AFCM : Notions de base

• Tableau de contingence de Burt

La données de 2 variables mises sous forme disjonctive complète permet d'aboutir au tableau de contingence utilisé pour l'AFC :

$$N_{I \times J} = Z_{1(n \times I)}^T Z_{2(n \times J)}$$

L'analyse du tableau croisant plus de deux partitions **se généralise** au cas $C > 2$.

74

5. L'AFCM : Notions de base

Notions de base

Z = tableau disjonctif complet

	V_1	V_2	V_C
Z (n,J)	0 1 0 0	1 0 0	0 1 0

Tableau de Burt

B = tableau de contingence de Burt
= juxtaposition de tableaux de contingence

$$B = Z'Z$$

(J,J)

	V_1	V_2	V_C
V_1	0		
V_2		0	
V_C			0

tableaux de contingence entre deux variables

effectifs des modalités de chaque variable

5. L'AFCM : Notions de base

• Tableau de contingence de Burt B

▪ B est une juxtaposition de tableaux de contingence ;

▪ B est formé de C^2 blocs ;

▪ ses **marges** sont pour tout $j \leq J$:

$$b_j = \sum_{j'=1}^J b_{jj'} = C \times z_{+j}$$

▪ L'**effectif total** de B vaut :

$$b = \sum_j b_j = C^2 n$$

76

5. L'AFCM : Notions de base

• Tableau de contingence de Burt B

- On désigne par **D** la matrice diagonale $J \times J$ ayant les mêmes éléments diagonaux que B (effectifs de chacune des modalités):

$$d_{jj} = b_{jj} = Z_{+j} \text{ et } d_{jj'} = 0 \forall j' \neq j.$$

- Suite exemple Lebart et al. (2006) Cf. slide 65

$J = 9$											
B =						D =					
(J,J)						(J,J)					
4	0	0	2	2	1	0	1	2	4	0	0
0	5	0	2	3	0	1	3	1	0	5	0
0	0	3	2	1	1	2	0	0	0	0	3
2	2	2	6	0	2	2	1	1	0	0	0
2	3	1	0	6	0	1	3	2	0	0	0
1	0	1	2	0	2	0	0	0	1	0	1
0	1	2	2	1	0	3	0	0	0	1	2
1	3	0	1	3	0	0	4	0	0	0	3
2	1	0	1	2	0	0	0	3	0	0	0

5. L'AFCM : Principes de base

- La problématique de l'AFCM est :

- apparentée à celle de l'ACP (tableau **individus** x **variables**) ; mais
- peut être considérée comme une **généralisation** de l'AFC (liaisons entre plusieurs variables qualitatives).

L'AFCM est l'AFC d'un tableau disjonctif complet $Z \in \mathcal{M}_{n \times J}$.

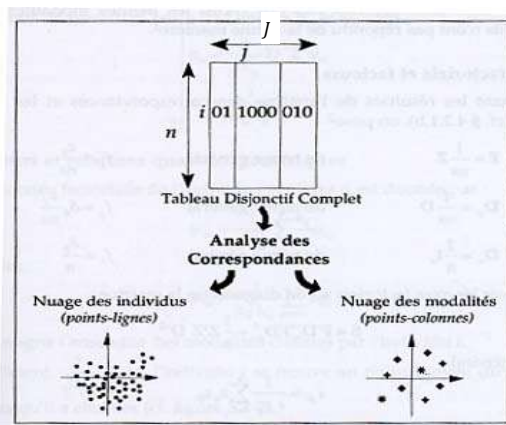
- Ses principes sont ceux de l'AFC :

- transformations** du tableau de données **en profils-lignes** et **profils-colonnes** ;
- pondération** des points par leurs profils marginaux ;
- distance** du χ^2 .

78

5. L'AFCM: Principes de base

Tiré de Lebart et al. (2006) p193



79

5. L'AFCM : Principes de base

- Cas de deux variables** à I et J modalités resp.

Le tableau que nous avons utilisé pour l'AFC s'écrit :

$$N_{I \times J} = Z_{1(n \times I)}^T \times Z_{2(n \times J)}$$

Soit **D** **matrice des effectifs marginaux** des modalités :

$$D = \begin{bmatrix} D_I & 0 \\ 0 & D_J \end{bmatrix}$$

où

- D_I matrice diagonale de dim I contenant les effectifs n_{i+}
- D_J matrice diagonale de dim J contenant les effectifs n_{+j} ;
- D matrice carrée d'ordre $I + J$.

On va voir que : l'AFC du tableau disjonctif **Z** est équivalente à l'AFC de **N**

80

5. L'AFCM : Principes de base

• Cas de 2 variables

- Le **tableau des profils lignes** L de Z est le **tableau des fréquences conditionnelles** (n_{ij}/n_{i+}) i fixé.
C'est exactement le tableau : $Z/2$
- Le **tableau des profils-colonnes** C de Z est le **tableau des fréquences conditionnelles** (n_{ij}/n_{+j}) j fixé.

C'est le tableau : $ZD^{-1} = \begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} D_I^{-1} & 0 \\ 0 & D_J^{-1} \end{bmatrix}$

- On sait que les **coordonnées des profils colonnes de Z sont les vecteurs propres de $C^T L$** (voir AFC p. 40) d'où :
 $C^T L = (ZD^{-1})^T Z / 2 = \frac{1}{2} D^{-1} Z^T Z = \frac{1}{2} D^{-1} B$; B tableau de Burt.

81

5. L'AFCM : Principes de base

• Cas de deux variables

- L'équation donnant les $I + J$ coordonnées des profils colonnes de Z est :

$$\frac{1}{2} \begin{bmatrix} D_I^{-1} & 0 \\ 0 & D_J^{-1} \end{bmatrix} \begin{bmatrix} D_I & N \\ N^T & D_J \end{bmatrix} \begin{bmatrix} F_L \\ F_C \end{bmatrix} = \mu \begin{bmatrix} F_L \\ F_C \end{bmatrix}$$

où $D^{-1} B = \begin{bmatrix} D_I^{-1} & 0 \\ 0 & D_J^{-1} \end{bmatrix} \begin{bmatrix} D_I & N \\ N^T & D_J \end{bmatrix}$

En notant F_L les I premiers composantes et F_C les J suivantes. Soit :

$$\begin{bmatrix} I_I & D_I^{-1} \\ D_J^{-1} & I_J \end{bmatrix} \begin{bmatrix} F_L \\ F_C \end{bmatrix} = 2\mu \begin{bmatrix} F_L \\ F_C \end{bmatrix}$$

- D'où les équations :

$$\begin{cases} F_L + D_I^{-1} N F_C = 2\mu F_L & \text{ou} \dots & D_J^{-1} N^T D_I^{-1} N F_C = (2\mu - 1)^2 F_C \\ D_J^{-1} N^T F_L + F_C = 2\mu F_C & & D_I^{-1} N D_J^{-1} N^T F_L = (2\mu - 1)^2 F_L \end{cases}$$

On reconnaît les équations de l'AFC de N (page 40) avec $\lambda = (2\mu - 1)^2$

82

5. L'AFCM : Principes de base

• Cas de deux variables

Soit $\begin{cases} D_J^{-1} N^T D_I^{-1} N F_C = (2\mu - 1)^2 F_C \\ D_I^{-1} N D_J^{-1} N^T F_L = (2\mu - 1)^2 F_L \end{cases}$

On reconnaît les équations de l'AFC vues précédemment dans lesquelles $\lambda = (2\mu - 1)^2$. Il y a au plus $I + J - 2$ vp non trivialement égales à 0 ou 1.

- Si μ_k est la $k^{\text{ième}}$ vp de l'AFC de Z , elle est liée à celle de l'AFC précédente par :

$$\mu_k = \frac{1}{2} (1 + \lambda_k^{1/2})$$

- On montre que l'inertie totale de Z vaut : $I_g = [(I + J)/2] - 1$.
- Les % d'inertie sont donc très différents entre AFC de Z et AFC de N et ne peuvent être interprétés sans précautions !

83

5. L'AFCM : Principes de base

• Cas général : $C > 2$ variables

L' AFCM est l'AFC de Z :

$$Z_{n \times J} = [Z_1(n \times J_1) \quad Z_2(n \times J_2) \quad \dots \quad Z_C(n \times J_C)]$$

où le nombre total de modalités est $J = \sum_{c=1}^C J_c$

Comme la somme de chaque ligne vaut C , Le tableau des profils lignes est : Z/C

Pour chaque valeur propre μ , on a l'équation donnant les coordonnées des modalités des C variables :

$$\frac{1}{C} D^{-1} B F = \mu F$$

Où $F = [F_1 \quad F_2 \quad \dots \quad F_C] \in \mathbb{R}^J$ le vecteur des coordonnées factorielles des modalités des C variables, il a J composantes.

84

5. L'AFCM : Principes de base

- **Cas général** : $C > 2$ variables

Comme la somme des élt de \mathbf{Z} vaut nC , on adopte comme **normalisation des vecteurs propres** \mathbf{F} :

$$\frac{1}{nC} \mathbf{F}^T \mathbf{D} \mathbf{F} = \mu$$

Une fois calculés les vecteurs propres, on peut calculer les coordonnées \mathbf{G} des individus (les lignes de \mathbf{Z}) par les **formules de transition** :

$$\mathbf{G} = \frac{1}{C\sqrt{\mu}} \mathbf{Z} \mathbf{F} ;$$

$$\text{on a aussi } \mathbf{F} = \frac{1}{\sqrt{\mu}} \mathbf{D}^{-1} \mathbf{Z}^T \mathbf{G}$$

Relations quasi-barycentriques ...

85

5. L'AFCM: Principes de base

- **Cas général** : $C > 2$ variables

- **Relations quasi-barycentriques**

$$\mathbf{G}^k(i) = \frac{1}{\sqrt{\mu_k}} \sum_{j=1}^J \frac{z_{ij}}{C} \mathbf{F}^k(j)$$

À $\frac{1}{\sqrt{\mu_k}}$ près, la **coordonnée d'un individu** i sur l'axe k est le **barycentre** des coordonnées des modalités auxquelles il appartient.

$$\mathbf{F}^k(j) = \frac{1}{\sqrt{\mu_k}} \sum_{i=1}^n \frac{z_{ij}}{z_{+j}} \mathbf{G}^k(i)$$

À $\frac{1}{\sqrt{\mu_k}}$ près la **coordonnée d'une modalité** j sur l'axe k est le **barycentre** des coordonnées des indiv. possédant cette modalité j .

- Ces relations justifient la représentation simultanée des variables et des individus.
- En éliminant l'axe trivial associé à $\mu_0 = 1$ on obtient **au maximum $J - C$ axes.**

86

5. L'AFCM : Principes de base

- Il est aussi possible de présenter l'AFCM sous forme de DVS et d'obtenir l'écriture la distance du chi-deux entre les profils lignes et celle entre profils colonnes.

- Pas abordé ici

87

5. L'AFCM : interprétations

- **Inertie**

L'inertie totale est égale à :

$$I_g = \frac{J}{C} - 1$$

- I_g ne dépend que du nombre total de variables C et de modalités J ;

- **Nb d'axes à conserver :**

Règle d'interprétation des valeurs propres

- Utiliser les sauts (règle du coude) ou
- La règle $\mu_k > 1/C$. $1/C$ constitue alors la **valeur seuil** pour conserver les facteurs de l'AFCM.
- Ex : avec 3 groupes de variables on ne regarde que les valeurs propres supérieures à 0.3.

L'interprétation de l'importance des coordonnées se fait principalement à l'aide des **coordonnées**, **contributions (ctr)** et **cosinus carrés (qlt)**.

88

5. L'AFCM : interprétations

• Règles d'interprétation

Avec l'AFC, on exprime:

- **La proximité entre individus en terme de ressemblances** : 2 individus proches ont globalement choisi les mêmes modalités.
- **La proximité entre modalités de variables différentes en terme d'association** : 2 modalités de variables différentes sont proches car elles concernent globalement les mêmes individus ou des individus semblables.
- **La proximité entre 2 modalités d'une même variable en terme de ressemblance** entre les groupes d'individus qui les ont choisis.

L'interprétation de l'importance des coordonnées se fait principalement à l'aide des **coordonnées**, **contributions (ctr)** et **cosinus carrés (qlt)**.

89

5. L'AFCM : interprétations

• Contributions

Pour la matrice Z_c qui est formée de J_c modalités, la **coordonnée de la $j^{\text{ème}}$ modalité de la variable c** étant $F^k(c_j)$ sur le $k^{\text{ème}}$ axe, sa **contribution absolue** vaut :

$$ctr_k(c_j) = \frac{1}{\mu_k} \frac{n_{c_j}}{nC} (F^k(c_j))^2$$

et

la **contribution absolue** de la **variable c** à l'axe k vaut :

$$ctr_k(c_+) = \sum_{j=1}^{J_c} ctr_k(c_j) = \frac{1}{\mu_k} \sum_{j=1}^{J_c} \frac{n_{c_j}}{nC} (F^k(c_j))^2$$

Reste à observer les **contributions à l'inertie**

90

5. L'AFCM: interprétations

• Contributions

L'**inertie totale de la variable c** correspondant à Z_c :

$$I_{Z_c} = \sum_{j=1}^{J_c} \frac{1}{C} \left(1 - \frac{n_{c_j}}{n} \right) = \frac{J_c - 1}{C}$$

La **contribution de Z_c à l'inertie totale** est égale à :

$$ctr_{Z_c} = \frac{I_{Z_c}}{I_g} = \frac{J_c - 1}{J - C}$$

- **Remarque** : Cette contribution est fonction du nombre de modalités de la variable.
- Pour éviter des contributions artificiellement élevées, préférable d'avoir des **nombre de modalités les plus voisins possible pour chacune des C variables**.

91

5. L'AFCM : exemples sous R

- **Library sous R** permettant de réaliser une **AFCM** :

MCA (FactoMineR) ou

dudi.acm(ade4) ou

mca(MASS)

92

5. L'AFCM : exemple credit sous R

Les données : (R Pour la statistique et la science des données, 2018)

Le tableau des données **credit.csv** contient 66 clients ayant souscrit un crédit à la consommation dans un organisme de crédit. Les 11 variables qualitatives et les modalités sont :

Marché : indique le bien pour lequel les clients ont fait un emprunt : rénovation d'un bien, voiture, scooter, moto, mobilier, side-car

Apport : indique si les clients possèdent un apport personnel avant de réaliser l'emprunt : oui, non

Impayé : nb d'échéances impayées par le client : 0, 1, 2, 3

Taux d'endettement : Taux discrétisé en 4 classes : 1 (faible), 2, 3, 4 (fort)

Assurance : type d'assurance : sans assurance, AID (assurance invalidité et décès), AID + chômage, sénior (pour les + de 60 ans)

Famille : union libre (concubinage), marié, veuf, célibataire, divorcé

Enfants à charge : 0, 1, 2, 3, 4 et plus

Logement : propriétaire, accédant à la propriété, locataire, logé par la famille, logé par l'employeur

Profession : ouvrier non qualifié, ouvrier qualifié, retraité, cadre moyen, cadre supérieur

Intitulé : M, Mme, Melle

Age : 20 (18 à 29 ans), 30 (30 à 39 ans), 40 (40 à 49 ans), 50 (50 à 59 ans), 60 et +⁹³

5. L'AFCM : exemple credit sous R

• Lecture des données :

```
> credit <- read.table("credit.csv", sep = ";", header=TRUE)
```

```
> summary(credit)
```

```
Marche      Apport      Impayé      Assurance      Endettement      Famille      Enfants      Logement
Mobilier / Ameublement:17  Apport : 0  AID : 13  End_1:18  Celibataire:17  End_0:39  Accédant + la propriétaire: 6
Moto : 8  pas_Apport:13  Imp_0 : 42  AID + Chômage :13  End_2:15  Divorcé : 5  End_1: 8  Locataire :123
Rénovation :18  Imp_1 et +:16  Sans Assurance:12  End_3:15  Marié :25  End_2:11  Logé par l'employeur : 3
Scooter : 5  Sénior :10  End_4:14  Union libre:13  End_3: 6  Logé par la famille : 6
Side-car : 1  Veuf : 6  End_4: 2  Propriétaire :18
Voiture :17
```

- Variable **Age** n'est pas considérée comme qualitative ! D'où transformation :

```
> credit[, "Age"] <- factor(credit[, "Age"])
```

- Etude des modalités rares : important car AFCM accorde bcp d'importance à ces modalités!

```
> for (i in 1:ncol(credit)){ # permet d avoir les graphes un par un
  par(ask=TRUE) # cliquer sur la fenêtre graphique
  plot(credit[,i]) # pour voir le graphe
}
```

Seule la variable **Marché** présente une modalité rare (1 seul individu pour Side-car) : on regroupe la modalité Side-car avec Moto

```
> plot(credit[, "Marche"])
```

```
> levels(credit[, "Marche"])[5] <- "Moto"
```

94

5. L'AFCM : exemple credit sous R

• Choix des variables et des individus actifs

Objectif : déterminer les profils de comportements bancaires

Variables actives : celles (les 5 premières) correspondant aux informations bancaires

Individus actifs : tous

Utilisation de la fonction **MCA** du package **FactoMineR** pour réaliser une AFCM

```
> library(FactoMineR)
```

```
> res.mca <- MCA(credit, quali.sup = 6:11, level.ventil = 0)
```

- AFCM construite sur les 5 premières variables
- Argument **level.ventil** par défaut à 0 : aucune ventilation effectuée. Si cet argument vaut par ex 5%, les modalités d'effectif $\leq 5\%$ du total des individus seront ventilés de façon automatique (répartition aléatoire) dans les autres modalités.
- L'objet **res.mca** contient l'ensemble des résultats

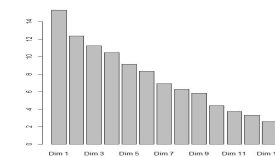
95

5. L'AFCM : exemple credit sous R

• Choix du nombre d'axes

Diagramme en barres des valeurs propres associées à chaque axe

```
> barplot(res.mca$eig[,2], names = paste("Dim", 1:nrow(res.mca$eig)))
```



- Décroissance régulière des vp (pas de coude ou de décrochage flagrant), difficile de choisir le nb d'axes

```
> summary(res.mca)
```

```
Eigenvalues      Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8  Dim.9  Dim.10  Dim.11  Dim.12  Dim.13
Variance      0.399  0.322  0.292  0.272  0.238  0.217  0.180  0.163  0.152  0.115  0.097  0.086  0.067
% of var.     15.330 12.380 11.247 10.471  9.141  8.329  6.935  6.278  5.838  4.419  3.746  3.316  2.570
Cumulative % of var. 15.330 27.710 38.957 49.428 58.569 66.898 73.833 80.111 85.949 90.368 94.114 97.430 100.000
```

- Faibles taux d'inertie et décroissance régulière classique en AFCM. Les 2 premiers axes expriment 28% de l'inertie totale, ce qui est relativement important -> on n'interprétera que ces **2 axes** MAIS il pourrait être intéressant d'analyser les suivants !

96

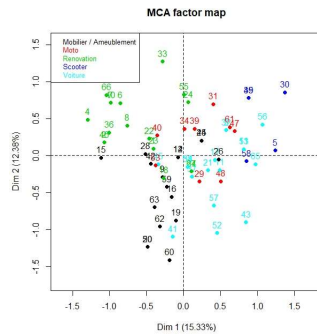
5. L'AFCM : exemple credit sous R

• Analyse des résultats

Interprétation des résultats à l'aide de la fonction `summary` et des graphes des individus et de l'ensemble des modalités dans le plan 1-2

```
> summary(res.mca, nbelements = 2, ncp = 2, nb.dec = 2)
```

```
> plot(res.mca, invisible = c("var", "quali.sup"), habillage = 1)
```



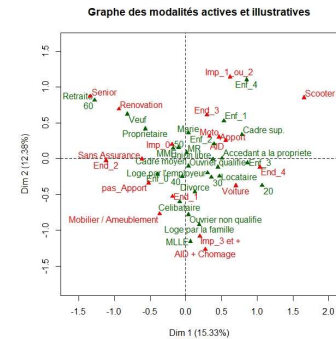
Représentation dans le plan 1-2 des individus, coloriés en fonction des modalités de la variable Marche

97

5. L'AFCM : exemple credit sous R

• Analyse des résultats

```
> plot(res.mca, invisible = "ind", title="Graphe des modalités actives et illustratives")
```

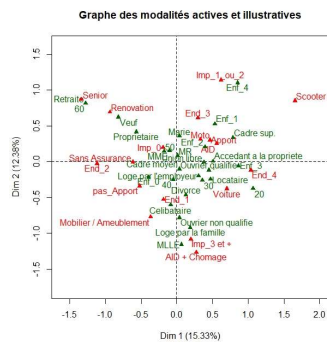


- 1^{er} axe : opposition
 - profil « jeunes » (droite) : crédit pour acheter un scooter par ex
 - profil « seniors » (gauche) : propriétaires ayant contracté un crédit pour financer des travaux de rénovation
 - 2^{eme} axe : difficultés financières
 - Grandes difficultés financières (en bas) : Impaye 3 et + ; AID + Chomage
 - Autres (en haut)
- ⇒ Affiner (s'assurer par ex. des q1 l) en regardant les résultats numériques du `summary` :
- ```
> summary(res.mca, nbelements = 2, ncp = 2, nb.dec = 2)
```

## 5. L'AFCM : exemple credit sous R

### • Analyse des résultats

```
> plot(res.mca, invisible = "ind", title="Graphe des modalités actives et illustratives")
```



- Règles générales d'interprétation de proximité entre modalités de 2 variables différentes ou non :
  - Après s'être assuré que les modalités sont bien représentées dans le plan, on pourra interpréter les proximités (bcp d'individus qui prennent la modalité xx) prennent aussi la modalité (xxx) ie les individus associés à ces modalités ont donc le même profil !

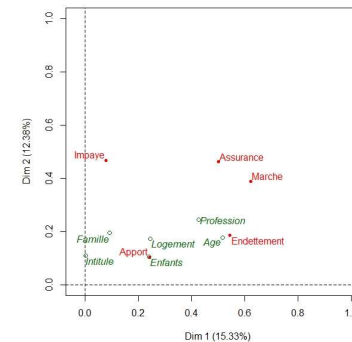
99

## 5. L'AFCM : exemple credit sous R

### • Analyse des résultats

Ctr des variables (et non des modalités des variables !) aux axes  $k$  :  $ctr_k(c_+)$

```
> plot(res.mca, choix="var")
```



- Pour interpréter les axes, on peut dessiner le graphe des ctr par variable et par axe :
- Variables actives :
    - Axe 1 : Marche
    - Axe 2 : Impaye
  - Variables supplémentaires :
    - Axe 1 : Age
    - Axe 2 : ??Profession

Description aussi avec la fonction `dimdesc` :

```
> dimdesc(res.mca)
```

100

## 5. L'AFCM : exemple credit sous R

### • Interface graphique Factoshiny

La fonction **MCashiny** du package **Factoshiny** permet d'utiliser une interface graphique interactive et conviviale pour réaliser les analyses précédentes !

```
> # Factoshiny
> library(Factoshiny)
> res.shiny <- MCashiny(credit)
```

101

## 5. L'AFCM : autre exemple sous R

Les données : © 2006, André Bouchier (20 Janvier 2006)

Le tableau des données **bledur.txt** contient 50 observations et 11 variables. Il contient les résultats d'un suivi agronomique sur 50 parcelles de blé dur :

**RDT** Rendement en grains  
**PLM** Nb de plantes par m<sup>2</sup>  
**ZON** Zone géographique  
**ARG** Taux d'argile de la parcelle  
**LIM** Taux de limon de la parcelle  
**SAB** Taux de sable de la parcelle  
**VRT** Variété cultivée  
**PGM** Poids de 1000 grains  
**MST** Matière sèche totale à la récolte  
**AZP** Azote dans la plante à la récolte  
**VRTC** Variété cultivée (codée en 3 classes)

102

## 5. L'AFCM: autre exemple sous R

### • Lecture des données :

```
> don<-read.table(file.choose(), sep=" ", header=T, dec=",")
```

|    | Numero | RDT    | PLM | ZON | ARG  | LIM  | SAB  | VRT | PGM   | MST   | AZP  | VRTC |
|----|--------|--------|-----|-----|------|------|------|-----|-------|-------|------|------|
| 1  | 1      | 6.490  | 84  | 1   | 21.5 | 60.6 | 17.9 | 3   | 43.10 | 34.49 | 3.82 | 2    |
| 2  | 2      | 15.580 | 112 | 1   | 21.0 | 58.3 | 20.7 | 3   | 38.30 | 39.18 | 3.78 | 2    |
| 3  | 3      | 7.290  | 68  | 1   | 26.2 | 47.6 | 26.2 | 3   | 45.30 | 26.89 | 2.61 | 2    |
| 4  | 4      | 1.090  | 88  | 1   | 29.7 | 54.5 | 15.8 | 3   | 29.09 | 23.09 | 3.78 | 2    |
| 5  | 5      | 5.100  | 174 | 1   | 22.8 | 59.0 | 18.2 | 3   | 42.80 | 18.10 | 3.41 | 2    |
| 6  | 6      | 2.030  | 63  | 1   | 19.6 | 68.0 | 12.4 | 3   | 41.26 | 20.43 | 3.04 | 2    |
| 7  | 7      | 6.330  | 92  | 1   | 26.7 | 53.7 | 19.6 | 3   | 38.57 | 20.93 | 2.26 | 2    |
| 8  | 8      | 17.300 | 117 | 1   | 34.0 | 44.9 | 21.1 | 1   | 31.80 | 40.70 | 3.85 | 1    |
| 9  | 9      | 6.970  | 58  | 1   | 16.7 | 57.6 | 25.7 | 4   | 42.40 | 29.97 | 2.69 | 3    |
| 10 | 10     | 20.350 | 116 | 1   | 37.4 | 49.0 | 13.6 | 6   | 42.90 | 40.75 | 3.62 | 3    |

.../.....

103

## 5. L'AFCM : : autre exemple sous R

### • Codage des données

- Ce tableau de données contient des **valeurs quantitatives et qualitatives**. Il faut, dans un premier temps, le **transformer** en données uniquement qualitatives.
- Chaque **variable quantitative** sera découpée en 3 classes d'effectifs égaux. Pour cela, nous utiliserons la fonction **codage()** - (voir prog après)

```
> source("C:/.../M1/AnalysedeDonnées/data/codage.R")
> RDT<-codage(don$RDT)
> PLM<-codage(don$PLM)
> ARG<-codage(don$ARG)
> LIM<-codage(don$LIM)
> SAB<-codage(don$SAB)
> PGM<-codage(don$PGM)
> MST<-codage(don$MST)
> AZP<-codage(don$AZP)
```

104

## 5. L'AFCM : autre exemple sous R

### • Codage des données

- Les **variables qualitatives** (ou non modifiées) seront transformées en facteurs

```
> ZON<-as.factor(don$ZON)
> VRTC<-as.factor(don$VRTC)
```

105

## 5. L'AFCM : autre exemple sous R

### • Mise en forme des données codées

- Le tableau des données codées :

```
> doncd<-data.frame(RDT,PLM,ARG,LIM,SAB,PGM,MST,AZP,ZON,VRTC)
> row.names(doncd)<-don$Numero
> doncd
```

|    | RDT | PLM | ARG | LIM | SAB | PGM | MST | AZP | ZON | VRTC |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| 1  | 1   | 1   | 2   | 3   | 1   | 3   | 2   | 3   | 1   | 2    |
| 2  | 3   | 2   | 2   | 3   | 1   | 2   | 3   | 3   | 1   | 2    |
| 3  | 1   | 1   | 2   | 3   | 1   | 3   | 1   | 1   | 1   | 2    |
| 4  | 1   | 1   | 3   | 3   | 1   | 1   | 1   | 3   | 1   | 2    |
| 5  | 1   | 3   | 2   | 3   | 1   | 3   | 1   | 2   | 1   | 2    |
| 6  | 1   | 1   | 2   | 3   | 1   | 2   | 1   | 2   | 1   | 2    |
| 7  | 1   | 1   | 2   | 3   | 1   | 2   | 1   | 1   | 1   | 2    |
| 8  | 3   | 2   | 3   | 3   | 1   | 1   | 3   | 3   | 1   | 1    |
| 9  | 1   | 1   | 1   | 3   | 1   | 3   | 2   | 1   | 1   | 3    |
| 10 | 3   | 2   | 3   | 3   | 1   | 3   | 3   | 3   | 1   | 3    |

106

## 5. L'AFCM : autre exemple sous R

### • Vérification du codage des données

```
> summary(doncd)
```

| RDT  | PLM  | ARG  | LIM  | SAB  | PGM  | MST  | AZP  | ZON  | VRTC |
|------|------|------|------|------|------|------|------|------|------|
| 1:17 | 1:17 | 1:17 | 1:17 | 1:17 | 1:17 | 1:17 | 1:17 | 1:17 | 1:24 |
| 2:16 | 2:16 | 2:16 | 2:16 | 2:16 | 2:16 | 2:16 | 2:16 | 2:15 | 2:21 |
| 3:17 | 3:17 | 3:17 | 3:17 | 3:17 | 3:17 | 3:17 | 3:17 | 3:18 | 3: 5 |

Pour être pertinent, un découpage en classes doit respecter 3 principes :

1. pas d'effectifs de classes trop déséquilibrés;
2. des nombres de classes semblables pour toutes les variables;
3. des découpages ayant une signification pour le chercheur.

107

## 5. L'AFCM: autre exemple sous R

### • Transformation des données en tableau disjonctif

- Utilisation de la library **ade4** :

```
> library(ade4)
```

- Création du tableau disjonctif :

```
> disj<-acm.disjonctif(doncd)
```

```
> dim(disj)
```

```
[1] 50 30
```

108

## 5. L'AFCM : autre exemple sous R

- La fonction `dudi.coa(ade4)`

- Les résultats de l'AFCM sont stockés dans la variable `z`

```
> z<-dudi.coa(df = disj, scannf = FALSE, nf = 3)
```

- L'éboulis des valeurs propres

```
> inertie<-z$eig/sum(z$eig)*100
```

```
> barplot(inertie,ylab="%
d'inertie",names.arg=round(inertie,2))
```

```
> title("Eboulis des valeurs propres en %")
```

- Les valeurs propres (20 vp)

```
> round(z$eig,3)
```

```
[1] 0.369 0.277 0.204 0.185 0.159 0.146 0.108 0.101
0.091 0.073 0.068 0.048
```

```
[13] 0.042 0.033 0.030 0.021 0.017 0.013 0.013 0.00
```

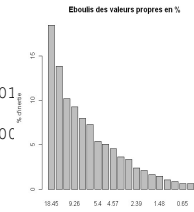
- Les valeurs propres en %

```
> round(z$eig/sum(z$eig)*100,2)
```

```
[1] 18.45 13.86 10.18 9.26 7.96 7.29 5.40 5.05
```

```
4.57 3.65 3.38 2.39
```

```
[13] 2.09 1.64 1.48 1.06 0.85 0.65 0.63 0.19
```

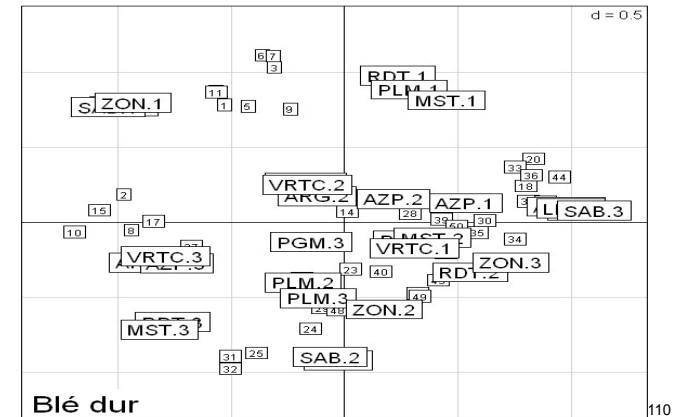


109

## 5. L'AFCM : autre exemple sous R

- Une représentation graphique du plan factoriel :

```
> scatter.coa(z, method = 1, sub = "Blé dur", posieig = "none")
```



110

## 5. L'AFCM : autre exemple sous R

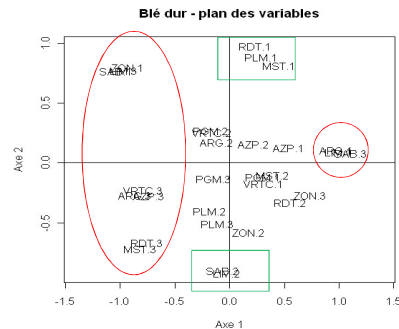
- Une autre représentation graphique du plan factoriel, les variables :

```
> plot(z$co[,1], z$co[,2], type="n", xlab="Axe 1", ylab="Axe 2",
xlim=c(-1.4,1.4))
```

```
> text(z$co[,1], z$co[,2], label= colnames(disj))
```

```
> title("Blé dur - plan des variables")
```

```
> abline(h=0,v=0)
```



111

## 5. L'AFCM : autre exemple sous R

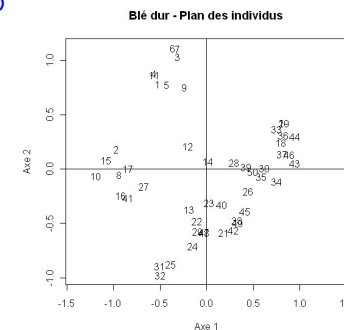
- Une autre représentation graphique du plan factoriel, les individus :

```
> plot(z$li[,1], z$li[,2], type="n", xlab="Axe 1", ylab="Axe
2", xlim=c(-1.4,1.4))
```

```
> text(z$li[,1], z$li[,2], label=row.names(disj))
```

```
> title("Blé dur - Plan des individus")
```

```
> abline(h=0,v=0)
```



112

## 5. L'AFCM : autre exemple sous R

### • Interprétation des axes

- Contributions des variables à la construction des axes :

```
> inertia.dudi(z,col.inertia = T)$col.abs
```

|       | Comp1 | Comp2 | Comp3 |
|-------|-------|-------|-------|
| RDT.1 | 52    | 1178  | 122   |
| RDT.2 | 265   | 128   | 125   |
| RDT.3 | 528   | 544   | 0     |
| PLM.1 | 76    | 966   | 305   |
| PLM.2 | 29    | 188   | 80    |
| PLM.3 | 12    | 317   | 683   |
| ARG.1 | 878   | 13    | 25    |
| ARG.2 | 12    | 35    | 251   |

Cf fig p 101

Axe 1: opposition sab1, LIM1 & arg1 à ...

Axe 2: opposition : RDT.1, PLM.1, MST.1 à LIM.2, SAB.2

|        |     |     |      |
|--------|-----|-----|------|
| ZON.2  | 25  | 367 | 1101 |
| ZON.3  | 528 | 94  | 873  |
| VRTC.1 | 123 | 52  | 3    |
| VRTC.2 | 30  | 96  | 270  |
| VRTC.3 | 169 | 19  | 882  |

Somme 10000 10000 10000

113

## 5. L'AFCM : autre exemple sous R

### • Interprétation des axes

- Contributions des lignes à la construction des axes :

```
> inertia.dudi(z,row.inertia = T)$row.abs
```

|   | Axis1 | Axis2 | Axis3 |
|---|-------|-------|-------|
| 1 | 151   | 444   | 159   |
| 2 | 511   | 25    | 67    |
| 3 | 51    | 774   | 136   |
| 4 | 173   | 555   | 76    |
| 5 | 98    | 434   | 9     |
| 6 | 72    | 900   | 0     |

|    |     |     |     |
|----|-----|-----|-----|
| 46 | 427 | 13  | 35  |
| 47 | 0   | 242 | 698 |
| 48 | 0   | 248 | 669 |
| 49 | 60  | 176 | 0   |
| 50 | 132 | 0   | 566 |

Somme 10000 10000 10000

114

## 5. L'AFCM : autre exemple sous R

### • Une aide à l'interprétation - axe 1

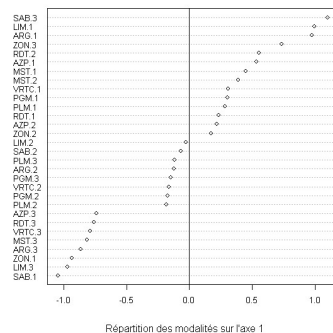
```
> modal<-as.data.frame(z$co)
```

```
> modal<-modal[sort.list(modal$Comp1),]
```

```
> dotchart(modal[,1],labels = row.names(modal),cex=0.8)
```

```
> title(sub="Répartition des modalités sur l'axe 1")
```

```
> abline(v=0)
```



115

## 5. L'AFCM : autre exemple sous R

### • Une aide à l'interprétation - axe 2

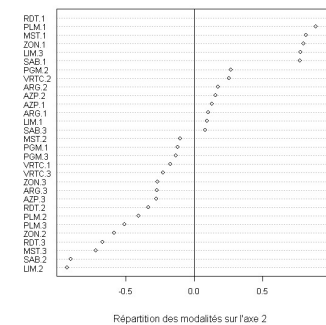
```
> modal<-as.data.frame(z$co)
```

```
> modal<-modal[sort.list(modal$Comp1),]
```

```
> dotchart(modal[,2],labels = row.names(modal),cex=0.8)
```

```
> title(sub="Répartition des modalités sur l'axe 2")
```

```
> abline(v=0)
```



116



## 5. L'AFCM : autre exemple sous R

### • La fonction `codage()`

```
codage<-function(nom)
#découpage en 3 classes d'effectifs égaux
{
#calcul des bornes
bornes<-quantile(nom, probs = c(0, 1/3,2/3,1), na.rm = TRUE,names = TRUE)
#description des bornes et effectifs
Amax<-aggregate(nom,list(Nom=cut(nom,bornes,include.lowest=T,label=F)),max)
Amin<-aggregate(nom,list(Nom=cut(nom,bornes,include.lowest=T,label=F)),min)
Afreq<-as.matrix(summary(as.factor(cut(na.omit(nom),bornes,
include.lowest=T,
label=F))))
limites<-as.data.frame(cbind(Amin[,1],Amin[,2],Amax[,2],Afreq))
names(limites)<-c("Classe","Mini","Maxi","Effectif")
#calcul du nombre de valeurs manquantes
manques<-length(nom)-length(na.omit(nom))
#impression des bornes
cat(paste("Découpage de la variable ",deparse(substitute(nom))," - Nb de
valeurs
manquantes : ",manques,"\\n"))
print(limites)
#découpage de la variable
varfac<-cut(nom,bornes,include.lowest=T,label=F)
#transformation en facteur
as.factor(varfac)
}
```

117

## 5. L'AFCM sous R : PrefConsom

- **PrefConsom**: résultats d'une enquête sur la préférence de consommateurs :
  - 3 variables ( $C=3$ ) : sexe, classe d'âge et produit ;
  - le total des dénombrements des personnes dans les 24 combinaisons vaut  $n = 1000$ .
- La matrice Z a donc 1000 lignes et 9 colonnes ; mais on peut traiter directement le *tableau de Burt* que l'on peut construire à partir des données ci-dessous :

| Classe d'âge    | Sexe | Produit |    |    |    |
|-----------------|------|---------|----|----|----|
|                 |      | A       | B  | C  | D  |
| A1<br>< 20 ans  | M    | 28      | 8  | 6  | 64 |
|                 | F    | 12      | 20 | 6  | 56 |
| A2<br>20-60 ans | M    | 120     | 50 | 40 | 80 |
|                 | F    | 60      | 90 | 80 | 80 |
| A3<br>> 60 ans  | M    | 50      | 12 | 10 | 8  |
|                 | F    | 70      | 28 | 10 | 12 |

118

## 5. L'AFCM sous R : PrefConsom

```
> Pref<-read.table("PrefConsom.txt",h=T)
> Pref
 Nb Produit Sexe Age
1 28 A M A1
2 8 B M A1
...
24 12 D F A3
> summary(Pref)
 Nb Produit Sexe Age
Min. : 6.00 A:6 F:12 A1:8
1st Qu.: 11.50 B:6 M:12 A2:8
Median : 34.00 C:6 A3:8
Mean : 41.67 D:6
3rd Qu. : 65.50
Max. :120.00
Tableau disjonctif complet d'un data frame ne contenant que les facteurs
(acm.disjonctif)
> disj<-acm.disjonctif(Pref[, -1]) #création du tableau disjonctif
> disj[,1]
 Produit.A Produit.B Produit.C Produit.D Sexe.F Sexe.M Age.A1 Age.A2 Age.A3
1 1 0 0 0 0 1 1 0 0
```

119

## 5. L'AFCM sous R : PrefConsom

```
AFCM du tableau des facteurs ; « row.w » fournit les pondérations,
ici les effectifs Nb
> Pref.acm<-dudi.acm(df = as.data.frame(Pref[, -1]),
row.w=as.vector(Pref$Nb), scannf = FALSE, nf = 6) #1'analyse factorielle
Eboulis des valeurs propres (Fig. 5A)
> inertie<-Pref.acm$eig/sum(Pref.acm$eig)*100
> barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))
> title("Eboulis des valeurs propres en %")
Valeurs propres
> round(Pref.acm$eig,4)
[1] 0.4585 0.4188 0.3598 0.3167 0.2414 0.2048
> round(Pref.acm$eig/sum(Pref.acm$eig)*100,2) #les valeurs propres en %
[1] 22.92 20.94 17.99 15.84 12.07 10.24

Plans factoriels (Nuages par modalité des facteurs; Fig, 5B)
> scatter(Pref.acm)
> par(mfrow=c(1,2)) # plan 1-2 et plan 1-3
> s.value(Pref.acm$li, Pref.acm$li[,2])
> s.value(Pref.acm$li, Pref.acm$li[,3])
```

120

## 5. L'AFCM sous R : PrefConsom

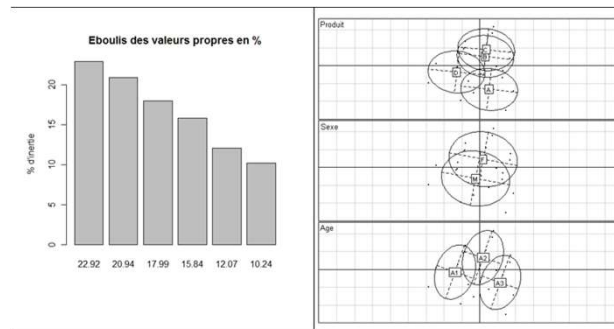


FIGURE 5A. PrefConsom : éboulis des valeurs propres.

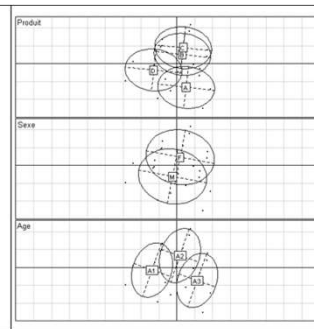


FIGURE 5B. PrefConsom : nuages par modalité des facteurs.

121

## 5. L'AFCM sous R : PrefConsom

```
Aide à l'interprétation : axe 1 (Fig.5C)
> modal<-as.data.frame(Pref.acm$co)
> modal<-modal[sort.list(modal$Comp1),]
> dotchart(modal[,1],labels = row.names(modal),cex=0.8)
> title(sub="Répartition des modalités sur l'axe 1") ; abline(v=0)
Aide à l'interprétation : axe 2 (Fig.5C)
> modal<-as.data.frame(Pref.acm$co)
> modal<-modal[sort.list(modal$Comp2),]
> dotchart(modal[,2],labels = row.names(modal),cex=0.8)
> title(sub="Répartition des modalités sur l'axe 2") ; abline(v=0)
```

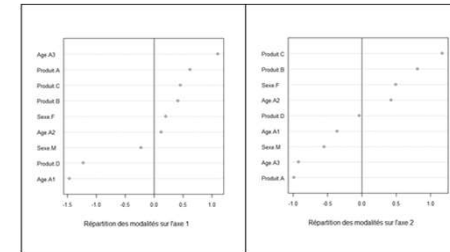


FIGURE 5C. PrefConsom : répartition des modalités des variables sur l'axe 1 et 2.

122

## 5. L'AFCM sous R : PrefConsom

```
Autre représentation : les variables (Fig.5D)
> plot(Pref.acm$co[,1],Pref.acm$co[,2],type="n",xlab="Axe 1",ylab="Axe 2",
xlim=c(-1.4,1.4))
> text(Pref.acm$co[,1], Pref.acm$co[,2], label= colnames(disj))
> title("Préférences consommateurs - plan des variables")
> abline(h=0,v=0)
Autre représentation : les individus (Fig.5D)
> plot(Pref.acm$li[,1],Pref.acm$li[,2],type="n",xlab="Axe 1",ylab="Axe 2",
xlim=c(-1.4,1.4))
> text(Pref.acm$li[,1], Pref.acm$li[,2], label=row.names(disj))
> title("Préférences consommateurs - Plan des individus") ; abline(h=0,v=0)
```

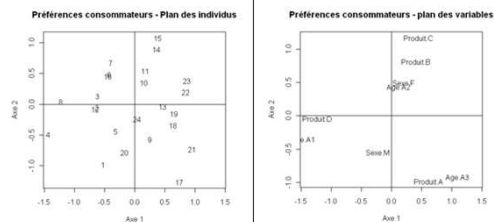


FIGURE 5D. PrefConsom : représentation des observations et des variables.

123

## 5. L'AFCM sous R : PrefConsom

```
Contributions (absolues) des modalités des variables
à la construction de chaque axe
> inertia.dudi(Pref.acm,col.inertia = T)$col.abs
```

|           | Comp1       | Comp2 | Comp3 | Comp4 | Comp5 | Comp6 |
|-----------|-------------|-------|-------|-------|-------|-------|
| Produit.A | 968         | 2632  | 114   | 0     | 604   | 2282  |
| Produit.B | 260         | 1088  | 1858  | 3436  | 1274  | 4     |
| Produit.C | 227         | 1666  | 1971  | 2671  | 1942  | 2     |
| Produit.D | <b>3280</b> | 4     | 50    | 131   | 1220  | 2315  |
| Sexe.F    | 158         | 1029  | 1406  | 887   | 936   | 344   |
| Sexe.M    | 174         | 1133  | 1547  | 976   | 1030  | 379   |
| Age.A1    | <b>3105</b> | 210   | 672   | 317   | 1536  | 2161  |
| Age.A2    | 63          | 879   | 1208  | 733   | 1113  | 4     |
| Age.A3    | 1765        | 1360  | 1175  | 847   | 345   | 2507  |

```
Contributions (absolues) des individus à la construction de chaque axe
> inertia.dudi(Pref.acm,row.inertia = T)$row.abs
```

|     | Axis1 | Axis2 | Axis3 | Axis4 | Axis5 | Axis6 |
|-----|-------|-------|-------|-------|-------|-------|
| 1   | 166   | 638   | 8     | 1     | 327   | 1133  |
| 2   | 68    | 0     | 68    | 151   | 498   | 66    |
| ... |       |       |       |       |       |       |
| 23  | 164   | 35    | 2     | 600   | 174   | 132   |
| 24  | 0     | 17    | 222   | 204   | 69    | 593   |

124

## 5. L'AFCM sous R : PrefConsom

**Interprétation :** fig 5 et code avant

- Le 1<sup>er</sup> facteur : préférence du produit D par la classe d'âge des moins de 20 ans, sans aucune différence entre hommes et femmes.
- Le 2<sup>nd</sup> facteur : opposition entre
  - le produit A, préféré par la classe d'âge de plus de 60 ans et par les hommes, et
  - les deux autres B et C préférés par la classe d'âge entre 20 et 60 ans et les femmes.

125

## 5. L'AFCM : exemple sous R

- Library sous R permettant de réaliser une AFCM :

`dudi.acm(ade4)` ou

`MCA(FactoMineR)` ou

`mca(MASS)`

126

## 5. L'AFCM sous R : PrefConsom

**Exercice :**

Effectuer AFCM à l'aide de la fonction `MCA` de la library `FactoMineR`

127

## Ce qu'il faut retenir

- L'AFC est une méthode factorielle du même type que l'ACP :
  - permet de décrire et de synthétiser sous forme de cartes l'information contenue dans un tableau de contingence.
- Les principales  $\neq$  entre AFC et ACP, outre la nature des données traitées, sont :
  - la métrique utilisée en AFC pour définir la proximité entre 2 lignes ou 2 colonnes est la métrique du  $\chi^2$  ;
  - la possibilité d'obtenir une **représentation superposée des lignes et des colonnes en AFC** (à utiliser avec précaution) ;
  - 2 graphes indépendants en ACP ; le biplot permet toutefois de s'en affranchir.
- L'AFCM est une extension du domaine d'application de l'AFC à l'étude de plus de 2 variables qualitatives. Les procédures de calcul et règles d'interprétations lui sont spécifiques. Elle permet d'obtenir :
  - une typologie des individus ;
  - une typologie des modalités ;
  - un résumé de l'ensemble des variables par un petit nombre de variables synthétiques, comme en ACP et AFC.

128

## Exercices Td/TP ch III: AFC « à la main », fonction sous R

### Données supplémentaires

| Données                                                    | Description                                                                  |
|------------------------------------------------------------|------------------------------------------------------------------------------|
| <a href="#">Heberg.txt</a>                                 | catégories socio-professionnelles avec des modes de résidence en vacances    |
| <a href="#">Mariage.txt</a>                                | mariages entre catégories professionnelles                                   |
| <a href="#">Cuisine.txt</a>                                | Corpus de 29 recettes décrites par 13 grands cuisiniers                      |
| <a href="#">Bledur.txt</a> ou <a href="#">BledurCD.txt</a> |                                                                              |
| <a href="#">PrefConsom.txt</a>                             | Préférences de consommateurs pour 4 produits en fonction de l'âge et du sexe |

129

## Références logiciels

- **Logiciel R :**
  - R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- **Bibliothèque [ade4](#) :**
  - Jean Thioulouse, Anne-Beatrice Dufour and Daniel Chessel (2004). [ade4](#): Analysis of Environmental Data : Exploratory and Euclidean methods in Environmental sciences. R package version 1.3-3.
  - Page web : <http://pbil.univ-lyon1.fr/ADE-4>
  - Mailing list: <http://pbil.univ-lyon1.fr/ADE-4/adelist.html>
- **Bibliothèque [FactoMineR](#)**
  - développée par F. Husson\*, J. Josse\*, S. Lê\*, d'Agrocampus Rennes, et J. Mazet.
  - Page web : [http://factominer.free.fr/index\\_fr.html](http://factominer.free.fr/index_fr.html)

130

## Références bibliographiques

- L. Bellanger, R. Tomassone, *Exploration de données et méthodes statistiques : Data analysis & Data mining avec R. Collection Références Sciences*, Editions Ellipses, Paris, 2014.
- A. Bouchier, Documents et supports de cours disponibles sur le site : <http://rstat.ouvaton.org/>
- J.-M. Bouroche & G. Saporta, *L'analyse des données*. Presses Universitaires de France : Que sais-je ? 85, Paris, 1992.
- J.-F. Durand, support intitulé « Elts de Calcul matriciel et d'Analyse Factorielle de Données » disponible sur le site: [www.math.univ-montp2.fr/~durand](http://www.math.univ-montp2.fr/~durand)
- F. Husson, S. Lê & J. Pagès, *Analyse de données avec R*. PUR, Rennes, 2009.
- L. Lebart, A. Morineau, M. Piron, *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 2006.
- G. Saporta, *Probabilités, Analyse des données*. Editions Technip, Paris, 2006.
- Statistics with R : [http://zoonek2.free.fr/UNIX/48\\_R/all.html](http://zoonek2.free.fr/UNIX/48_R/all.html)

131