Analyse de données

L. BELLANGER

Master 1 Ingénierie Statistique
Dpt de Mathématiques - Université de Nantes

Plan

0. 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

Plan ch. III

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

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 $\mathbf{N}$ à travers des représentations graphiques
  - ….

Origine : (Guttman, 1940 – Benzécri, 1964)

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.

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 $\mathbf{N}$ 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.

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, …

Brève introduction …

Types de tableaux traités par l'*AFC*

- **Tableau de contingence $\mathbf{N}$ 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*.

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* !

1. Notions de base

- **Tableau de contingence** $\mathbf{N}_{I\times J} = \lfloor n_{ij} \rfloor \,;\, i = 1,\cdots,I \,;\, j = 1,\cdots,J$

		Variable 2						
		Modalité 1	Modalité 2	…	Modalité j	…	Modalité J	Effectif marginal
Variable 1	Modalité 1	n_{11}	n_{12}		n_{1j}		n_{1J}	n_{1+}
	Modalité 2	n_{21}	n_{22}				n_{2J}	n_{2+}
	⋮							
	Modalité i	n_{i1}	n_{i2}		n_{ij}		n_{iJ}	n_{i+}
	⋮							
	Modalité I	n_{I1}	n_{I2}		n_{Ij}		n_{IJ}	n_{I+}
	Effectif marginal	n_{+1}	n_{+2}		n_{+j}		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.

1. Notions de base

- **Tableau de contingence**

On appelle:

- **Tableau des fréquences associé à $\mathbf{N}$** : $\mathbf{P} = \mathbf{N}/n_{++}$ où $n_{++} = \sum_{i,j} n_{ij}$, la matrice $I \times J$, d'élt courant $f_{ij} = \frac{n_{ij}}{n_{++}}$.
  - On a : $\sum_{i,j} f_{ij} = 1$.
- Les **effectifs marginaux** (resp. fréquences marginales) du tableau $\mathbf{N}$ (resp.$\mathbf{P}$), 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 $\mathbf{N}$)
  - 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 $\mathbf{N}$)
- $\mathbf{D_I} = \mathbf{diag}\{n_{1+}, n_{2+}, \cdots, n_{I+}\}$ et $\mathbf{D_J} = \mathbf{diag}\{n_{+1}, n_{+2}, \cdots, n_{+J}\}$, les deux matrices diagonales dont les éléments sont les effectifs marginaux des 2 variables du tableau $\mathbf{N}$.
- $\widetilde{\mathbf{D}}_\mathbf{I} = \mathbf{diag}\{f_{1+}, f_{2+}, \cdots, f_{I+}\}$ et $\widetilde{\mathbf{D}}_\mathbf{J} = \mathbf{diag}\{f_{+1}, f_{+2}, \cdots, f_{+J}\}$, les deux matrices diagonales dont les éléments sont les fréquences marginales des 2 variables du tableau $\mathbf{P}$.

1. Notions de base

- **Tableau de contingence**

$\Longrightarrow$ **Analyse du tableau de contingence :**

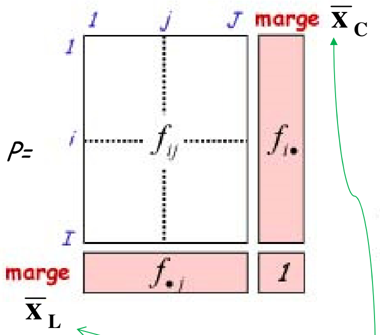
- *Ce ne sont pas les effectifs bruts qui nous intéressent ; mais les répartitions en % à l'intérieur d'une ligne ou d'une colonne.*
- On parle de **profils-lignes** et de **profils-colonnes** (freq. Cond.) :
  - Un profil ligne (resp. colonne) est déterminé en divisant la ligne (colonne) par le total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

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

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

1. Notions de bases

Transformation → Tableau des fréquences relatives
définit une mesure de probabilité



$$f_{ij} = n_{ij} / n$$

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

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

$$\sum_i f_{i\bullet} = \sum_j f_{\bullet j} = \sum f_{ij} = 1$$

$f_{i\bullet}$ = profil-colonne moyen ou marge ligne

$f_{\bullet j}$ = profil-ligne moyen ou marge colonne

Notation « • » = « + »

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 :

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

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

$$\bar{\mathbf{x}}_C = \mathbf{P}\mathbf{1_J} = \begin{bmatrix} f_{1+} \\ \vdots \\ f_{I+} \end{bmatrix}$$

Notation : $n = n_{++}$

1. Notions de base

- *Indépendance de 2 variables qualitatives*

Il y a **indépendance** entre les deux variables si $\boxed{f_{ij} = f_{i\bullet} f_{\bullet j}}$

⟺

toutes les **lignes** sont proportionnelles $\dfrac{f_{ij}}{f_{i\bullet}} = f_{\bullet j}$

toutes les **colonnes** sont proportionnelles $\dfrac{f_{ij}}{f_{\bullet j}} = f_{i\bullet}$

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

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 $t_{ij} = \dfrac{n_{i+} \times n_{+j}}{n_{++}}$, on définit :

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

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:

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

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.

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 $n = (3-1) \times (3-1) = 4$.
- 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.

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

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. $d_{ij} = \left(t_{ij} - n_{ij}\right)^2 / t_{ij}$ - 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* $\Longrightarrow$ supposer $\exists$ une liaison entre les 2 variables étudiées.
- Sous R : `chisq.test (stats)`

A faire ex. Td page 2

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 $i \in \{1, \ldots, I\}$ fixé :

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

- **Tableau $I \times J$ des profils-lignes**, le **tableau des fréquences conditionnelles** f_{ij}/f_{i+}) par $\mathbf{L} = \mathbf{D_I^{-1}N} \in \mathcal{M}_{I\times J}$:
  - 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* $\mathbf{D_I}/n_{++} = \widetilde{\mathbf{D}}_\mathbf{I}$) ;
  - le **centre de gravité** (**profil-ligne moyen**) de ce nuage est :

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

  - *C'est le profil marginal des lignes.*

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

2. Principe de l'*AFC*

- **Analyse d'un tableau de contingence : pondération**

On appelle :

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

$$\mathbf{C}^j = x^j = [n_{1j}/n_{+j}, \cdots, n_{Ij}/n_{+j}]^{\mathrm{T}} = [f_{1j}/f_{+j}, \cdots, f_{Ij}/f_{+j}]^{\mathrm{T}}; j = 1, \cdots, J.$$

- **Tableau $I \times J$ des profils-colonnes**, le **tableau des fréquences conditionnelles** (f_{ij}/f_{+j}) par $\mathbf{C} = \mathbf{ND_J^{-1}} \in \mathcal{M}_{I\times J}$:
  - 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* $\mathbf{D_J}/n_{++} = \widetilde{\mathbf{D}}_J$) ;
  - le **centre de gravité** (**profil-colonne moyen**) de ce nuage est :

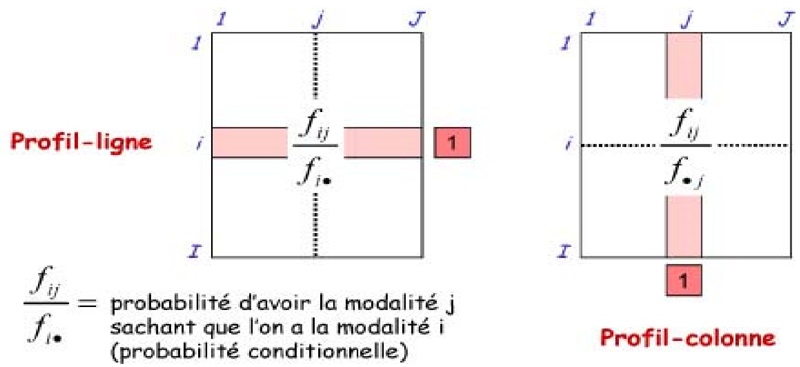
$$\bar{x}_C = \frac{1}{n_{++}}(\mathbf{ND_J^{-1}})\mathbf{D_J}\mathbf{1}_J = [f_{1+}, \cdots, f_{I+}]^{\mathrm{T}} \in \mathbb{R}^I.$$

  - *C'est le profil marginal des colonnes.*

2. Principe de l'*AFC*

Principe

Transformations des données en profils



$\frac{f_{ij}}{f_{i\bullet}}$ = probabilité d'avoir la modalité j sachant que l'on a la modalité i (probabilité conditionnelle)

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

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 :

$$\boldsymbol{L}_i = \boldsymbol{x}_i = \lfloor f_{i1}/f_{i+}, \cdots, f_{iJ}/f_{i+} \rfloor = \lfloor f_{+1}, \cdots, f_{+J} \rfloor = \overline{\boldsymbol{x}}_L \ ; \ \forall i$$

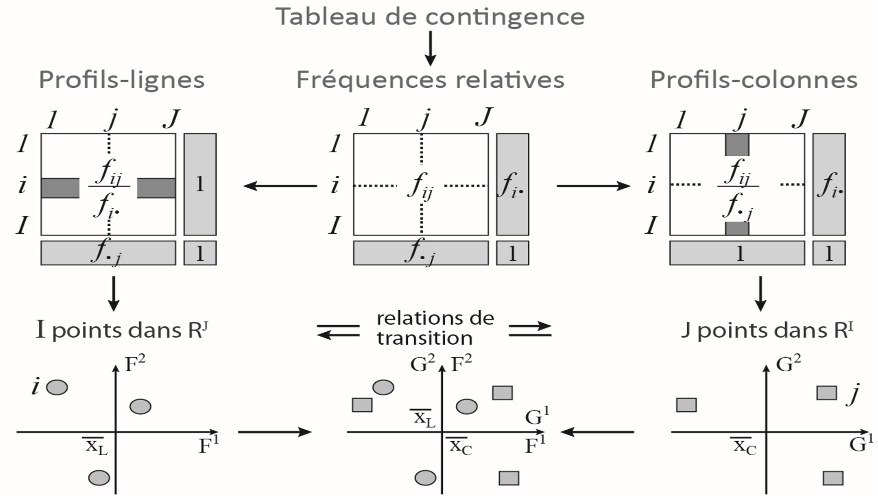
et

$$\mathbf{C}^j = \boldsymbol{x}^j = \lfloor f_{1j}/f_{+j}, \cdots, f_{Ij}/f_{+j} \rfloor^{\mathrm{T}} = \lfloor f_{1+}, \cdots, f_{I+} \rfloor^{\mathrm{T}} = \overline{\boldsymbol{x}}_C \ ; \ \forall j$$

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

2. Principe de l'*AFC*

Schéma général



ACP = étude de la dispersion du nuage des I profils dans $\mathbb{R}^J$ autour de $\overline{x}_L$.

Représentation simultanée

ACP = étude de la dispersion du nuage des J profils dans $\mathbb{R}^I$ autour de $\overline{x}_C$.

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) :

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

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

$\Longrightarrow$ 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.

2. Principe de l'*AFC*

- **Caractéristiques des nuages**
  - ***Nuage des I profils-lignes N_J :*** $\left(\mathbf{L} = \tilde{\mathbf{D}}_\mathbf{I}^{-1}\mathbf{P}, \mathbf{Q}_{\mathbf{J}*\mathbf{J}} = \tilde{\mathbf{D}}_\mathbf{J}^{-1}\right)$

  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 :**

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

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

2. Principe de l'*AFC*

- **Caractéristiques des nuages**

Nuage des profils-lignes : $\left(\mathbf{L} = \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}\mathbf{P}, \mathbf{Q}_{\mathbf{J*J}} = \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}\right)$

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

La **matrice des poids « $\mathbf{D}$ »** est $\tilde{\mathbf{D}}_{\mathbf{I}}$

- On définit le **nuage pondéré** des profils-lignes N_J

Nécessaire pour le calcul du centre de gravité du nuage $\overline{\mathbf{x}}_{\mathbf{L}}$ 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 :

$$\left(\mathbf{L} = \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}\mathbf{P}, \mathbf{Q} = \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}, \mathbf{D} = \tilde{\mathbf{D}}_{\mathbf{I}}\right)$$

2. Principe de l'*AFC*

- **Caractéristiques des nuages**
  - *Nuage des profils-colonnes* N_I : $\left(\mathbf{C}^{\mathrm{T}} = \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}\mathbf{P}^{\mathrm{T}}, \mathbf{Q}_{\mathbf{I*I}} = n_{++}\mathbf{D}_{\mathbf{I}}^{-1} = \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}\right)$

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**

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

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

2. Principe de l'*AFC*

- **Caractéristiques des nuages**

Nuage des profils-colonnes : $\left(\mathbf{C}^{\mathrm{T}} = \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}\mathbf{P}^{\mathrm{T}}, \mathbf{Q}_{\mathbf{I*I}} = \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}\right)$

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

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

- On définit le **nuage pondéré** des profils-colonnes N_I

Nécessaire pour le calcul du centre de gravité du nuage $\overline{\mathbf{x}}_{\mathbf{C}}$ et de l'inertie $I(N_I)$

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

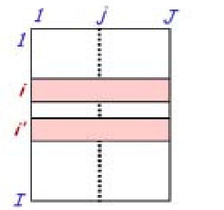
$$\left(\mathbf{C}^{\mathrm{T}} = \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}\mathbf{P}^{\mathrm{T}}, \mathbf{Q} = \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}, \mathbf{D} = \tilde{\mathbf{D}}_{\mathbf{J}}\right)$$

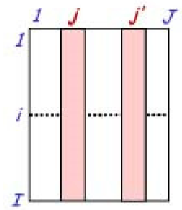
2. Principe de l'*AFC*

- **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**





distance entre 2 profils-lignes

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

distance entre 2 profils-colonnes

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

2. Principe de l'*AFC*

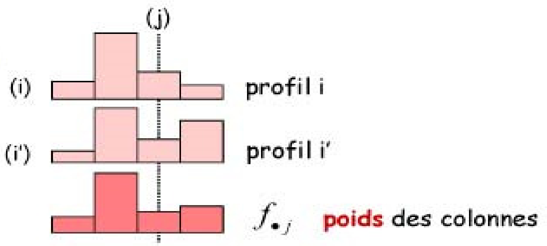
Principe — Ressemblance entre profils

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

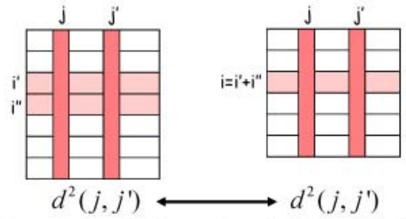
$\frac{1}{f_{i\bullet}}$ équilibre l'influence des lignes sur la distance entre les colonnes

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 $\mathbf{N}$ 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

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 $\overline{\mathbf{x}}_L$
  - 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 $\overline{\mathbf{x}}_C$

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 :

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

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

  - 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.

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

- **Conséquences :** on peut aussi voir l'*AFC* comme une DVS du triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ et $I = tr(\mathbf{X^T DXQ}) = tr(\mathbf{S}_I)$

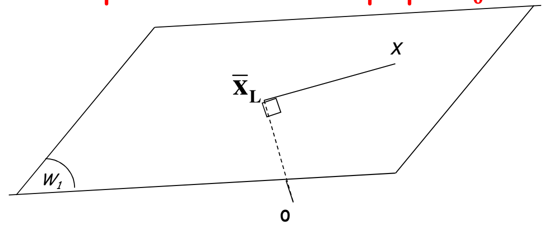
2. Principe de l'*AFC*

ACP non centrée et facteur trivial

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

- pour les lignes $\mathbf{S}_L = \mathbf{X^T DXQ} - \mathbf{\bar{x}_L \bar{x}_L^T}$ et
- pour les colonnes $\mathbf{S}_C = \mathbf{X^T DXQ} - \mathbf{\bar{x}}_C \mathbf{\bar{x}}_C^{\mathbf{T}}$

- En fait ce centrage n'est pas utile, car ∃ 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 1ère valeur propre $\lambda_0 = 1$.**



2. Principe de l'*AFC*

ACP non centrée et facteur trivial

Centrage ou non centrage du nuage ?

DVS de $(L_0 = D_I^{-1}P - I_I\,{}^tF_J, Q = D_J^{-1}, D = D_I)$

avec ${}^tI_I = [1 \ldots 1]$ et ${}^tF_J = [p_{.1} \ldots p_{.J}]$ (centre de gravité du nuage)

Matrice d'inertie : $({}^tP D_I^{-1} - F_J\,{}^tI_I) D_I (D_I^{-1}P - I_I\,{}^tF_J) D_J^{-1}$

$= {}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_I D_I D_I^{-1} P D_J^{-1} - {}^tP D_I^{-1} D_I I_I\,{}^tF_J D_J^{-1} + F_J\,{}^tI_I D_I I_I\,{}^tF_J D_J^{-1}$

$= {}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_I P D_J^{-1} - {}^tP I_I\,{}^tF_J D_J^{-1} + F_J\,{}^tI_I D_I I_I\,{}^tF_J D_J^{-1}$

$= {}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_I P D_J^{-1} - {}^tP I_I\,{}^tF_J D_J^{-1} + F_J\,{}^tI_I D_I I_I\,{}^tF_J D_J^{-1}$

$= {}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_J - F_J\,{}^tI_J + F_J\,{}^tI_J$

$= {}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_J$

DVS: $({}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_J)\, U_{L0} = U_{L0} \Lambda_{L0}$ avec ${}^tU_{L0} D_J^{-1} U_{L0} = I$

On remarque que F_J est vecteur propre de $({}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_J)$ associé à la valeur propre 0 :

$${}^tP D_I^{-1} P D_J^{-1} F_J - F_J\,{}^tI_J F_J = F_J - F_J$$

On remarque que pour tout vecteur propre u de ${}^tP D_I^{-1} P D_J^{-1}$, *u est vecteur propre de* ${}^tP D_I^{-1} P D_J^{-1} - F_J\,{}^tI_J$ *associé à la même valeur propre car* ${}^tF_J\ D_J^{-1}\ u = 0$ *(U_L matrice de vecteurs propre D_J^{-1} orthonormé) donc* $F_J\,{}^tI_J\,u = [0]_{J \times J}$

Conclusion : La DVS de $(L_0 = D_I^{-1}P - I_I\,{}^tF_J, Q = D_J^{-1}, D = D_I)$ conduit aux mêmes valeurs propres à l'exception de 1 et aux même 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.

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	$\mathbf{D_J^{-1}N^T D_I^{-1}N}$	$\boldsymbol{F_L}$ de $\mathbf{D}_I^{-1}\mathbf{N}\mathbf{D}_J^{-1}\mathbf{N}^T = \mathbf{LC^T}$	$\boldsymbol{F}_L^T \frac{\boldsymbol{D}_I}{n_{++}} \boldsymbol{F}_L = \lambda$
Colonnes	$\mathbf{D}_I^{-1}\mathbf{N}\mathbf{D}_J^{-1}\mathbf{N}^T$	$\boldsymbol{F_C}$ de $\mathbf{D}_J^{-1}\mathbf{N}^T\mathbf{D}_I^{-1}\mathbf{N} = \mathbf{C^T L}$	$\boldsymbol{F}_C^T \frac{\boldsymbol{D}_J}{n_{++}} \boldsymbol{F}_C = \lambda$

- 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.

2. Principe de l'*AFC*

- **Compléments sur l'ajustement des nuages en *AFC***

$D_I \to \tilde{D}_I$
$D_J \to \tilde{D}_J$

COMPLEMENTS SUR L' AJUSTEMENT DES NUAGES EN AFC

(X, Q, D)

DVS de $(L = D_I^{-1}P, Q = D_J^{-1}, D = D_I)$.

Matrice d'inertie $S_I = X'DXQ$

$${}^tP D_I^{-1} D_I D_I^{-1} P D_J^{-1} = {}^tP D_I^{-1} P D_J^{-1}$$
$$D_I^{-1} P D_J^{-1}\,{}^tP D_I^{-1} D_I = D_I^{-1} P D_J^{-1}\,{}^tP$$

DVS :

$X'DXQ\,U = U\Lambda$ avec $U'QU = I_n$

$${}^tP D_I^{-1} P D_J^{-1} U_L = U_L \Lambda_L \text{ avec } {}^tU_L D_J^{-1} U_L = I$$
$$D_I^{-1} P D_J^{-1}\,{}^tP V_L = V_L \Lambda_L \text{ avec } {}^tV_L D_I V_L = I$$

$XQX'DV = V\Lambda$ avec $V'DV = I$

$$L = V_L \Lambda_L^{1/2}\,{}^tU_L$$ → reconstitution $X = V\Lambda U'$

Transition

$$V_L = D_I^{-1} P D_J^{-1} U_L \Lambda_L^{-1/2}$$ ie $V = XQU\Lambda^{-1/2}$

Composantes principales

$$F_L = L D_J^{-1} U_L = D_I^{-1} P D_J^{-1} U_L$$ project° des lignes suivant les axes $= XQU$

DVS de $(C = D_J^{-1}\,{}^tP, Q = D_I^{-1}, D = D_J)$.

Matrice d'inertie

$$P D_J^{-1} D_J D_J^{-1}\,{}^tP D_I^{-1} = P D_J^{-1}\,{}^tP D_I^{-1}$$
$$D_J^{-1}\,{}^tP D_I^{-1} P D_J^{-1} D_J = D_J^{-1}\,{}^tP D_I^{-1} P$$

DVS :

$$P D_J^{-1}\,{}^tP D_I^{-1} U_C = U_C \Lambda_C \text{ avec } {}^tU_C D_I^{-1} U_C = I$$
$$D_J^{-1}\,{}^tP D_I^{-1} P\, V_C = V_C \Lambda_C \text{ avec } {}^tV_C D_J V_C = I$$

$$C = V_C \Lambda_C^{1/2}\,{}^tU_C$$

Transition

$$V_C = D_J^{-1}\,{}^tP D_I^{-1} U_C \Lambda_C^{-1/2}$$

Composante principale

$$F_C = D_J^{-1}\,{}^tP D_I^{-1} U_C$$ project° des colonnes suivant les axes

2. Principe de l'*AFC*

- **Compléments sur l'ajustement des nuages en *AFC***

$= XQU$

Lien entre les deux DVS

$$P D_J^{-1}\,{}^tP D_I^{-1} U_C = U_C \Lambda_C \text{ avec } {}^tU_C D_I^{-1} U_C = I \;\Rightarrow\; D_I^{-1} P D_J^{-1}\,{}^tP D_I^{-1} U_C = D_I^{-1} U_C \Lambda_C \text{ avec } {}^tU_C D_I^{-1} D_I D_I^{-1} U_C = I$$

$$\Rightarrow \quad V_L = D_I^{-1} U_C \quad \text{et} \quad V_C = D_J^{-1} U_L \quad \text{et} \quad \Lambda_C = \Lambda_L$$ mêmes valeurs propres

On en déduit les formules de transition

$$U_L = D_J V_C = D_J D_J^{-1}\,{}^tP D_I^{-1} U_C \Lambda^{-1/2} = {}^tP D_I^{-1} U_C \Lambda^{-1/2}$$

$$F_L = D_I^{-1} P D_J^{-1} U_L = D_I^{-1} P D_J^{-1}\,{}^tP D_I^{-1} U_C \Lambda^{-1/2} \text{ avec } F_C = D_J^{-1}\,{}^tP D_I^{-1} U_C \;\Rightarrow\; F_L = D_I^{-1} P F_C \Lambda_C^{-1/2}$$ et par symétrie $F_C = D_J^{-1}\,{}^tP F_L \Lambda^{-1/2}$

> **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.

2. Principe de l'*AFC*

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

$$\left(\mathbf{L} = \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}\mathbf{P}, \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}, \tilde{\mathbf{D}}_{\mathbf{I}}\right) \text{et} \left(\mathbf{C}^T = \tilde{\mathbf{D}}_{\mathbf{J}}^{-1}\mathbf{P}^{\mathrm{T}}, \tilde{\mathbf{D}}_{\mathbf{I}}^{-1}, \tilde{\mathbf{D}}_{\mathbf{J}}\right)$$

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

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

$\Leftrightarrow$ *Au sens où les DVS ont les mêmes valeurs singulières – Voir ch ACP p. 28 lien DVS*

$$\left(\mathbf{Z} = \left[\frac{f_{ij}}{\sqrt{f_{i+}}\sqrt{f_{+j}}}\right], \mathbf{I}_{\mathbf{J}}, \mathbf{I}_{\mathbf{I}}\right)$$

On a alors :

- *Coordonnées des profils-lignes sur l'axe k :* $\mathbf{F}_X^k = \mathbf{L}\tilde{\mathbf{D}}_J^{-1/2}(\mathbf{U}_Z)_k = \mathbf{F}_L^k$

et en utilisant la formule de transition

- *Coordonnées des profils-colonnes sur l'axe k :* $\mathbf{G}_X^k = \sqrt{\lambda_k}\tilde{\mathbf{D}}_J^{-1/2}(\mathbf{U}_Z)_k = \mathbf{F}_C^k$

2. Principe de l'*AFC*

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

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

Le nuage définit 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 définit par ${}^tX$ possède les mêmes distances entre points que C compte tenu des métriques. Les distances entre deux lignes de ${}^tX$ sont les mêmes qu'entre deux lignes de C. → $(X', Q = D_I, D = D_J)$ pour les colonnes

Matrice d'inertie S_I

lignes $X'QXD$ → $D_J^{-1}\,{}^tP D_I^{-1} D_I D_I^{-1} P D_J^{-1} D_J = D_J^{-1}\,{}^tP D_I^{-1} P$

colonnes → $D_I^{-1} P D_J^{-1} D_J D_J^{-1}\,{}^tP D_I^{-1} D_I = D_I^{-1} P D_J^{-1}\,{}^tP$ ($X\ D\ X'Q$)

DVS :

$$D_J^{-1}\,{}^tP D_I^{-1} P\, U_X = U_X \Lambda_X \text{ avec } {}^tU_X D_J U_X = I \quad \Rightarrow \quad U_X = V_C \quad \Lambda = \Lambda_X$$
$$D_I^{-1} P D_J^{-1}\,{}^tP\, V_X = V_X \Lambda_X \text{ avec } {}^tV_X D_I V_X = I \quad \Rightarrow \quad V_X = V_L$$

$$X = V_X \Lambda_X^{1/2}\,{}^tU_X \quad \text{et} \quad {}^tX = U_X \Lambda_X^{1/2}\,{}^tV_X$$

Transition

$$V_X = D_I^{-1} P D_J^{-1} D_J U_X \Lambda^{-1/2} = D_I^{-1} P U_X \Lambda^{-1/2} \quad \text{et} \quad U_X = D_J^{-1}\,{}^tP D_I^{-1} D_I V_X \Lambda^{-1/2} = D_J^{-1}\,{}^tP V_X \Lambda^{-1/2}$$

Composantes principales

$$F_X = D_I^{-1} P D_J^{-1} D_J U_X = D_I^{-1} P U_X = D_I^{-1} P V_C = D_I^{-1} P D_J^{-1} U_L \Rightarrow F_L = F_X$$
$$G_X = D_J^{-1}\,{}^tP D_I^{-1} D_I V_X = D_J^{-1}\,{}^tP V_X = D_J^{-1}\,{}^tP V_L = D_J^{-1}\,{}^tP D_I^{-1} U_C \Rightarrow F_C = G_X$$

> **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 $(U_\alpha, U_{\alpha'})$ 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 :

$$I_g = \lambda_1 + \lambda_2 + \cdots + \lambda_K$$

  - 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 $\mathbf{N}$.

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 $\boldsymbol{F_X}^k$, on en déduit les K vecteurs coord. des pts-colonnes (et $\boldsymbol{G_X}^k$ réciproquement) :

$$\begin{cases} \mathbf{G}_X{}^k = \mathbf{D}_\mathbf{J}^{-1}\mathbf{N}^\mathbf{T}\mathbf{F}_X{}^k / \sqrt{\lambda_k} & \text{soit } \mathrm{G}_X^k(j) = \dfrac{1}{\sqrt{\lambda_k}} \displaystyle\sum_{i=1}^{i=I} \dfrac{n_{ij}}{n_{+j}} F_X^k(i) \\ \mathbf{F}_X{}^k = \mathbf{D}_\mathbf{I}^{-1}\mathbf{N}\mathbf{G}_X{}^k / \sqrt{\lambda_k} & \text{soit } \mathrm{F}_X^k(i) = \dfrac{1}{\sqrt{\lambda_k}} \displaystyle\sum_{j=1}^{j=J} \dfrac{n_{ij}}{n_{i+}} G_X{}^k(j) \end{cases}$$

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

3. Eléments principaux de l'*AFC*

- **Relations quasi-barcentriques**

Principe — **Représentation simultanée ligne-colonnes**

Au coefficient $\dfrac{1}{\sqrt{\lambda_s}}$ 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

3. Eléments principaux de l'*AFC*

- **Reconstitution des données**
  - Comme en *ACP*, on peut reconstituer les valeurs du tableau analysé par :

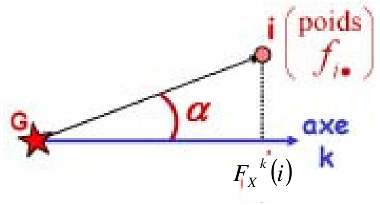
$$n_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}\left(1 + \sum_{k=1}^{K} F_X{}^k(i)\, G_X{}^k(j) / \sqrt{\lambda_k}\right)$$

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

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)

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



Même principe pour j.

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 :**

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

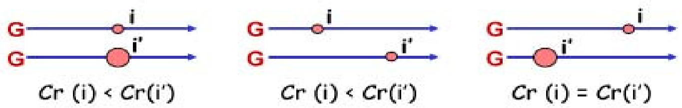
      - **Contribution du profil-colonne j à l'inertie de l'axe k :**

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

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 $ctr_k(j) > \frac{n_{+j}}{n_{++}}$)

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 × 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.**

4. Exemple sous R

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'*AFC* simples nous permettra de projeter sur un plan les structures d'un tableau de contingence.
- **Attention** : l'*AFC* étudie la structure des écarts à l'indépendance, pas leur intensité.

4. Exemple sous R

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

• Le tableau des données d'exemple : **housetasks(ade4)**. Il contient 13 « **taches 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.*

4. Exemple sous R

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

4. Exemple sous R

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*

4. Exemple sous R

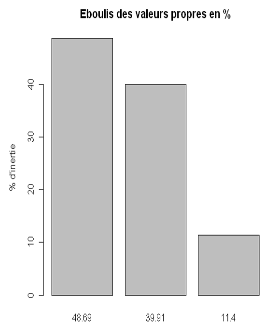
4. L'*AFC* - fonction `dudi.coa()` de la bibliothèque `ade4` :

- 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.arg=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,

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

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


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**

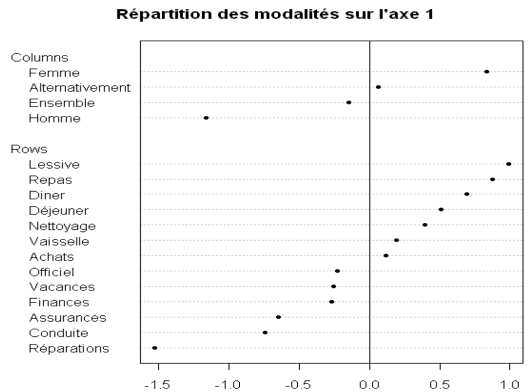
- Les *qlt* ne sont pas indiquées *; mais il faut les regarder avec attention avant de commenter !*

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


4. Exemple sous R

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

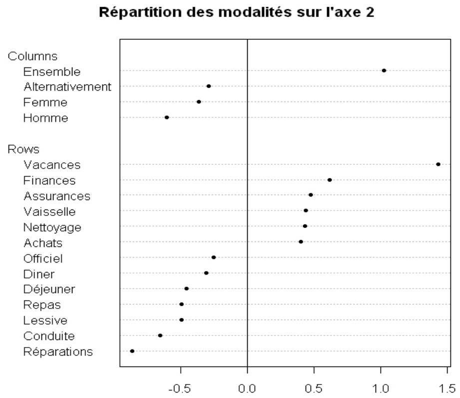
```
# Aide graphique à l'interprétation des axes
> score.coa (TacheMenage.afc,xax = 1,dotchart = TRUE) >
title("Répartition des modalités sur l'axe 1")
> abline(v=0)
```



4. Exemple sous R

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

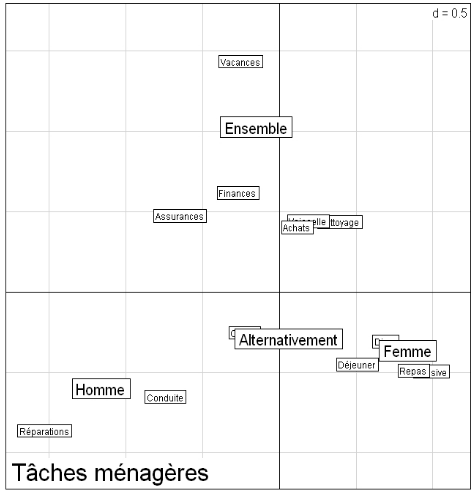
```
> score.coa (TacheMenage.afc,xax = 2,dotchart = TRUE)
> title("Répartition des modalités sur l'axe 2")
> abline(v=0)
```



4. Exemple sous R

9. Représentation du plan factoriel

```
> scatter.coa(TacheMenage.afc, method=1,sub="Tâches ménagères",posieig="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 !

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*.

4. Exemple sous R

11. Avec `CA(FactoMineR)`

```
library(FactoMineR)
CA(TacheMenage)
```

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

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** × **variables qualitatives**

Exemple : enquêtes socio-économiques

Individus		Variable 1	...	Variable j	...	Variable C
	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
	n	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**

5. L'*AFCM* : Notions de base

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

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

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

5. L'*AFCM* : Notions de base

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

$$
\underset{(n,C)}{\mathbf{R}} = \overbrace{\left|\begin{array}{c|c|c}
2 & 2 & 4\\
2 & 1 & 3\\
3 & 1 & 2\\
1 & 2 & 4\\
1 & 2 & 3\\
2 & 2 & 3\\
3 & 1 & 1\\
1 & 1 & 1\\
2 & 1 & 2\\
2 & 2 & 3\\
3 & 2 & 2\\
1 & 1 & 4
\end{array}\right|}^{C=3}
\Rightarrow
\underset{(n,J)}{\mathbf{Z}} = \overbrace{\left|\begin{array}{ccc|cc|cccc}
0&1&0&0&1&0&0&0&1\\
0&1&0&1&0&0&0&1&0\\
0&0&1&1&0&0&1&0&0\\
1&0&0&0&1&0&0&0&1\\
1&0&0&0&1&0&0&1&0\\
0&1&0&0&1&0&0&1&0\\
0&0&1&1&0&1&0&0&0\\
1&0&0&1&0&1&0&0&0\\
0&1&0&1&0&0&1&0&0\\
0&1&0&0&1&0&0&1&0\\
0&0&1&0&1&0&1&0&0\\
1&0&0&1&0&0&0&0&1
\end{array}\right|}^{J=9}
$$

(lignes 1 à n)

5. L'*AFCM* : Notions de base

- **codage disjonctif** (d'après Lebart *et al*. (2006)
  - *Le tableau* $\mathbf{Z}$ *est appelé* ***tableau disjonctif complet***, de terme général :
    - Selon que le sujet i a choisi la modalité j : $z_{ij} = 1$ ou $z_{ij} = 0$

	Variable 1: 1 … J_1	…	Variable k: 1 … k … J_k	…	Variable C: 1 … J_C	Marges en ligne
1			.			C
			.			C
i	0 1 0	…	z_{ik}	…	0 1	C
			.			C
n			.			C
Marges en colonnes	z_{+1} … z_{+J1}	…		…	z_{+J}	nC

Individus (lignes) ; *Effectif total du tableau Z* : nC

5. L'*AFCM* : Notions de base

- **Codage disjonctif** (d'après Lebart *et al*. (2006)
  - *Le tableau* $\mathbf{Z}$ *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* $\mathbf{Z}_j$, *l'effectif total est bien* n ;
    - *la somme des marges donne* ***l'effectif total*** *de* $\mathbf{Z}$: nC

5. L'*AFCM* : Notions de base

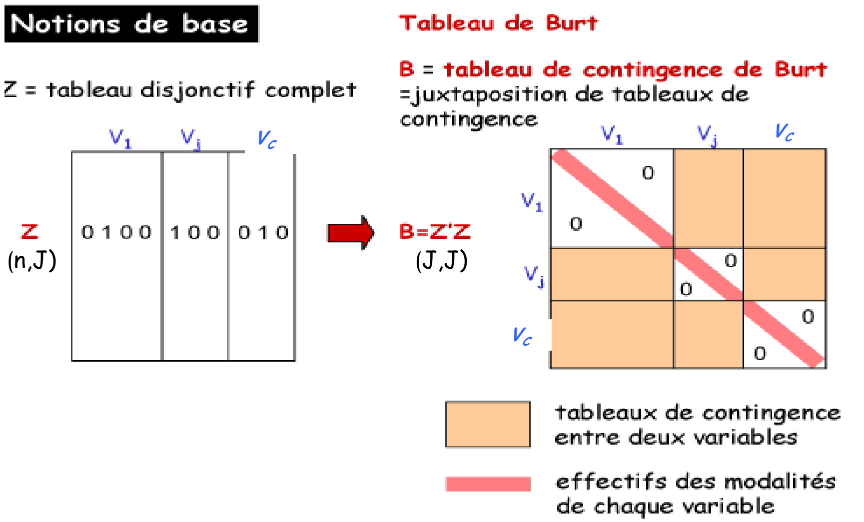
- ***Tableau de contingence de Burt***

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

$$\mathbf{N}_{I\times J} = \mathbf{Z}^{T}_{\mathbf{1}(n\times I)}\mathbf{Z}_{\mathbf{2}(n\times J)}$$

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

5. L'*AFCM* : Notions de base



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 $j \le J$:

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

  - **L'effectif total** de **B** vaut :

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

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):

$$d_{jj} = b_{jj} = \mathbf{Z}_{+j} \text{ et } d_{jj'} = 0 \; \forall j' \neq j.$$

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

$J = 9$

$$\underset{(J,J)}{\mathbf{B}} = \left|\begin{array}{ccc|cc|cccc}
4 & 0 & 0 & 2 & 2 & 1 & 0 & 1 & 2 \\
0 & 5 & 0 & 2 & 3 & 0 & 1 & 3 & 1 \\
0 & 0 & 3 & 2 & 1 & 1 & 2 & 0 & 0 \\
\hline
2 & 2 & 2 & 6 & 0 & 2 & 2 & 1 & 1 \\
2 & 3 & 1 & 0 & 6 & 0 & 1 & 3 & 2 \\
\hline
1 & 0 & 1 & 2 & 0 & 2 & 0 & 0 & 0 \\
0 & 1 & 2 & 2 & 1 & 0 & 3 & 0 & 0 \\
1 & 3 & 0 & 1 & 3 & 0 & 0 & 4 & 0 \\
2 & 1 & 0 & 1 & 2 & 0 & 0 & 0 & 3
\end{array}\right|
\qquad
\underset{(J,J)}{\mathbf{D}} = \left|\begin{array}{ccc|cc|cccc}
4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3
\end{array}\right|$$

5. L'*AFCM* : Principes de base

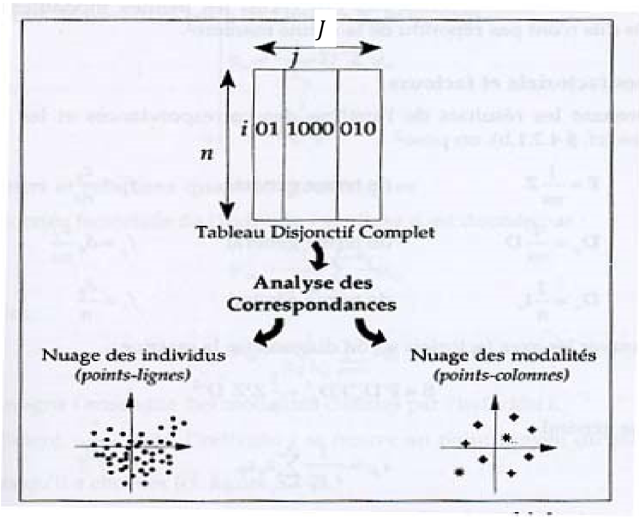
- La problématique de l'*AFCM* est :
  - apparentée à celle de l'*ACP* (tableau individus x variables) ; mais
  - peut être considérée comme une **généralisation** de l'*AFC* (liaisons entre plusieurs variables qualitatives).

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

- 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.

5. L'*AFCM*: Principes de base

Tiré de Lebart *et al*. (2006) p193



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 :

$$\mathbf{N}_{I\times J} = \mathbf{Z}^{\mathrm{T}}_{\mathbf{1}(n\times I)} \times \mathbf{Z}_{\mathbf{2}(n\times J)}$$

Soit $\mathbf{D}$ **matrice des effectifs marginaux** des modalités :

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_I & 0 \\ 0 & \mathbf{D}_J \end{bmatrix}$$

où

$\mathbf{D}_I$ matrice diagonale de dim I contenant les effectifs n_{i+}

$\mathbf{D}_J$ matrice diagonale de dim J contenant les effectifs n_{+j} ;

$\mathbf{D}$ matrice carrée d'ordre $I+J$.

On va voir que : l'AFC du tableau disjonctif $\mathbf{Z}$ est équivalente à l'AFC de $\mathbf{N}$

5. L'*AFCM* : Principes de base

- ***Cas de 2 variables***
  - Le **tableau des profils lignes** $\mathbf{L}$ de $\mathbf{Z}$ est le **tableau des fréquences conditionnelles** (n_{ij}/n_{i+}) i fixé.

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

    C'est le tableau : $\mathbf{ZD}^{-1} = [\mathbf{Z}_1 \quad | \mathbf{Z}_2] \begin{bmatrix} \mathbf{D}_I^{-1} & 0 \\ 0 & \mathbf{D}_J^{-1} \end{bmatrix}$
  - On sait que les *coordonnées des profils colonnes* de $\mathbf{Z}$ sont les vecteurs propres de $\mathbf{C}^{\mathrm{T}}\mathbf{L}$ (voit AFC p. 40) d'où :

    $\mathbf{C}^{\mathrm{T}}\mathbf{L} = (\mathbf{ZD}^{-1})^{\mathrm{T}}\mathbf{Z}/2 = \frac{1}{2}\mathbf{D}^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \frac{1}{2}\mathbf{D}^{-1}\mathbf{B}$; $\mathbf{B}$ tableau de Burt.

5. L'*AFCM* : Principes de base

- ***Cas de deux variables***
  - L'équation donnant les $I+J$ coordonnées des profils colonnes de $\mathbf{Z}$ est :

$$\frac{1}{2}\begin{bmatrix} \mathbf{D}_I^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_J^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{D}_I & \mathbf{N} \\ \mathbf{N}^T & \mathbf{D}_J \end{bmatrix}\begin{bmatrix} \boldsymbol{F}_L \\ \boldsymbol{F}_C \end{bmatrix} = \mu \begin{bmatrix} \boldsymbol{F}_L \\ \boldsymbol{F}_C \end{bmatrix}$$

En notant $\boldsymbol{F}_L$ les I premiers composantes et $\boldsymbol{F}_C$ les J suivantes. Soit :

$$\begin{bmatrix} I_I & \mathbf{D}_I^{-1} \\ \mathbf{D}_J^{-1} & I_J \end{bmatrix}\begin{bmatrix} \boldsymbol{F}_L \\ \boldsymbol{F}_C \end{bmatrix} = 2\mu \begin{bmatrix} \boldsymbol{F}_L \\ \boldsymbol{F}_C \end{bmatrix}$$

- D'où les équations :

$$\begin{cases} \boldsymbol{F}_L + \mathbf{D}_I^{-1}\mathbf{N}\boldsymbol{F}_C = 2\mu \boldsymbol{F}_L \\ \mathbf{D}_J^{-1}\mathbf{N}^{\mathrm{T}}\boldsymbol{F}_L + \boldsymbol{F}_C = 2\mu \boldsymbol{F}_C \end{cases} \text{ou ..} \begin{cases} \mathbf{D}_J^{-1}\mathbf{N}^{\mathrm{T}}\mathbf{D}_I^{-1}\mathbf{N}\boldsymbol{F}_C = (2\mu-1)^2\boldsymbol{F}_C \\ \mathbf{D}_I^{-1}\mathbf{N}\mathbf{D}_J^{-1}\mathbf{N}^{\mathrm{T}}\boldsymbol{F}_L = (2\mu-1)^2\boldsymbol{F}_L \end{cases}$$

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

5. L'*AFCM* : Principes de base

- ***Cas de deux variables***

Soit $\begin{cases} \mathbf{D}_J^{-1}\mathbf{N}^{\mathrm{T}}\mathbf{D}_I^{-1}\mathbf{N}\boldsymbol{F}_C = (2\mu-1)^2\boldsymbol{F}_C \\ \mathbf{D}_I^{-1}\mathbf{N}\mathbf{D}_J^{-1}\mathbf{N}^{\mathrm{T}}\boldsymbol{F}_L = (2\mu-1)^2\boldsymbol{F}_L \end{cases}$

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 $k^{ième}$ vp de l'*AFC* de $\mathbf{Z}$, elle est liée à celle de l'AFC précédente par :

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

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

5. L'*AFCM* : Principes de base

- ***Cas général :*** $C>2$ ***variables***

L' AFCM est l'AFC de $\mathbf{Z}$ *:*

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

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 : $\mathbf{Z}/C$

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

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

Où $\boldsymbol{F} = [\boldsymbol{F}_1 \quad \boldsymbol{F}_2 \quad \cdots \quad \boldsymbol{F}_C] \in \mathbb{R}^J$ le vecteur des coordonnées factorielles des *modalités des* C *variables*, il a J composantes.

5. L'*AFCM* : Principes de base

- ***Cas général :*** $C>2$ ***variables***

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

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

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

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

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

Relations quasi-barycentriques …

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 indiv.** 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**.

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

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 ou
  - La règle $\mu_k > 1/C$. $1/C$ *cons*titue 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)**.

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)**.

5. L'*AFCM* : interprétations

- ***Contributions***

Pour la matrice $\mathbf{Z}_c$ qui est formée de J_c modalités, la **coordonnée de la $j^{ème}$ modalité de la variable c** étant $F^k(c_j)$ sur le $k^{ème}$ 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 :

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

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

5. L'*AFCM*: interprétations

- *Contributions*

L'inertie totale de la variable c correspondant à $\mathbf{Z}_c$:

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

La **contribution** de $\mathbf{Z}_c$ **à l'inertie totale** est égale à :

$$ctr_{\mathbf{Z}_c} = \frac{I_{\mathbf{Z}_c}}{I_g} = \frac{J_c - 1}{J - C}$$

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

5. L'*AFCM* : 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^2
- **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)

5. L'*AFCM*: 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
…../………..
```


5. L'*AFCM* : : 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)
```

5. L'*AFCM* : 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)
```


5. L'*AFCM* : 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

………………………./……………………….

5. L'*AFCM* : 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.

5. L'*AFCM*: 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
```

5. L'*AFCM* : 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

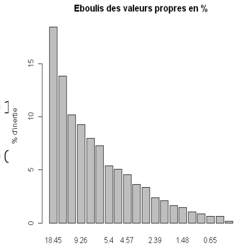
```
> inertie<-z$eig/sum(z$eig)*100
> barplot(inertie,ylab="% d'inertie",names.arg=round(inertie,2))
> title("Eboulis des valeurs propres en %")
```

  - 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 %

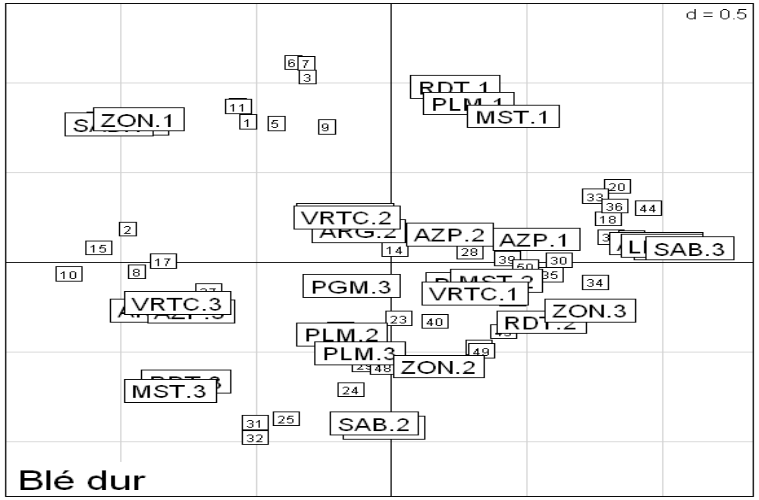
```
> round(z$eig/sum(z$eig)*100,2)
[1] 18.45 13.86 10.18  9.26  7.96  7.29  5.40  5.05
 4.57  3.65  3.38  2.39
[13]  2.09  1.64  1.48  1.06  0.85  0.65  0.63  0.19
```



5. L'*AFCM*: exemple sous R

- Une représentation graphique du plan factoriel :

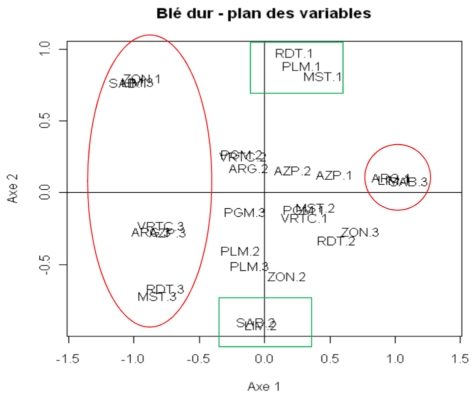
```
> scatter.coa(z, method = 1, sub = "Blé dur", posieig = "none")
```



5. L'*AFCM* : exemple sous R

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

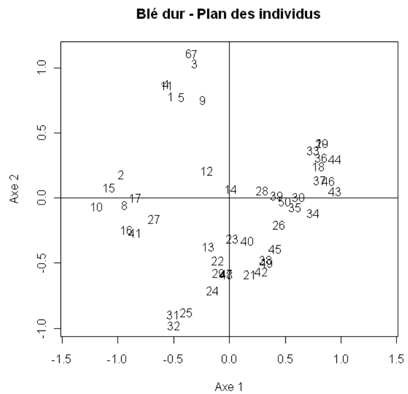
```
> plot(z$co[,1],z$co[,2],type="n",xlab="Axe 1",ylab="Axe 2",
   xlim=c(-1.4,1.4))
> text(z$co[,1], z$co[,2], label= colnames(disj))
> title("Blé dur – plan des variables")
> abline(h=0,v=0)
```



5. L'*AFCM* : 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)
```



5. L'AFCM : 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
……………… .			
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

Cf fig p 101
Axe 1: opposition sabl1, LIM1 & arg1 à …
Axe 2 : opposition : RDT.1, PLM.1, MST.1 à LIM.2, SAB.2

5. L'AFCM : exemple sous R

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

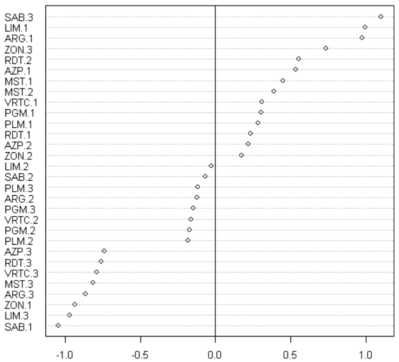
```
> inertia.dudi(z,row.inertia = T)$row.abs
```

	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

5. L'AFCM : 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)
```

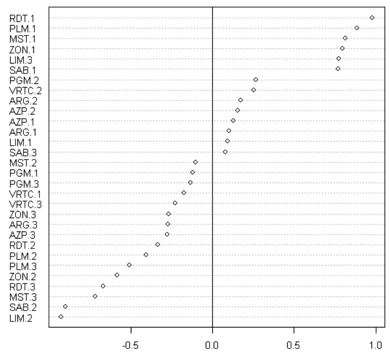


Répartition des modalités sur l'axe 1

5. L'AFCM : 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)
```



Répartition des modalités sur l'axe 2

5. L'*AFCM* : 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)
}
```


5. L'*AFCM* sous R : PrefConsom

- **PrefConsom:** résultats d'une enquête sur la préférence de consommateurs :
  - 3 variables ($\mathcal{C}$=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	Produit B	Produit C	Produit D
A1 < 20 ans	M	28	8	6	64
	F	12	20	6	56
A2 20-60 ans	M	120	50	40	80
	F	60	90	80	80
A3 > 60 ans	M	50	12	10	8
	F	70	28	10	12

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
```


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$Nb), scannf = FALSE, nf =6) #l'analyse factorielle
#  Eboulis des valeurs propres (Fig.5A)
> inertie<-Pref.acm$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])
```

5. L'*AFCM* sous R : PrefConsom

•

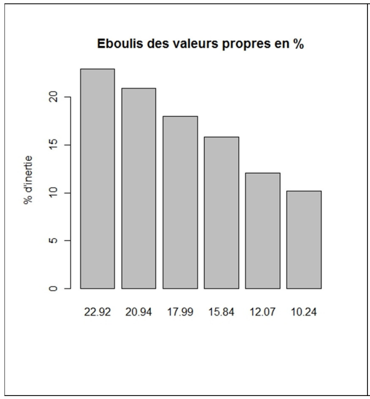


FIGURE 5A. **PrefConsom** :
éboulis des valeurs propres.

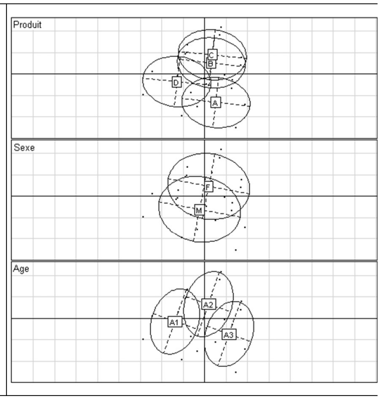


FIGURE 5B. **PrefConsom** :
nuages par modalité des facteurs.

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)
```

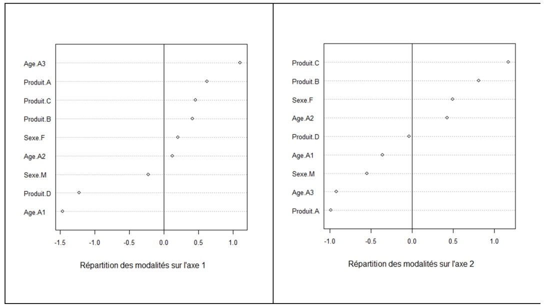


FIGURE 5 . **PrefConsom** *: répartition des modalités des variables sur l'axe 1 et 2.*

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)
```

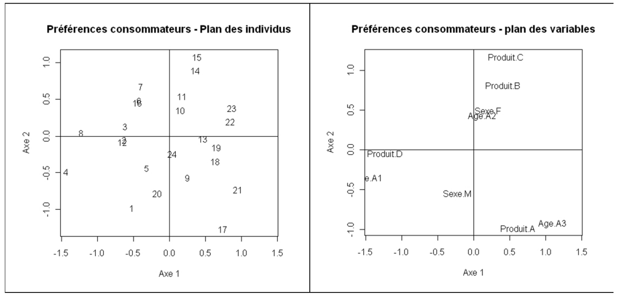


FIGURE 5D. **PrefConsom** *: représentation des observations et des variables.*

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	**3280**	4	50	131	1220	2315
Sexe.F	158	1029	1406	887	936	344
Sexe.M	174	1133	1547	976	1030	379
Age.A1	**3105**	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

5. L'*AFCM* sous R : PrefConsom

Interprétation : fig 5 et code avant

- Le 1er facteur : préférence du produit D par la classe d'âge des moins de 20 ans, *sans aucune différence entre hommes et femmes.*
- Le 2nd 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.

5. L'*AFCM* : exemple sous R

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

`dudi.acm(ade4)` `ou`

`MCA(FactoMineR)` `ou`

`mca(MASS)`

5. L'*AFCM* sous R : PrefConsom

Exercice :

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

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 ≠ 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** χ^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.*

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

Données supplémentaires

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

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.htmlactoMineR

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
- 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