Analyse Factorielle des Correspondances (AFC) et des Correspondances Multiples (AFCM)

Cours 3

Analyse Factorielle des Correspondances (AFC)

•**L'analyse des correspondances** (**AFC**) étudie les proximités entre individus décrits par deux ou plusieurs variables qualitatives ainsi que les proximités entre les modalités de ces variables.

Analyse Factorielle des Correspondances (AFC)

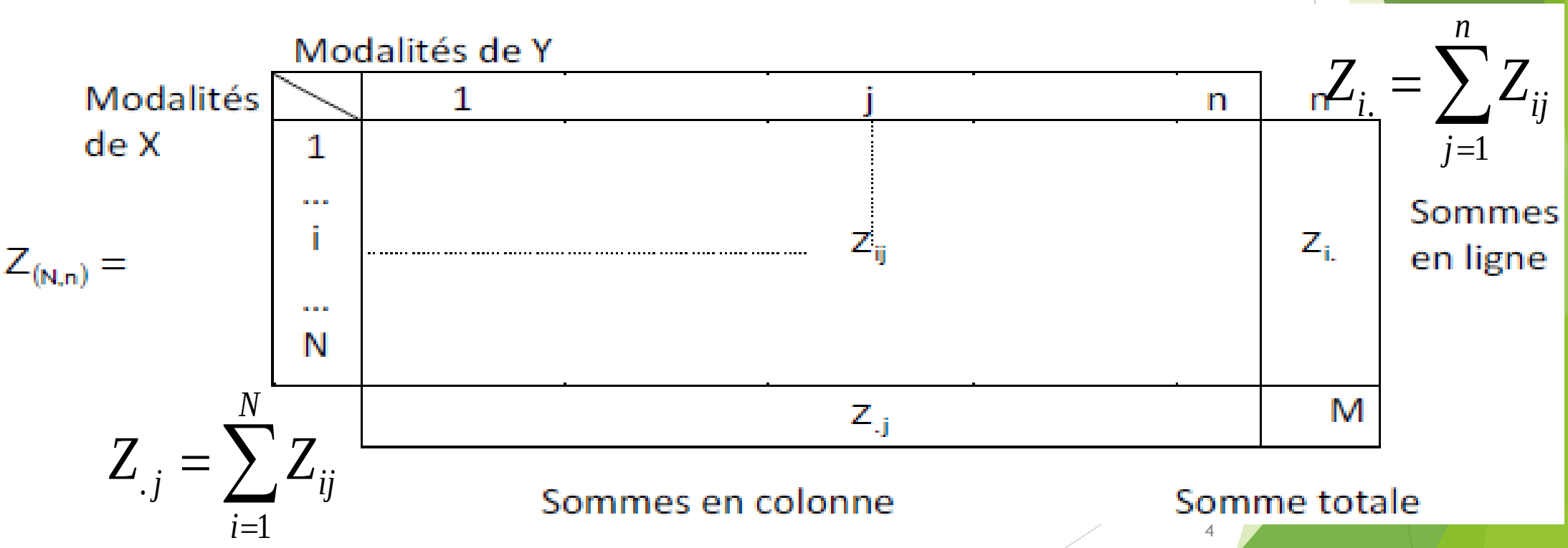
- ACP = étude des liaisons contenues dans un tableau individus*variables, lorsque toutes les variables sont **quantitatives**.
- AFC (Analyse Factorielle des Correspondances) et l'ACM (Analyse des Correspondances Multiples) = étude des liaisons contenues dans un tableau individus*variables, lorsque toutes les variables sont **qualitatives**.
- **L'AFC** est l'étude des correspondances entre les modalités de **deux variables** qualitatives.
- **L'ACM** est une généralisation de l'AFC à **plus de deux variables** qualitatives.

Analyse Factorielle des Correspondances (AFC)

L'AFC s'applique essentiellement à des tableaux de contingence. C'est un tableau d'effectifs qui contient à l'intersection de la ligne i et de la colonne j des zij individus.

Zij: nombre d'individu appartenant à l'élement i de modalité X et j de modalité Y

Zi. : marge ligne Z.j: marge colonne



EXEMPLE: Le tableau de contingence

OBSEVATION

Individu	bac	sexe
1	S	HOMME
2	TM	FEMME
3	TM	HOMME
4	L	FEMME
5	S	FEMME
6	TM	Femme

M=6

Z=

TABLEAU DE CONTINGENCE

MODALITE	HOMME	FEMME	Zi.
S	1	1	2
TM	1	2	3
L	0	1	1
Z.j	2	4	6

Tableau de contingence (exemple)

La fréquence(probabilité) associée au terme x_{ij} est:

$$P_{.j} = \frac{Z_{.j}}{M} \qquad P_{i.} = \frac{Z_{i.}}{M} \qquad P_{ij} = \frac{Z_{ij}}{M}$$

Probabilité marginale

Exemple précèdent: $p_{22}=2/6$

M: nombre des individus

Tableau de fréquence (probabilité)

J / I	1	....	i	...	m	Total
1	p_{11}	...	p_{1j}	...	p_{1m}	$p_{1.}$
⋮						
i	p_{i1}	...	p_{ij}	...	p_{im}	$p_{i.}$
⋮						
n	p_{n1}	...	p_{nj}	...	p_{nm}	$p_{n.}$
Total	$p_{.1}$		$p_{.j}$		$p_{.m}$	1

33

Z=

TABLEAU DE CONTINGENCE

MODALITE	HOMME	FEMME	
S	1	1	2
TM	1	2	3
L	0	1	1
	2	4	6

TABLEAU DE FREQUENCE

P=

			Pi.
	1/6	1/6	2/6
	1/6	2/6	3/6
	0/6	1/6	1/6
P.j	2/6	4/6	6/6=1

P_{ij} sont les fréquences relatives du tableau (les pourcentages)

Liaison et indépendance entre deux variables qualitatives

Model d'indépendance: $P(A \text{ et } B)=P(A)\times P(B)$

Variables qualitatives indépendantes: $P_{ij}= P_{i.} \times P_{.j}$

Analyse Factorielle des Correspondances (AFC)

L'AFC s'intéresse plus particulièrement aux effectifs marginaux des tableaux que l'on appelle **profils. Le tableau Z peut être alors transformé selon deux autres tableaux appelés** tableaux de profils.

Ainsi, de $Z_{(N,n)}$ on peut déduire deux matrices $X_{(N,n)}$ et $Y_{(N,n)}$:

Analyse Factorielle des Correspondances (AFC)

$$X_{(N,n)}(I) = \left\{\frac{z_{ij}}{z_{i.}}\right\} = \left\{\frac{p_{ij}}{p_{i.}}\right\}$$

Fréquences relatives des effectifs lignes

Tableau des pourcentages lignes

$$Y_{(N,n)}(J) = \left\{\frac{z_{ij}}{z_{.j}}\right\} = \left\{\frac{p_{ij}}{p_{.j}}\right\}$$

Fréquences relatives des effectifs colonnes

Tableau des pourcentages colonnes

L'AFC s'intéresse plus particulièrement aux effectifs marginaux des tableaux que l'on appelle **profils. Le tableau Z peut être alors transformé selon deux autres tableaux appelés** tableaux de profils.

Ainsi, de $Z_{(N,n)}$ on peut déduire deux matrices $X_{(N,n)}$ et $Y_{(N,n)}$:

PROFILS LIGNE)

$$X_{ij} = \frac{P_{ij}}{P_{i.}}$$ dans cet espace on utilise le tableau des profils lignes.

	1	…	j	…	n	
1						
…						
i			$p_{ij}/p_{i.}$			Coordonnées du point i dans R^n
…						
N						

PROFILS COLONNE

L'espace R^N (modalités lignes) dans lequel on peut représenter le nuage des n points (modalité colonne).

$$Y_{ij} = \frac{P_{ij}}{P_{.j}}$$ dans cet espace on utilise le tableau des profils colonnes.

	1	…	j	…	n
1					
…					
i			$p_{ij}/p_{.j}$		
…					
N					

Coordonnées de j dans R^N

PROFILS LIGNE – PROFILS COLONNE

PROFILS LIGNE

x=

1/2	1/2
1/3	2/3
0	1

PROFILS COLONNE

Y=

1/2	1/4
1/2	2/4
0	1/4

TABLEAU DE FREQUENCE

P=

	1/6	1/6	Pi. 2/6
	1/6	2/6	3/6
	0/6	1/6	1/6
P.j	2/6	4/6	6/6=1

La transformation initiale des données

L'information est donnée par la distance Euclidienne entre les points des nuages des deux espaces $\mathbf{R^n}$ et $\mathbf{R^N}$

par exemple dans $\mathbf{R^n}$

Calculons la distance euclidienne entre deux points quelconques : **x(i)** et **x(i')** de cet espace.

$$d^2(x(i), x(i')) = \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}} \right)^2$$

Analyse Factorielle des Correspondances (AFC)

En AFC, on n'utilise pas cette distance euclidienne. Plus précisément, on l'utilise mais après avoir effectué une transformation préalable des coordonnées des points du nuages. Dans l'espace $\mathbf{R^n}$ cette transformation s'écrit :

$$X_{ij} = \frac{1}{\sqrt{P_{.j}}} \frac{P_{ij}}{P_{i.}}$$

Distance du x2

En définitive, dans l'espace Rn on calcule la distance entre deux points x(i) et x(i') par la formule :

$$d^2(x(i), x(i')) = \sum_{j=1}^{n} \left(\frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}} - \frac{1}{\sqrt{p_{.j}}} \frac{p_{i'j}}{p_{i'.}} \right)^2 = \sum_{j=1}^{n} \frac{1}{p_{.j}} \left(\frac{p_{ij}}{p_{i.}} - \frac{p_{i'j}}{p_{i'.}} \right)^2$$

Analyse Factorielle des Correspondances (AFC)

On procède de façon équivalente pour l'espace $\mathbf{R^N}$
Considérons dans cet espace deux points du nuage y(j) et y(j')

$$Y_{ij} = \frac{1}{\sqrt{P_{i.}}} \frac{P_{ij}}{P_{.j}}$$

En définitive, dans l'espace Rn on calcule la distance entre deux points x(i) et x(i') par la formule :

$$d^2(y(j), y(j')) = \sum_{i=1}^{N} \left(\frac{1}{\sqrt{p_{i.}}} \frac{p_{ij}}{p_{.j}} - \frac{1}{\sqrt{p_{i.}}} \frac{p_{ij'}}{p_{.j'}} \right)^2 = \sum_{i=1}^{N} \frac{1}{p_{i.}} \left(\frac{p_{ij}}{p_{.j}} - \frac{p_{ij'}}{p_{.j'}} \right)^2$$

Détermination des composantes principales dans R^n

On se place dans l'espace des variables, on utilise donc la matrice des profils lignes. On vient de voir que dans cet espace les N points du nuage ont pour coordonnées :

$$X_{ij} = \frac{1}{\sqrt{P_{.j}}} \frac{P_{ij}}{P_{i.}}$$

On calcule la moyenne et la covariance de ces variables (notées xj pour j=1 à n)

Détermination des composantes principales dans R^n

On calcule la moyenne et la covariance de ces variables (notées xj pour j=1 à n)

La moyenne

$$\overline{X}_j = \sum_i p_{i.} X_{ij}$$

Moyenne arithmétique pondérée

$$\overline{X}_j = \sum_i p_{i.} \frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}} = \sum_i \frac{p_{ij}}{\sqrt{p_{.j}}} = \frac{1}{\sqrt{p_{.j}}} \sum_i p_{ij} = \frac{1}{\sqrt{p_{.j}}} p_{.j}$$

$$\overline{X}_j = \sqrt{p_{.j}}$$

Détermination des composantes principales dans R^n

La covariance

La covariance entre deux variables xj et xj' est :

$$\operatorname{cov}(x_j, x_i) = v_{jj'} = \sum_i p_{i.} \left[\left(\frac{1}{\sqrt{p_{.j}}} \frac{p_{ij}}{p_{i.}} - \sqrt{p_{.j}} \right) \left(\frac{1}{\sqrt{p_{.j'}}} \frac{p_{ij'}}{p_{i.}} - \sqrt{p_{.j'}} \right) \right]$$

$$V_{jj'} = \sum_i \frac{p_{ij} p_{ij'}}{\sqrt{p_{.j}} \sqrt{p_{.j'}} p_{i.}} - \sqrt{p_{.j}} \sqrt{p_{.j'}}$$

Analyse Factorielle des Correspondances Multiples (AFCM)

L'AFC est une méthode factorielle qui ne concerne que deux variables d'une population de n individus.

•L'AFCM (Analyse Factorielle des Correspondances Multiples) est une extension de l'AFC au cas ou l'on dispose de plus de deux variables sur cette population.

•Il s'agit d'une extension de l'AFC, les concepts utilisés dans l'AFC (comme ceux de l'ACP) sont repris par l'AFCM.

•Une particularité : on pourra ici représenter les individus, contrairement a l'AFC.

AFCM ou **ACM** = Analyse des Correspondances Multiples

Analyse Factorielle des Correspondances Multiples (AFCM)

Les données :

•On dispose de p variables qualitatives (Y1,…,Yp) mesurées sur n individus. La variable Yj a mj modalités.

		Sexe	Nationnalité	Couleur Yeux
Individus	1	homme	Français	Bleu
	2	femme	Etranger	Marron
	3	femme	Etranger	Noir
	4	homme	Etranger	Bleu
	5	femme	Français	Marron
	6	homme	Français	Noir
	N	femme	Français	Bleu

Tableau disjonctif complet

•La variable Yj est transformée en mj variables binaires.
•Le tableau disjonctif est une matrice de dimension n×M,

dans laquelle on affecte à l'individu i la valeur **1** à la colonne **m** si i possède la modalité **m** de Yj et **0** sinon.

Tableau disjonctif complet(TDC)

	1								l							M	
1																	p
i									K_{il}								$K_{i.} = p$
n																	p
									$K_{.l}$								np

$$K_{.l} = \sum_{i=1}^{n} K_{il} = \text{nombre d'individus possédant la modalité } l$$

$$K_{i.} = \sum_{l=1}^{M} K_{il} = p = \text{nombre de modalités possédées par i}$$

$$\sum_{i=1}^{n}\sum_{l=1}^{M} K_{il} = np$$

$$M = \sum_{j=1}^{p} mj$$

Tableau disjonctif complet(TDC)

		Sexe	Nationnalité	Couleur Yeux
Individus	1	homme	Français	Bleu
	2	femme	Etranger	Marron
	3	femme	Etranger	Noir
	4	homme	Etranger	Bleu
	5	femme	Français	Marron
	6	homme	Français	Noir
	N	femme	Français	Bleu

Tableau disjonctif						
Homme	Femme	Français	Etranger	Yeux bleus	Marron	Noir
1	0	1	0	1	0	0
0	1	0	1	0	1	0
0	1	0	1	0	0	1
1	0	0	1	1	0	0
0	1	1	0	0	1	0
1	0	1	0	0	0	1
0	1	1	0	1	0	0

Tableau des fréquences(probabilité)

	1			m_1					l							M	
1																	$\frac{1}{n}$
i									f_{il}								$f_{i.} = \frac{1}{n}$
n																	$\frac{1}{n}$
									$f_{.l}$								1

$$f_{il} = \frac{K_{il}}{np}, f_{i.} = \frac{K_{i.}}{np} = \frac{1}{n}, f_{.l} = \frac{K_{.l}}{np}$$

Donc distribution marginale ligne est une constante

$$f_{i.} = \frac{1}{n}$$

$$\sum_{i=1}^{n}\sum_{m=1}^{M} f_{il} = \sum_{i=1}^{n} f_{i.} = \sum_{m=1}^{M} f_{.l} = 1$$

Le tableau de Burt

A partir du tableau TDC on peut construire le tableau de Burt :

$$BURT_{(p,p)} = TDC'_{(p,n)} \times TDC_{(n,p)}$$

Le tableau de Burt est donc le produit matriciel entre la transposée du tableau disjonctif complet et lui même.

Le tableau de Burt est donc une matrice carrée et symétrique qui croise les questions entre elles. Sur sa **diagonale principale** on trouve le croisement des questions entre elles (le tris à plat) et de part et d'autre de la diagonale principale **les croisements entre deux questions** distinctes (tris croisés).

Le tableau de Burt

TCD’

Homme	1	0	0	1	0	1
Femme	0	1	1	0	1	0
Algérien	1	0	0	0	1	1
Etranger	0	1	1	1	0	0
Yeux bleu	1	0	0	1	0	0
Marron	0	1	0	0	1	0
Noir	0	0	1	0	0	1

TCD

Homme	Femme	Algérien	Etranger	Yeux bleu	Marron	Noir
1	0	1	0	1	0	0
0	1	0	1	0	1	0
0	1	0	1	0	0	1
1	0	0	1	1	0	0
0	1	1	0	0	1	0
1	0	1	0	0	0	1

On peut ainsi voir sur l’exemple que :
Tris à plat : Nombre d’hommes=3 ; nombre de femmes=3
Etrangers=3 ; Algerien =3
Tris croisés : Parmi les hommes, il y a 2 Algérien et 1 étranger
Parmi les femmes il y a 1 Algérienne et 2 étrangères

	Homme	Femme	Algérien	Etranger	Yeux bleu	Marron	Noir
Homme	3	0	2	1	2	0	1
Femme	0	3	1	2	0	2	1
Algérien	2	1	3	0	1	1	1
Etranger	1	2	0	3	1	1	1
Yeux bleu	2	0	1	1	2	0	0
Marron	0	2	1	1	0	2	0
Noir	1	1	1	1	0	0	2

Tris à plat

Tris croisé

Tableaux des profils

- Profils lignes :

		l			
i		$f_{il}/f_{i.}$			
n					

$$f_{il} = \frac{K_{il}}{np}$$

$$f_{.l} = \frac{1}{n}$$

$$\frac{f_{il}}{fi.} = \frac{K_{il}}{p}$$

Tableaux des profils

- Profils colonne :

		l			
i		$f_{il}/f_{.j}$			
n					

$$f_{il} = \frac{K_{il}}{np}$$

$$f_{.l} = \frac{k.l}{np}$$

$$\frac{f_{il}}{f.l} = \frac{K_{il}}{K.l}$$

Passage de AFCM vers AFC

➔Comme pour l'AFC on calcule la part de la variance expliquée par les composante principales