Analyse Factorielle des Correspondances (AFC)

Prof. Mohamed El Merouani
Faculté Polydisciplinaire de Tétouan
Département de Statistique et Informatique

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**.

2

®El Merouani FP Tetouan

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:

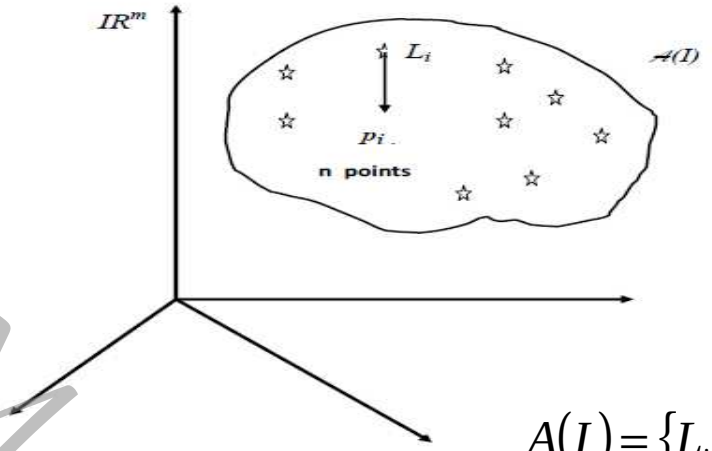
$$p_{ij} = p_{i\cdot} \times p_{\cdot j}$$

3

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

4

®El Merouani FP Tetouan



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

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

5

Distance du χ^2

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

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

- Pourquoi une telle distance?

6

®El Merouani FP Tetouan

Pourquoi la distance du χ^2?

- La distance euclidienne ne prend pas compte complètement de tous les caractères étudiés!
- Il a été alors proposé de modifier la distance euclidienne en tenant compte des écarts entre deux probabilités de deux individus d'avoir un caractère en donnant de l'importance aux probabilités que l'individu ait tous les caractères étudiés.
- Cela, donc, par multiplication par l'inverse de la probabilité d'avoir tous les caractères.

7

Pourquoi la distance du χ^2?

- Aussi, parce que la distance du χ^2 a une propriété qui s'appelle «**la propriété d'équivalence distributionnelle** » et que la distance euclidienne ne vérifie pas!
- Si deux colonnes j et j' de J correspond au même ligne i, il est logique de les regrouper en une seule de probabilité ($p_{ij}+p_{ij'}$), il faut alors que cette opération ne modifie pas les distances entre les i.

8

®El Merouani FP Tetouan

Pourquoi la distance du χ^2?

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

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

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

9

- Ce sont les points qu'on a noté M_i dans le cours

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

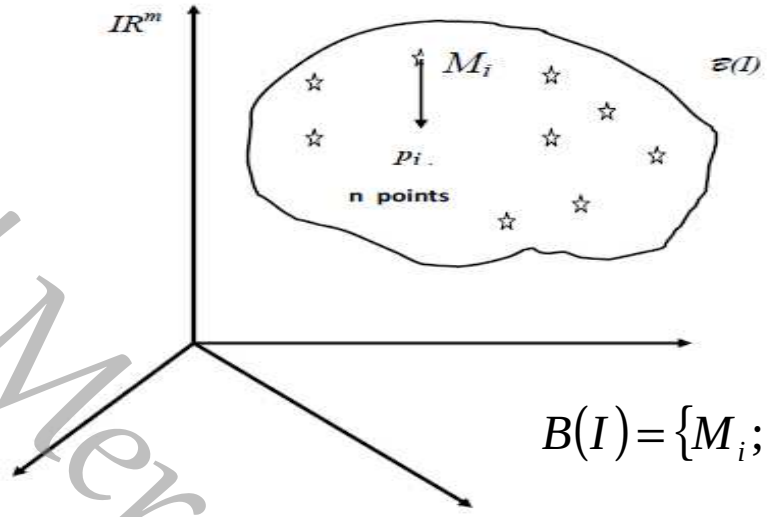
- Avec

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

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

10

®El Merouani FP Tetouan



$$B(I) = \{M_i ; p_{i.}\}$$

11

- Ainsi la distance du χ2 entre deux points M_i et $M_{i'}$ est:

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

12

®El Merouani FP Tetouan

Projection du nuage $\mathcal{B}(I)$ sur un axe:

- On projet orthogonalement le nuage $\mathcal{B}(I)$ sur un axe (espace vectoriel de dim 1) de vecteur unitaire u, de telle façon que l'information perdue soit minime.
- Comme en ACP, ce qui revient à max $u'Wu$, sous la condition $u'u$=1, avec W est la matrice des variances-covarainces de $\mathcal{B}(I)$.
- Ce qui revient à trouver la valeur propre la plus grande λ_{max} de W.

13

Matrice des variances-covariances W:

- La matrice des variances-covariances W du nuage $\mathcal{B}(I)$ relativement à un paramètre j est:

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

14

®El Merouani FP Tetouan

Matrice des variances-covariances W:

- La variance v_{jj} 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

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

15

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

16

®El Merouani FP Tetouan

Matrice des variances-covariances W:

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

$$W = R'R$$

où R' est la transposée de R.

• Maximiser $u'Wu$ revient à maximiser $u'R'Ru$ sous la condition $u'u=1$, c'est-à-dire déterminer les vecteurs propres associés aux valeurs propres de la matrice $R'R$.

17

Variabilité totale du nuage $\mathcal{B}(I)$:

- On appelle la variabilité totale du nuage $\mathcal{B}(I)$, la trace de la matrice W:

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

- On parle aussi de la variabilité totale du nuage projeté $\mathcal{G}(I)$ qui sera $Vc=\lambda_{max}$

18

®El Merouani FP Tetouan

Variabilité expliquée du nuage $\mathcal{B}(I)$:

- La partie de variabilité expliquée par la projection de $\mathcal{B}(I)$, sur u est alors:

$$\delta = \frac{Vc}{V_B}$$

- Soit encore:

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

19

Projection du nuage $\mathcal{B}(I)$ sur un plan:

- Comme en ACP, les vecteurs propres de W s'appellent « axes factoriels » du nuage.
- La détermination des axes factoriels se fait en diagonalisant la matrice symétrique W.
- En pratique, on se contente des valeurs propres les plus grands.

20

®El Merouani FP Tetouan

Recherche des facteurs:

- Les points du nuage $\mathcal{T}(I)$ possèdent un nombre réduit de coordonnées dans le référentiel formé les axes factoriels.
- Ces coordonnées sont les valeurs de nouvelles variables qui s'appellent: Facteurs.
- Le premier facteur correspond aux coordonnées sur le premier axe factoriel.

21

Recherche des facteurs:

- On peut démontrer que les facteurs sont non-corrélés entre eux et s'expriment comme combinaisons linéaires des données.
- Réciproquement, les données ont des coefficients qui sont des combinaisons linéaires des facteurs.
- Ainsi, à partir des facteurs, il est possible de reconstruire un tableau de données avec une minime perte d'information, c'est-à-dire obtenir un tableau plus facilement accessible à l'analyse descriptive.

22

®El Merouani FP Tetouan

Proximité en IR^m et en IR^n:

- On a vu précédemment les proximités entre n points de IR^m.
- Par des calculs symétriques, on peut étudier les proximités de m points de IR^n.
- Sauf qu'il existe des relations entre les facteurs de IR^m et les facteurs de IR^n.
- Il est alors possible de représenter, sur le même graphique, dans le plan des deux premiers axes factoriels, les proximités entre les individus et les proximités entre les caractères.
- Cette simultanéité de représentation suggère parfois une interprétation économique, sociale, politique,…des axes factoriels.

23

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.

24

®El Merouani FP Tetouan