Analyse Factorielle des Correspondances: AFC

Mme L.HAMDAD

Plan

Introduction

L'AFC est une méthode de l'analyse de données qui consiste à analyser des tableaux de contingence. Le tableau de contingence résume le résultat d'observations de N individus selon deux variables qualitatives X à n modalités et Y à p modalités. Le but de l'AFC est d'étudier les liaisons entre les modalités de X et Y.

Exemple

- Observation des types de diplômes obtenus selon les catégories socio-professionnelles des parents. Le but est de connaître la relation qui existe entres les modalités de ces variables. L'AFC permet entre autres de visualiser ces données et ainsi d'en tirer une information visuelle.
- Pour pouvoir améliorer la programmation d'emissions de télévision, un nombre N d'individus sont observés selon différentes emissions de Télé et les Appréciation qu'ils donnent à ces emissions (excellent, mauvais, inconnus,...).

Tableau de données

Tableau de données :

$K =$

	Y^1	$\cdots$	Y^j	$\cdots$	Y^p	$k_{i.}$
X1	k_{11}	$\cdots$	k_{1j}			
$\vdots$						
X_i	k_{i1}	$\cdots$	k_{ij}	$\cdots$	k_{ip}	
$\vdots$						
X_n	k_{n1}	$\cdots$	k_{nj}	$\cdots$	k_{np}	
$k_{.j}$						N

k_{ij} : représente le nombre d'individus ayant pris la modalité i pour X et j pour Y.
$k_{i.}$: représente le nombre d'individus ayant la modalité i pour X.
$k_{.j}$: représente le nombre d'individus ayant la modalité j pour Y.
A ce tableau de donnée est associé un tableau de fréquences

$$F = \frac{K}{n} = \left[f_{ij} = \frac{k_{ij}}{n} \right]_{\substack{i=1,\ldots,n \\ j=1,\ldots,p.}}.$$

Détermination des tableaux de profils et les métriques associées

Tableau des profils lignes

$$X_L = \begin{pmatrix} \frac{k_{11}}{k_{1.}} & \frac{k_{12}}{k_{1.}} & \frac{k_{1p}}{k_{1.}} \\ \frac{k_{21}}{k_{2.}} & & \\ & & \\ \frac{k_{n1}}{k_{n.}} & & \frac{k_{np}}{k_{n.}} \end{pmatrix}$$

A la $i$ème ligne de X_L, on associe le vecteur $X_L(i) = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^P$,

avec $X_L(ij) = \frac{k_{ij}}{k_{i.}}$.

Tableau des profils colonnes

$$X_C = \begin{array}{|c|c|c|c|} \hline \frac{k_{11}}{k_{.1}} & \frac{k_{21}}{k_{.1}} & & \frac{k_{n1}}{k_{.1}} \\ \hline \frac{k_{12}}{k_{.2}} & & & \\ \hline & & & \\ \hline \frac{k_{1p}}{k_{.P}} & & & \frac{k_{np}}{k_{.p}} \\ \hline \end{array}$$

A la $j^{ème}$ colonne de X_C , on associe le vecteur $X_C(j) = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$,

avec $X_C(ji) = \frac{k_{ij}}{k_{.j}}$.

Nuages de points associé aux profils

:

a- **Profils lignes** : On note par $N(I)$ le nuage de points associé au profils lignes donnée par

$$N(I) = \{(X_L(i), f_{i.}), i = 1, ..., n\}$$

$f_{i.}$ est le poids associé au ième profil ligne. On définit ainsi la métrique des poids $D_p = \begin{pmatrix} f_{1.} & & 0 \\ & \ddots & \\ 0 & & f_{n.} \end{pmatrix}$. Le centre de gravité de $N(I)$ est donné par

$$g_n = \sum_{i=1}^{n} f_{i.} X_L(i) = \begin{pmatrix} f_{.1} \\ \vdots \\ f_{.p} \end{pmatrix}.$$

b- **Profils colonne** : On note par $N(J)$ le nuage de points associé aux profils colonnes donné par

$$N(J) = \{(X_C(j), f_{.j}), j = 1, ..., p\}$$

$f_{.i}$ est le poids associé au $j$ème profil colonne.

la métrique des poids $D_n = \begin{pmatrix} f_{.1} & & 0 \\ & \ddots & \\ 0 & & f_{.p} \end{pmatrix}$

Le centre de gravité de $N(J)$:

$$g_p = \sum_{j=1}^{p} f_{.j} X_C(j) = \begin{pmatrix} f_{1.} \\ \vdots \\ f_{n.} \end{pmatrix}.$$

L'écriture de X_L et de X_C en fonction du tableau des fréquences donne respectivement : $X_L = D_n^{-1} F$ et $X_L = D_p^{-1} F^t$.

Problème de l'AFC

Formellement, l'AFC consiste à :
- Effectuer l'ACP des profils lignes.

- Effectuer l'ACP des profils colonnes.

- Déduire les relations dites barycentriques entre les projections des deux profils.

- Et ainsi, visualiser ces profils sur des sous espaces de dimension réduite.

Avant d'entamer l'analyse de $N(I)$, on commence par définir la distance du $KHI2$ entre deux profils lignes par :

$$d^2(i,s) = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{k_{ij}}{k_{i.}} - \frac{k_{sj}}{k_{s.}} \right)^2$$

Ainsi la métrique choisie sur l'espace $\mathbb{R}^p$ est $Q_p = D_p^{-1}$. Cette distance peut être réecrite comme suite

$$d^2(i,s) = \sum_{j=1}^{p} \left(\frac{k_{ij}}{k_{i.}\sqrt{f_{.j}}} - \frac{k_{sj}}{k_{s.}\sqrt{f_{.j}}} \right)^2$$

- A partir de l'écriture précedente, nous remarquons qu'au lieu d'analyser le $N(I) = \{(X_L(i), f_{i.}), i = 1, ..., n / X_L(i) \in (R^p, Q_p)\}$, on peut analyser le nuage $N(\widetilde{I}) = \left\{ (\widetilde{X}_L^j(i), f_{i.}), i = 1, ..., n / \widetilde{X}_L(i) \in (R^p, I) \right\}$, tel que $\widetilde{X}_L^j(i) = \frac{k_{ij}}{k_{i.}\sqrt{f_{.j}}}$.

- L'écriture de $\widetilde{X}_L$ en fonction du tableau des fréquences donne $\widetilde{X}_L = D_n^{-1} F D_p^{-\frac{1}{2}}$.

Avantage de la distance du *KHI2*

Cette distance satisfait au critère d'équivalence distributionnelle, c'est à dire que si deux profils colonnes sont identiques, alors la distance entre deux profils lignes quelconques ne change pas en regroupant les profils colonnes en un profil y^* de poids égale à la somme de leur deux poids. L'analyse factorielle du nuage $N(\widetilde{I})$ conduit à la diagonalisation de la matrice $V = \widetilde{X}_L^t D_p \widetilde{X}_L$, tel que $\widetilde{X}_L^j = \begin{pmatrix} \widetilde{X}_L^t(1) \\ \vdots \\ \widetilde{X}_L^t(n) \end{pmatrix}$

Les résultats de l' Analyse factorielle(AF)

- La matrice V a toutes ses valeurs propres réelles comprises entre 0 et 1.
- Le centre de gravité g_n est un vecteur normé.
- Le centre de gravité g_n est orthogonal à tout profil ligne centré.
- g_n est vecteur propre de V associé à la valeur propre $\lambda = 1$.
- Les axes factoriels qui ajustent au mieux le nuage des profils lignes sont engendrés par $u_1 = g_n, u_2, ..., u_p$ orthonormés associés aux valeurs propres $\lambda_1 = 1 \geq ... \geq \lambda_p \geq 0$.
- Le nombre de valeurs propres est égales au $min(n-1, p-1)$.
- L'axe engendré par le centre de gravité est appelé, l'axe trivial.

Projection des profils lignes

La projection du $i$ème profil ligne centré $(\widetilde{X}_L(i) - g_n)$ sur l'axe factoriel Δu_r est donnée par :

$$w_{ir} = \left\langle \widetilde{X}_L(i) - g_n, u_r \right\rangle$$

Comme u_r et g_n sont orthogonaux alors

$$w_{ir} = \widetilde{X}_L^t(i) u_r$$

Ainsi le vecteur de projection est donné par

$$W_r = \begin{pmatrix} \widetilde{X}_L^t(1) u_r \\ \vdots \\ \widetilde{X}_L^t(n) u_r \end{pmatrix} = \widetilde{X}_L u_r$$

: $A_r = Q_p^{\frac{1}{2}} u_r$ est appelé facteur associé aux profils lignes.

Analyse factorielle de N(J)

On définit la distance du $KHI2$ entre deux profils colonnes par :

$$d^2(j,d) = \sum_{i=1}^{n} \frac{1}{f_{i.}} \left(\frac{k_{ij}}{k_{.j}} - \frac{k_{id}}{k_{.d}} \right)^2$$

Ainsi la métrique choisie sur l'espace $\mathbb{R}^n$ est $Q_n = D_n^{-1}$. Cette distance peut être réecrite

$$d^2(j,d) = \sum_{j=1}^{p} \left(\frac{k_{ij}}{k_{.j}\sqrt{f_{i.}}} - \frac{k_{sd}}{k_{.d}\sqrt{f_{i.}}} \right)^2$$

Comme pour l'AF du nuage $N(I)$ précédente, au lieu d'analyser le $N(J) = \{(X_C(j), f_{.j}), i = 1, ..., p / X_C(j) \in (R^n, Q_n)\}$, on peut analyser le nuage $N(\widetilde{J}) = \left\{ (\widetilde{X}_C^j(j), f_{.j}), i = 1, ..., p / \widetilde{X}_C(j) \in (R^n, I) \right\}$, tel que $\widetilde{X}_{iL}(j) = \frac{k_{ij}}{k_{.j}\sqrt{f_{i.}}}$.

- L'écriture de $\widetilde{X}_C$ en fonction du tableau des fréquences donne $\widetilde{X}_C = D_p^{-1} F^t D_n^{-\frac{1}{2}}$.

Résultat de l'AF de $N(J)$

Nous avons les résultats suivants : L'analyse factorielle du nuage $N(\widetilde{J})$ conduit à la diagonalisation de la matrice $S = \widetilde{X}_C^t D_n \widetilde{X}_C$, tel que

$$\widetilde{X}_C^j = \begin{pmatrix} \widetilde{X}_C^t(1) \\ \vdots \\ \widetilde{X}_C^t(p) \end{pmatrix}$$

- La matrice S a toutes ses valeurs propres réelles comprises entre 0 et 1.
- Le centre de gravité g_p est normé.
- Le centre de gravité g_p est orthogonal à tout profil colonne centré.
- g_p est vecteur propre de S associé à la valeur propre $\lambda = 1$.
- Les axes factoriels qui ajustent au mieux le nuage des profils colonnes, sont engendrés par $v_1 = g_p, v_2, ..., v_p$ orthonormés associés aux valeurs propres $\lambda_1 = 1 \geq ... \geq \lambda_p \geq 0$.
- Le nombre de valeurs propres est égales au $min(n-1, p-1)$.
- L'axe engendré par le centre de gravité g_p est apelé axe trivial.

Projection des profils colonnes

La projection du $j$ème profil colonne centré $(\widetilde{X}_C(j) - g_p)$ sur l'axe factoriel Δv_r est donnée par :

$$V_{jr} = \left\langle \widetilde{X}_C(j) - g_p, v_r \right\rangle$$

Comme v_r et g_p sont orthogonaux alors

$$V_{jr} = \widetilde{X}_C^t(j) v_r$$

Ainsi le vecteur de projection est donnée par

$$V_r = \begin{pmatrix} \widetilde{X}_C^t(1) v_r \\ \vdots \\ \widetilde{X}_C^t(p) v_r \end{pmatrix} = \widetilde{X}_C v_r$$

: $B_r = Q_n^{\frac{1}{2}} v_r$ est appelé facteur associé aux profils colonnes.

Formules de transitions

Dans cette section, nous nous intéressons à la relation qui existe entre les analyses factorielles de $N(I)$ et $N(J)$. Cette relation découle de la proposition suivante :

Proposition

Les matrices V et S ont les mêmes valeurs propres.

Formules de transitions entre vecteurs A partir de la preuve, nous obtenons les formules de transition entre vecteurs.
La première formule de transition entre vecteurs :

$$v = \frac{1}{\sqrt{\lambda}} D_n^{-\frac{1}{2}} F D_p^{-\frac{1}{2}} u$$

la deuxième formule :

$$u = \frac{1}{\sqrt{\lambda}} D_p^{-\frac{1}{2}} F^t D_n^{-\frac{1}{2}} v$$

Formule de transitions entre facteurs : A partir des formules de transition entre vecteurs, nous avons

$$A = \frac{1}{\sqrt{}} D_p^{\frac{-1}{2}} D_p^{-\frac{1}{2}} F^t D_n^{-\frac{1}{2}} v = \frac{1}{\sqrt{}} Q_p F^t B$$

Formules barycentrique

Dans ce qui suit, on donne la relation barycentrique entre la projection des profil lignes et celle des profils colonnes :

Proposition

à $\frac{1}{\sqrt{\lambda}}$ prés, la projection des profils lignes est le barycentre des projections des profils colonnes de masses $\frac{k_{ij}}{k_{i.}}$, c'est à dire que :

$$W_r(i) = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{k_{ij}}{k_{i.}} V_r(j).$$

De même à $\frac{1}{\sqrt{\lambda}}$ prés La projection des profils colonne est le barycentre des projections des profils lignes de masses $\frac{k_{ij}}{k_{.j}}$, c'est à dire que :

$$V_r(j) = \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{n} \frac{k_{ij}}{k_{.j}} W_r(i)$$

Formules de reconstitution et profils supplémentaires

Le tableau des fréquences $F = \frac{K}{n} = \left[f_{ij} = \frac{k_{ij}}{n}\right]_{\substack{i=1,\dots,n \\ j=1,\dots,p.}}$ est reconstituées

Ã *cent* pour *cent* cellule par cellule, i.e :

$$f_{ij} = f_{i.} f_{.j}(1 + \sum_{\alpha} \frac{1}{\sqrt{\lambda_\alpha}} Y_\alpha(i) V_\alpha(j)).$$

ou bien

$$f_{ij} = f_{i.} f_{.j}(1 + \sum_{\alpha} \sqrt{\lambda_\alpha} B_\alpha(i) A_\alpha(j)).$$

tels que A_α et B_α représentent les facteurs des profils lignes et des profils colonnes respectivement.

Projection de profils supplémentaires :Supposons un profil ligne (colonne) supplémentaire $X_L(n+1)(X_C(p+1))$, sa projection sur Δ_α est donnÃ©e par :

- Profil ligne supplémentaire :

$$W_r(n+1) = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{p} \frac{k_{n+1j}}{k_{n+1.}} V_r(j).$$

- Profil colonne supplémentaire :

Aides à l'interprétation

Qualité globale de représentation La quantité

$$I_{\Delta u_\alpha} = \frac{\lambda_\alpha}{\sum \lambda_k} \quad \text{tel que } \lambda_k \neq 1.$$

appellée inertie expliquée par l'axe factoriel Δu_αdifférent de l'axe trivial mesure la qualité globale de représentation de ce dernier.
Contribution absolue Elle mesure la part du profil ligne ou colonne à la construction des facteurs
a- **profil ligne** :

$$C_{ab}^r(i) = f_{i.} \frac{(w_r(i))^2}{\lambda_r}$$

b- **profil colonne** :

$$C_{ab}^r(j) = f_{.j} \frac{(V_r(j))^2}{\lambda_r}$$

$\sum_i C_{ab}^r(i) = \sum_j C_{ab}^r(j) = 1.$
- Dés que $C_{ab}^r(i) > f_{i.}$, on dira que le profil ligne i contribue à la construction de Δu_r et de même pour les profils colonnes.

Contribution relatives

Elle mesure la qualité de représentation des profils par l'axe factoriel.
a- **profil ligne** :

$$C_{re}^{r}(i) = \cos_i^2(\theta) = \frac{(w_r(i))^2}{d^2_{Q_p}(g_n, X_L(i))}$$

b- **profil colonne**

$$C_{re}^{r}(j) = \cos_j^2(\theta) = \frac{(V_r(j))^2}{d^2_{Q_n}(g_p, X_C(j))}$$

Dés que $\cos_i^2(\theta)\left(\cos_j^2(\theta)\right) \simeq 1$, on dira que le profil est très bien représenté par l'axe factoriel.

Récapitulation : Algorithme AFC

1. Calculer le tableau des profils lignes $\widetilde{X_L}$.
2. Calculer la matrice $\tilde{A}$ diagonaliser $V = \widetilde{X_L}_t D_n \widetilde{X_L}$.
3. Calculer les valeurs propres et vecteurs propres de V.
4. Considérer que les valeurs propres supérieures à 1 et calculer la projections des profils lignes.
5. Calculer les projections des profils colonnes en utilisant les formules de transitions entre facteurs.
6. Représenter graphiquement les profils lignes et les profils colonnes.
7. Interpréter