L'ANALYSE FACTORIELLE DES CORRESPONDANCES

Mr Z.Bouafia

Master RID

2016 / 2017

Introduction

- L’AFC a pour objet le traitement de l’information contenue dans un tableau appelé de contingence ou de dépendance, relatif a deux ensembles de nature quelconque, en relation par moyen d’un processus naturel ou expérimental plus ou moins bien connu.
- Les données sont ici pondérées. Les fréquences de répétitions s’interprète facilement en termes de probabilités.
- Le tableau de dépendance peut être ainsi représenté dans un espace approprié par un nuage de points affectés de probabilités.

Tableau de contingence

- Considérons un tableau a double entrée :

Ensemble J (paramèttres) / *Ensemble I (individus)*	1	....	J	...	*m*
1	X_{11}	...	X_{1j}	...	X_{1m}
i	X_{i1}	...	X_{ij}	...	X_{im}
n	X_{n1}	...	X_{nj}	...	X_{nm}

Tableau de contingence (exemple)

Observations		
individu	bac	sexe
1	S	homme
2	ES	femme
3	ES	homme
4	A	femme
5	S	femme
5	ES	femme

Tableau de contingence		
modalité	homme	femme
S	1	1
ES	1	2
A	0	1

Tableau de contingence

- Dans le cas qualitatif, le tableau précédent se présente sous la forme d'un tableau des uns et des zéros (suivant si l'individu i possède ou non le paramètre j).
- La probabilité associée au terme $\mathbf{x_{ij}}$ est:

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}}$$

Probabilité

- Ou les probabilités marginales sont:

$$p_{i\bullet} = \sum_{j=1}^{m} p_{ij} \quad avec \quad i = 1, \cdots, n$$

$$p_{\bullet j} = \sum_{i=1}^{n} p_{ij} \quad avec \quad j = 1, \cdots, m$$

- qui vérifient les propriétés:

$$\sum_{i=1}^{n} p_{i\bullet} = 1 \quad et \quad \sum_{j=1}^{m} p_{\bullet j} = 1$$

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

C’est quoi « les correspondances »?

- Lorsque les variables sont **quantitatives**, on fait une étude de **corrélation**.
- Mais, lorsqu’on a aussi des variables **qualitatives**, on doit faire une étude des **correspondances**.

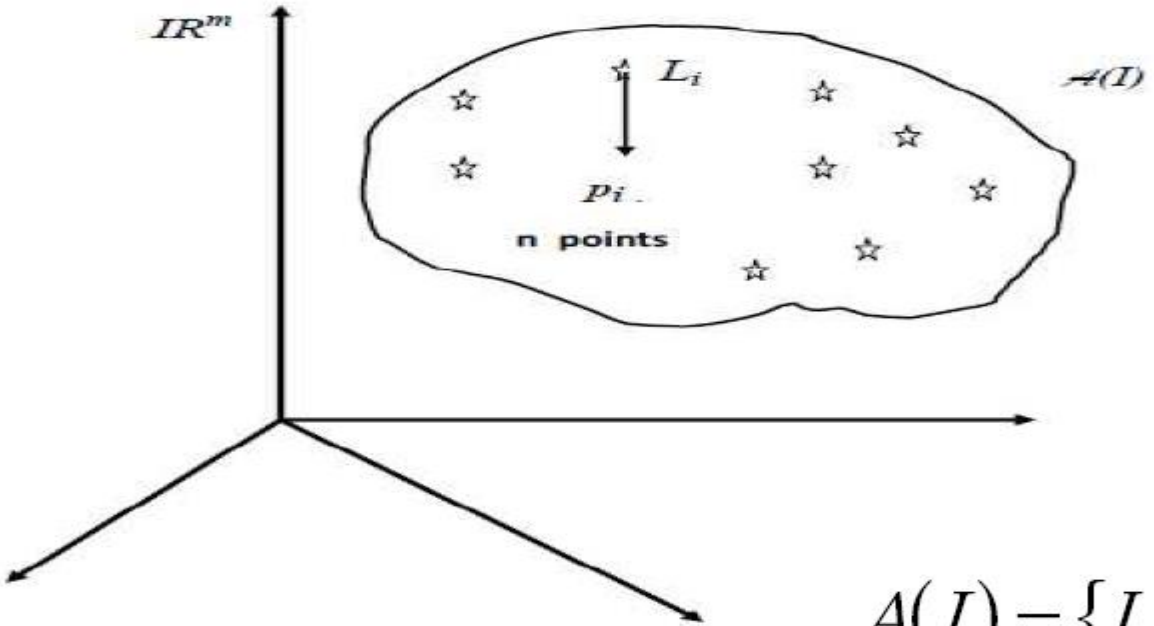
Indépendance?

- Probabilités conditionnelles, dans ce cas:

$$\frac{p_{ij}}{p_{i\cdot}} = p_{\cdot j} \Leftrightarrow \frac{p_{ij}}{p_{\cdot j}} = p_{i\cdot}$$

- Formule d'indépendance:

$$p_{ij} = p_{i\cdot} \times p_{\cdot j}$$



$$A(I) = \{L_i;\ p_{i\cdot}\}$$

$$L_i = \left(\frac{p_{i1}}{p_{i\cdot}}, \frac{p_{i2}}{p_{i\cdot}}, \ldots, \frac{p_{ij}}{p_{i\cdot}}, \ldots, \frac{p_{im}}{p_{i\cdot}} \right)$$

Distance du χ^2

- Pour deux individus quelconques *i* et *i'*:

$$d^2(L_i, L_{i'}) = \sum_j \frac{1}{p_{\cdot j}} \left(\frac{p_{ij}}{p_{i\cdot}} - \frac{p_{i'j}}{p_{i'\cdot}} \right)^2$$

- Plus généralement, la distance du χ^2 est égale à la distance euclidienne entre:

$$\left(\frac{p_{i1}}{p_{i\cdot}\sqrt{p_{\cdot 1}}}, \frac{p_{i2}}{p_{i\cdot}\sqrt{p_{\cdot 2}}}, \cdots, \frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}, \cdots, \frac{p_{im}}{p_{i\cdot}\sqrt{p_{\cdot m}}} \right)$$

$$\left(\frac{p_{i'1}}{p_{i'\cdot}\sqrt{p_{\cdot 1}}}, \frac{p_{i'2}}{p_{i'\cdot}\sqrt{p_{\cdot 2}}}, \cdots, \frac{p_{i'j}}{p_{i'\cdot}\sqrt{p_{\cdot j}}}, \cdots, \frac{p_{i'm}}{p_{i'\cdot}\sqrt{p_{\cdot m}}} \right)$$

Distance du χ^2

- Ce sont les points qu'on a noté $\mathbf{M_i}$ dans le cours

$$M_i = \left(\beta_{i1}, \beta_{i2}, \cdots, \beta_{ij}, \cdots, \beta_{im}\right)$$

- Avec

$$\beta_{ij} = \frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}}$$

- $p_{i\cdot}$ étant toujours la pondération

Distance du χ^2

- Ainsi la distance du χ^2 entre deux points M_i et $M_{i'}$ est:

$$d^2(M_i, M_{i'}) = \sum_j \left(\beta_{ij} - \beta i'j\right)^2$$

Projection du nuage B(I) sur un axe

- On projet orthogonalement le nuage **B(I)** sur un axe (espace vectoriel de dim 1) de vecteur unitaire **u**, de telle façon que l'information perdue soit minime.
- Ce qui revient a trouver la valeur propre la plus grande λ_{max} de **W**. avec **W** la matrice des variances-covariances de B(I),
- La matrice des variances-covariances **W** du nuage **B(I)** relativement à un paramètre **j** est:

$$W = \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1m} \\ V_{21} & V_{22} & \cdots & V_{2m} \\ \vdots & & \ddots & \vdots \\ V_{m1} & V_{m2} & \cdots & V_{mm} \end{pmatrix}$$

Matrice des variances-covariances W

- La variance V_{ij} caractérise la dispersion du nuage tout au long de l'axe j:

$$v_{jj} = \sum_i p_{i\cdot} \left(\beta_{ij} - \sqrt{p_{\cdot j}} \right)^2$$

- La covariance V_{jk} est :

$$v_{jk} = \sum_i p_{i\cdot} \left(\beta_{ij} - \sqrt{p_{\cdot j}} \right) \left(\beta_{ik} - \sqrt{p_{\cdot k}} \right)'$$

Matrice des variances-covariances W

- Soit encore, en remplaçant β_{ij} par sa valeur:

$$V_{jk} = \sum_i \left(\frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot} p_{\cdot k}}} \right) \left(\frac{p_{ik} - p_{i\cdot} p_{\cdot k}}{\sqrt{p_{i\cdot} p_{\cdot k}}} \right)'$$

- Posons $\dfrac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot} p_{\cdot k}}} = r_{ij} \;\; ; \;\; i = 1, \cdots, n \, , \;\; j = 1, \cdots, m$

Matrice des variances-covariances W

$$\left(r_{ij}\right)_{\substack{1\le i\le n \\ 1\le j\le m}} = R$$

$$W = R' R$$

- ou *R'* est la transposée de *R*.
- Maximiser **u'Wu** revient à maximiser **u'R'Ru** sous la condition **u'u=1**, c'est-a-dire déterminer les vecteurs propres associes aux valeurs propres de la matrice **R'R**.

Variabilité totale du nuage B(I)*:*

- On appelle la variabilité totale du nuage **B(I)**, la trace de la matrice **W**:

$$V_B = tr(W) = \sum_j v_{jj}$$

- On parle aussi de la variabilité totale du nuage projeté **C(I)** qui sera $\mathbf{V_c = \lambda_{max}}$

- La partie de variabilité expliquée par la projection de B*(I),* sur *u* est alors:

$$\delta = \frac{Vc}{V_B} \qquad \delta = \frac{\lambda_{\max}}{tr(W)}$$

Inconvénients et avantages de l'AFC

- Les inconvénients sont les défauts de toute analyse factorielle: déformation inévitable du nuage durant la projection et la signification ou interprétation des axes.
- L'avantage essentiel réside dans l'étude des caractères qualitatifs.