Analyse des données

Chapitre 3

Analyse factorielle des correspondances (AFC)

1. Introduction

L'analyse factorielle des correspondances (AFC) consiste à étudier les *relations (dépendances) existant entre les modalités de deux variables qualitatives* observées sur un ensemble d'individus. Par conséquent, nous nous intéressons au problème de réduction de la dimension des données, de la représentation graphique simultanée des modalités et de l'interprétation.

2. Présentation des données

Considérons deux variables qualitatives X et Y observées sur n individus telles que :

- X possède p modalités : $x_1, x_2, \dots, x_p$,
- Y possède q modalités : $y_1, y_2, \dots, y_q$,

Alors, les $p \times q$ observations se présentent dans un tableau (matrice) $K = (k_{ij})_{i,j}$ pour $i = 1,2, \dots, p$ et $j = 1,2, \dots, q$, appelé tableau (matrice) de ***contingences*** croisant les deux variables X et Y :

$K =$

X \ Y	y_1	…	y_j	…	y_q	*Total*
x_1	k_{11}		k_{1j}		k_{1q}	k_{1+}
⋮			⋮			
x_i	k_{i1}	…	k_{ij}	…	k_{iq}	k_{i+}
⋮			⋮			
x_p	k_{p1}		k_{pj}		k_{pq}	k_{p+}
Total	k_{+1}		k_{+j}		k_{+q}	n

Figure 1. Tableau des contingences.

- k_{ij} represente le ***nombre d'individus*** ayant simultanément la modalité i de X et la modalité j de Y.
- $\mathbf{k_{i+}}$ et k_{+j} sont appelés les ***coefficients marginaux***. Ils sont définis pour tout $i = 1,2, \dots, p$ et $j = 1,2, \dots, q$ par :

$$k_{i+} = \sum_{l=1}^{q} k_{il}, \quad k_{+j} = \sum_{l=1}^{p} k_{lj}.$$

- De plus, la somme des $p \times q$ observations n'est que le nombre d'individus :

$$\sum_{i,j} k_{ij} = \sum_{i=1}^{p} \sum_{j=1}^{q} k_{ij} = n.$$

Analyse des données

Remarques

1) Chaque individu ***représente une modalité et une seule*** de chaque variable.
2) Il ***n'y a pas d'ordre à respecter sur les modalités*** des variables.

Exemples

1) Considérons les résultats obtenus lors d'un concours organisé au sein d'une université pour les étudiants d'une filière donnée. Ces étudiants sont répartis en plusieurs sections. Nous supposons donc que nous disposons de **6** sections et **4** modules, alors les résultats peuvent être présentés par la matrice des contingences suivante :

$K =$

Modules / *Sections*	S_1	...	S_3	...	S_6	*Total*
M_1	k_{11}	...	k_{13}	...	k_{16}	k_{1+}
M_2	k_{21}		k_{23}		k_{26}	
M_3	k_{31}	...	k_{33}	...	k_{36}	k_{3+}
M_4	k_{41}	...	k_{43}	...	k_{46}	
Total	k_{+1}	...	k_{+2}	...	k_{+4}	n

Ou, k_{ij} *représente le nombre d'étudiants de la section* ***j*** *ayant le module* ***i***,

2) Supposons que nous voulons étudier la relation existante entre :
la ***couleur des cheveux*** et celle des ***yeux*** pour un ensemble d'individus.
Pour ce faire, nous pouvons présenter les résultats de cette analyse effectuée sur les individus dans un tableau (matrice) telle que par exemple :
sur la première ligne sera consacrée pour la couleur des cheveux et la première colonne sera consacrée pour la couleur des yeux.
Les modalités de la 1ere variable sont :

Noir, **Châtain**, **Roux**, **Blond**.

Les modalités de la 1ere variable sont :

Vert, **Bleu**, **Marron**, **Noisette.**

Donc, nous aurons : **4*4 = 16 relations (correspondances).**

$K =$

C. Cheveux / *C.Yeux*	*Noir*	*Châtain*	*Roux*	*Blond*	*Total*
Vert	*68*	*119*	*26*	*7*	*220*
Bleu	*15*	*54*	*14*	*10*	*93*
Marron	*5*	*29*	*14*	*16*	*64*
Noisette	*20*	*84*	*17*	*94*	*215*
Total	*108*	*286*	*71*	*127*	**592**

Analyse des données

3. Transformation des données

Pour réaliser les buts de l'analyse factorielle des correspondances (AFC), nous appliquons tout d'abord quelques transformations linéaires sur la matrice des données initiale.

i) Le premier traitement consiste à transformer la matrice K en une autre matrice F dite matrice ***des fréquences relatives***. Cette matrice est obtenue en divisant les coefficients de la matrice initiale par le nombre d'individus n.
C'est-à-dire :

$$K \longrightarrow F$$
$$k_{ij} \rightarrow f_{ij}$$

Avec pour tout $i = 1, 2, \dots, p$ et $j = 1, 2, \dots, q$, nous avons :

$$\boldsymbol{f_{ij} = \frac{k_{ij}}{n}}.$$

Ce qui donne :

$F =$

X \ Y	y_1	…	y_j	…	y_q	*Total*
x_1	f_{11}		f_{1j}		f_{1q}	f_{1+}
⋮			⋮			
x_i	f_{i1}	…	f_{ij}	…	f_{iq}	f_{i+}
⋮			⋮			
x_p	f_{p1}		f_{pj}		f_{pq}	f_{p+}
Total	f_{+1}		f_{+j}		f_{+q}	1

Figure 2. Tableau des fréquences relatives.

Dans ce cas, les fréquences marginales sont définies pour tout $i = 1, 2, \dots, p$ et pour tout $j = 1, 2, \dots, q$ par :

$$f_{i+} = \sum_{l=1}^{q} f_{il}, \quad f_{+j} = \sum_{l=1}^{p} f_{lj}.$$

De plus, elles vérifient :

$$\sum_{i=1}^{p} f_{i+} = 1, \quad \sum_{j=1}^{q} f_{+j} = 1.$$

Remarque

La matrice des fréquences relatives peut être interprétée par la matrice des probabilités associées a la matrice de contingences.

ii) Extraire de F ***la matrice des profils lignes*** (P.L), notée $\boldsymbol{X_L} = (\boldsymbol{X_L})_{ij}$, telle que pour tout $i = 1, 2, \dots, p$ et $j = 1, 2, \dots, q$, le coefficient $(\boldsymbol{X_L})_{ij}$ est défini comme suit :

Analyse des données

$$(X_L)_{ij} = \frac{f_{ij}}{f_{i+}}.$$

Par conséquent, le i^{ème} profil ligne n'est que la i^{ème} nouvelle ligne de la matrice des fréquences relatives, il est défini pour tout $i = 1, 2, \dots, p$ par :

$$(X_L)_i = \left(\frac{f_{i1}}{f_{i+}}, \frac{f_{i2}}{f_{i+}}, \dots, \frac{f_{ij}}{f_{i+}}, \dots, \frac{f_{iq}}{f_{i+}}\right).$$

D'où,

$X_L =$

X \ Y	y_1	…	y_j	…	y_q	
x_1	$\frac{f_{11}}{f_{1+}}$		$\frac{f_{1j}}{f_{1+}}$		$\frac{f_{1q}}{f_{1+}}$	
⋮			⋮			
x_i	$\frac{f_{i1}}{f_{i+}}$	…	$\frac{f_{ij}}{f_{i+}}$	…	$\frac{f_{iq}}{f_{i+}}$	$i^{ème}$ *ligne*
⋮			⋮			
x_p	$\frac{f_{p1}}{f_{p+}}$		$\frac{f_{pj}}{f_{p+}}$		$\frac{f_{pq}}{f_{p+}}$	

Figure 3. Présentation des coordonnées du i^{ème} point dans $\Re^q$.

Chaque coefficient de la matrice des profils ligne est défini comme suit :

$$(X_L)_{ij} = \frac{f_{ij}}{f_{i+}}.$$

iii) Extraire de F, la matrice des profils colonnes (P.C) notée $X_C = (X_C)_{ij}$. Pour tout $i = 1, 2, \dots, p$ et $j = 1, 2, \dots, q$, nous avons :

$$(X_C)_{ij} = \frac{f_{ij}}{f_{+j}}.$$

Par suite, le j^{ème} profil colonne est donné par :

$$(X_C)_j = \left(\frac{f_{1j}}{f_{+j}}, \frac{f_{2j}}{f_{+j}}, \dots, \frac{f_{ij}}{f_{+j}}, \dots, \frac{f_{pj}}{f_{+j}}\right)^t.$$

D'où,

Analyse des données

$X_C =$

X \ Y	y_1	…	y_j	…	y_q
			$j^{ème}$ colonne		
x_1	$\frac{f_{11}}{f_{+1}}$		$\frac{f_{1j}}{f_{+j}}$		$\frac{f_{1q}}{f_{+q}}$
⋮			⋮		
x_i	$\frac{f_{i1}}{f_{+1}}$	…	$\frac{f_{ij}}{f_{+j}}$	…	$\frac{f_{iq}}{f_{+q}}$
⋮			⋮		
x_p	$\frac{f_{p1}}{f_{+1}}$		$\frac{f_{pj}}{f_{+j}}$		$\frac{f_{pq}}{f_{+q}}$

Figure 4. Présentation des coordonnées du j^{ème} point dans $\Re^p$.

Remarque
La réalisation d'une AFC n'est qu'une ***double analyse en composantes principales (ACP)*** appliquée sur les deux matrices X_L et X_C extraites de la matrice des fréquences relatives F avec des ***distances particulières***.

4. Mesure de similarité

Pour mesurer la ressemblance ou bien la proximité entre deux profils, lignes ou colonnes, nous appliquons la distance dite ***distance*** ou ***test du Khi-deux*** χ^2**.** Il est défini comme suit :

i) Profils lignes

Soient $(X_L)_k$ et $(X_L)_l$ deux points de $\Re^q$ c'est-à-dire deux profils lignes :

$$(X_L)_k = \left(\frac{f_{k1}}{f_{k+}}, \frac{f_{k2}}{f_{k+}}, \dots, \frac{f_{kj}}{f_{k+}}, \dots, \frac{f_{kq}}{f_{k+}}\right).$$

Et

$$(X_L)_l = \left(\frac{f_{l1}}{f_{l+}}, \frac{f_{l2}}{f_{l+}}, \dots, \frac{f_{lj}}{f_{l+}}, \dots, \frac{f_{lq}}{f_{l+}}\right).$$

Alors :

$$d^2_{\chi^2}((X_L)_k, (X_L)_l) = \sum_{j=1}^{q} \frac{1}{f_{+j}} \left(\frac{f_{kj}}{f_{k+}} - \frac{f_{lj}}{f_{l+}}\right)^2.$$

Cette proximité ou bien cette distance peut être exprimée moyennant la distance Euclidienne. En effet,

$$d^2_{\chi^2}((X_L)_k,(X_L)_l) = \sum_{j=1}^{q} \frac{1}{f_{+j}}\left(\frac{f_{kj}}{f_{k+}} - \frac{f_{lj}}{f_{l+}}\right)^2 = \sum_{j=1}^{q} \frac{1}{\left(\sqrt{f_{+j}}\right)^2}\left(\frac{f_{kj}}{f_{k+}} - \frac{f_{lj}}{f_{l+}}\right)^2.$$

Ce qui est équivalent à :

$$d^2_{\chi^2}((X_L)_k,(X_L)_l) = \sum_{j=1}^{q} \left(\frac{f_{kj}}{\sqrt{f_{+j}} \times f_{k+}} - \frac{f_{lj}}{\sqrt{f_{+j}} \times f_{l+}}\right)^2.$$

Cette dernière formule n'est que la ***distance Euclidienne***.

C'est-à-dire :

$$d^2_{\chi^2}((X'_L)_k,(X'_L)_l) = \left\|(X'_L)_k - (X'_L)_l\right\|^2.$$

Où,

X'_L est la transformée de la matrice X_L dont le terme général est défini pour tout $i = 1, 2, \dots, p$ et $j = 1, 2, \dots, q$ par :

$$(X'_L)_{ij} = \frac{f_{ij}}{\sqrt{f_{+j}} \times f_{i+}} = \frac{1}{\sqrt{f_{+j}}}\left(\frac{f_{ij}}{f_{i+}}\right).$$

ii) Profils colonnes

De même, le test Khi-deux peut être appliqué sur les profils colonnes.

Soient $(X_C)_j$ et $(X_C)_k$ deux points de $\Re^p$ c'est-à-dire deux profils colonnes :

$$(X_C)_j = \left(\frac{f_{1j}}{f_{+j}}, \frac{f_{2j}}{f_{+j}}, \dots, \frac{f_{ij}}{f_{+j}}, \dots, \frac{f_{pj}}{f_{+j}}\right)^t.$$

Et

$$(X_C)_k = \left(\frac{f_{1k}}{f_{+k}}, \frac{f_{2k}}{f_{+k}}, \dots, \frac{f_{ik}}{f_{+k}}, \dots, \frac{f_{pk}}{f_{+k}}\right)^t.$$

Alors :

$$d^2_{\chi^2}((X_C)_j,(X_C)_k) = \sum_{j=1}^{q} \frac{1}{f_{i+}}\left(\frac{f_{ij}}{f_{+j}} - \frac{f_{ik}}{f_{+k}}\right)^2.$$

En suivant un raisonnement analogue a , nous obtenons :

$$d^2_{\chi^2}((X_C)_j,(X_C)_k) = \sum_{j=1}^{q} \left(\frac{f_{ij}}{\sqrt{f_{i+}} \times f_{+j}} - \frac{f_{ik}}{\sqrt{f_{i+}} \times f_{+k}}\right)^2.$$

Ce qui est équivalent à :

$$d^2_{\chi^2}((X'_C)_j,(X'_C)_k) = \left\|(X'_C)_j - (X'_C)_k\right\|^2.$$

Où,

X'_C est la transformée de la matrice X_C dont le terme général est défini pour tout $i = 1, 2, \dots, p$ et $j = 1, 2, \dots, q$ par :

$$(X'_C)_{ij} = \frac{f_{ij}}{\sqrt{f_{i+}} \times f_{+j}} = \frac{1}{\sqrt{f_{i+}}}\left(\frac{f_{ij}}{f_{+j}}\right).$$

5. Nuages et représentation des deux profils

La présentation des deux matrices : matrice des profils lignes et celle des profils colonnes nous conduit à considérer deux nuages de points : nuage des points profils lignes et celui des profils colonnes.

i) Nuage des profils lignes

Quand nous étudions les modalités de la première variable, nous considérons les données (individus) comme étant les profils lignes, pondérées par les fréquences marginales des lignes de la matrice F. Ainsi, pour p lignes de X_L, nous associons le nuage $\boldsymbol{N(I)}$ défini par :

$$\boldsymbol{N(I) = \{((X_L)_i, f_{i+}), i = 1, 2, \dots, p\} \subset \mathfrak{R}^q.}$$

Où,

$\boldsymbol{f_{i+}}$ représente le poids associé au i^{ème} profile ligne $(X_L)_i$ pour tout $i = 1, 2, \dots, p$.

ii) Nuage des profils colonnes

Aux q colonnes de X_C c'est-à-dire aux modalités de la variable Y, nous associons le nuage $N(J)$ défini par :

$$\boldsymbol{N(J) = \{((X_C)_j, f_{+j}), j = 1, 2, \dots, q\} \subset \mathfrak{R}^p.}$$

Où,

$\boldsymbol{f_{+j}}$ n'est que le poids associé au j^{ème} profile colonne $(X_C)_j$ pour tout $j = 1, 2, \dots, q$.

6. Ajustement des deux nuages

6.1 Ajustement du nuage des profils lignes (PL)

Le but de cette étude est de réaliser une ACP sur le nuage des profils lignes afin de résumer les liaisons entre les modalités de la première variable. Il s'agit donc de considérer les profils lignes en tant qu'individus.

Afin de suivre les mêmes étapes vues dans le chapitre précèdent, nous considérons le nuage des profils lignes transformés dans l'espace $\mathfrak{R}^q$ muni de la distance Euclidienne. Par conséquent, nous avons :

i) Centrage du nuage

Soit le nuage des profils lignes transformé $\boldsymbol{N'(I)}$ défini par :

$$\boldsymbol{N'(I) = \{((X'_L)_i, f_{i+}), i = 1, 2, \dots, p\} \subset \mathfrak{R}^q.}$$

Avec :

Analyse des données

$$(X'_L)_i = \left(\frac{f_{i1}}{f_{i+}\cdot\sqrt{f_{+1}}}, \frac{f_{i2}}{f_{i+}\cdot\sqrt{f_{+2}}}, \dots, \frac{f_{ij}}{f_{i+}\cdot\sqrt{f_{+j}}}, \dots, \frac{f_{iq}}{f_{i+}\cdot\sqrt{f_{+q}}}\right).$$

Alors, le centre de gravité de ce nuage transformé noté g_I est un point de $\Re^q$ dont sa j^{ème} composante est donnée pour tout $j = 1,2,\dots,q$ par :

$$(g_I)_j = \sum_{i=1}^{p} f_{i+} \times \frac{1}{\sqrt{f_{+j}}}\left(\frac{f_{ij}}{f_{i+}}\right) = \frac{1}{\sqrt{f_{+j}}}\sum_{i=1}^{p} f_{ij} = \frac{1}{\sqrt{f_{+j}}}\cdot f_{+j} = \sqrt{f_{+j}}.$$

Donc, le centre de gravité de $N'(I)$ est donné par :

$$g_I = \left(\sqrt{f_{+1}}, \sqrt{f_{+2}}, \dots, \sqrt{f_{+j}}, \dots, \sqrt{f_{+q}}\right) \in \Re^q.$$

Par suite, la matrice centrée des ***profils-lignes transformés et centrés***, notée Y_L est définie par son terme général $(Y_L)_{ij}$ donné pour tout $i = 1,2,\dots,p$ et pour tout $j = 1,2,\dots,q$ par :

$$(Y_L)_{ij} = \frac{f_{ij}}{f_{i+}\cdot\sqrt{f_{+j}}} - \sqrt{f_{+j}} = \frac{1}{\sqrt{f_{+j}}}\left(\frac{f_{ij}}{f_{i+}}\right) - \sqrt{f_{+j}}.$$

Autrement dit,

le i^{ème} profil ligne transformé centré est défini pour tout $i = 1,2,\dots,p$ par :

$$(Y_L)_i = \left(\frac{f_{i1}}{f_{i+}\cdot\sqrt{f_{+1}}} - \sqrt{f_{+1}}, \frac{f_{i2}}{f_{i+}\cdot\sqrt{f_{+2}}} - \sqrt{f_{+2}}, \dots, \frac{f_{iq}}{f_{i+}\cdot\sqrt{f_{+q}}} - \sqrt{f_{+q}}\right).$$

ii) Matrice d'inertie

La ***matrice d'inertie*** du nuage des profils lignes ***transformées et centrées*** n'est que la matrice ***variances-covariances*** entre les différentes modalités de la première variable.

Comme chaque profil-ligne a un poids spécifique pour lui, alors nous définissons la matrice des poids notée D_I comme suit :

$$D_I = \begin{pmatrix} f_{1+} & 0 & \cdots & 0 \\ 0 & f_{2+} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & f_{p+} \end{pmatrix}.$$

Par suite, la matrice d'inertie est donnée par :

$$V_I = Y_L^t \cdot D_I \cdot Y_L.$$

Ce qui donne :

Pour tout $i,j = 1,2,\dots,q$:

$$v_{ij} = Cov((Y_L)_i, (Y_L)_j).$$

Et pour tout $j = 1,2,\dots,q$:

$$v_{jj} = Var(Y_L)_j.$$

Par conséquent, l'inertie totale du nuage des points profils lignes est donnée par :

$$I = \sum_{j=1}^{q} v_{jj} = Trace(V_I).$$

iii) Analyse en composantes principales

Une fois la matrice d'inertie déterminée, nous pouvons déterminer le meilleur sous espace de dimension s ajustant le nuage de points centré. Pour ce faire, il suffit juste de suivre la procédure présentée et expliquée dans le chapitre précédent pour réaliser une ACP.

Donc, il s'agit de déterminer les **s vecteurs normés générateurs des axes principaux** constituant le sous espace factoriel ajustant le nuage de points.

Ces axes principaux passent par le centre de gravité et leurs vecteurs générateurs ne sont les vecteurs propres normés associés aux s plus grandes valeurs propres de la matrice d'inertie V_I. De plus ces axes sont orthogonaux deux à deux.

6.2 Ajustement du nuage des profils colonnes (PC)

D'une manière duale, nous pouvons appliquer une analyse en composantes principales (ACP) sur le nuage des modalités de la deuxième variable. Il suffit de considérer le nuage des profils-colonnes transformé :

$$N'(J) = \{((X'_C)_j, f_{+j}), j = 1, 2, \dots, q\} \subset \Re^p.$$

Avec :

$$(X'_C)_j = \left(\frac{f_{1j}}{f_{+j}\cdot\sqrt{f_{1+}}}, \frac{f_{2j}}{f_{+j}\cdot\sqrt{f_{2+}}}, \dots, \frac{f_{ij}}{f_{+j}\cdot\sqrt{f_{i+}}}, \dots, \frac{f_{pj}}{f_{+j}\cdot\sqrt{f_{p+}}}\right)^t.$$

Alors, le centre de gravité de ce nuage transformé noté g_J est un point de $\Re^p$ dont sa i^{ème} composante est donnée pour tout $i = 1,2,\dots,p$ par :

$$(g_J)_i = \sum_{j=1}^{q} f_{+j} \times \frac{1}{\sqrt{f_{i+}}}\left(\frac{f_{ij}}{f_{+j}}\right) = \frac{1}{\sqrt{f_{i+}}}\sum_{j=1}^{q} f_{ij} = \frac{1}{\sqrt{f_{i+}}}\cdot f_{i+} = \sqrt{f_{i+}}.$$

Donc, le centre de gravité de $N'(J)$ est donné par :

$$g_J = \left(\sqrt{f_{1+}}, \sqrt{f_{2+}}, \dots, \sqrt{f_{i+}}, \dots, \sqrt{f_{p+}}\right) \in \Re^p.$$

Par suite,

la matrice centrée des ***profils-lignes transformés et centrés***, notée Y_L est définie par son terme général $(Y_C)_{ij}$ donné pour tout $i = 1, 2, \dots, p$ et pour tout $j = 1, 2, \dots, q$ par :

$$(Y_C)_{ij} = \frac{f_{ij}}{f_{+j}\cdot\sqrt{f_{i+}}} - \sqrt{f_{i+}} = \frac{1}{\sqrt{f_{i+}}}\left(\frac{f_{ij}}{f_{+j}}\right) - \sqrt{f_{i+}}.$$

Autrement dit,

le j[ème] profil colonne transformé centré est défini pour tout $j = 1, 2, \dots, q$ par :

$$(\boldsymbol{Y_C})_j = \left(\frac{\boldsymbol{f_{1j}}}{\boldsymbol{f_{+j}} \cdot \sqrt{\boldsymbol{f_{1+}}}} - \sqrt{\boldsymbol{f_{1+}}}, \dots, \frac{\boldsymbol{f_{ij}}}{\boldsymbol{f_{+j}} \cdot \sqrt{\boldsymbol{f_{i+}}}} - \sqrt{\boldsymbol{f_{i+}}}, \dots, \frac{\boldsymbol{f_{pj}}}{\boldsymbol{f_{+j}} \cdot \sqrt{\boldsymbol{f_{p+}}}} - \sqrt{\boldsymbol{f_{p+}}} \right)^{\boldsymbol{t}}.$$

Par suite, la matrice d'inertie est donnée par :

$$\boldsymbol{V_J = Y_C \cdot D_J \cdot Y_C^t}.$$

Et la matrice des poids est définie par :

$$D_J = \begin{pmatrix} \boldsymbol{f_{+1}} & 0 & \cdots & 0 \\ 0 & \boldsymbol{f_{+2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{f_{+q}} \end{pmatrix}.$$

Remarque

Nous pouvons récapituler les différentes étapes dans la réalisation d'une AFC dans le diagramme suivant :

Analyse des données

Tableau de contingence (K)

j^{ème} colonne

i^{ème} ligne

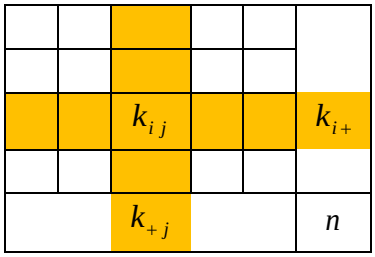
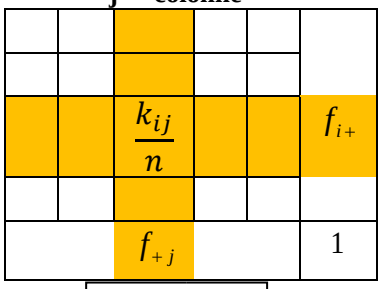


Tableau des fréquences relatives (F)

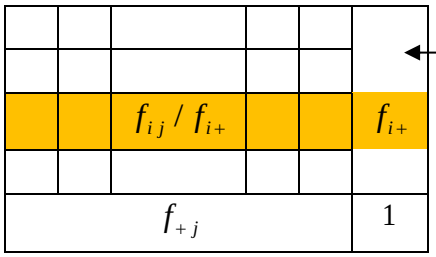
j^{ème} colonne

i^{ème} ligne

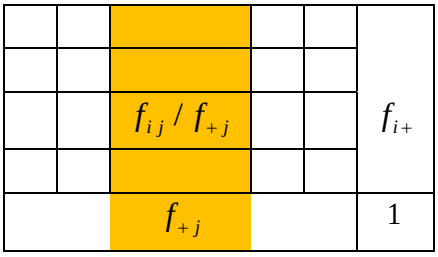
$f_{ij} = k_{ij}/n$



Nuage des profils-lignes



Nuage des profils- colonnes



(X_L)

Nuage des p points dans $\Re^q$

$$g_I = \left(\sqrt{f_{+1}}, \sqrt{f_{+2}}, ..., \sqrt{f_{+j}}, ..., \sqrt{f_{+q}}\right) \in \Re^q$$

(X_C)

Nuage des q points dans $\Re^p$

$$g_J = {}^t\left(\sqrt{f_{1+}}, \sqrt{f_{2+}}, ..., \sqrt{f_{i+}}, ..., \sqrt{f_{p+}}\right) \in \Re^p$$

Les profils transformés

$$(X'_L)_i = \left(\frac{f_{i1}}{f_{i+} \cdot \sqrt{f_{+1}}}, \frac{f_{i2}}{f_{i+} \cdot \sqrt{f_{+2}}}, \dots, \frac{f_{ij}}{f_{i+} \cdot \sqrt{f_{+j}}}, ..., \frac{f_{iq}}{f_{i+} \cdot \sqrt{f_{+q}}}\right)$$

$$(\mathbf{X'_C})_j = \left(\frac{\mathbf{f_{1j}}}{\mathbf{f_{+j}} \cdot \sqrt{\mathbf{f_{1+}}}}, \frac{\mathbf{f_{2j}}}{\mathbf{f_{+j}} \cdot \sqrt{\mathbf{f_{2+}}}}, \dots, \frac{\mathbf{f_{ij}}}{\mathbf{f_{+j}} \cdot \sqrt{\mathbf{f_{i+}}}}, \dots, \frac{\mathbf{f_{pj}}}{\mathbf{f_{+j}} \cdot \sqrt{\mathbf{f_{p+}}}}\right)^{\mathbf{t}}$$

Analyse des données

Exemple

On considère la matrice de contingences suivante :

$$K = \begin{pmatrix} 2 & 2 \\ 1 & 3 \\ 3 & 1 \end{pmatrix}.$$

1) Donner le tableau des fréquences relatives.
2) Donner le nuage des profils-lignes $N(I)$.
3) Déterminer la matrice des variances-covariances associée au nuage $N(I)$ transformé.
4) Déterminer le sous espace de dimension 1 ajustant le nuage des profils-lignes.
5) Que constatez-vous ?
6) Calculer la proximité entre les profils-lignes 1 et 2.

Solution

De la matrice des données, le nombre des individus est égal à $n = \sum_{i,j} k_{ij} = 12$.

1) Le tableau des fréquences relatives est donné par :

F =

Y / X			f_{i+}
	1/6	1/6	1/3
	1/12	1/4	1/3
	1/4	1/12	1/3
f_{+j}	1/2	1/2	1

fij= kij/12

2) Le nuage des profils-lignes est défini par :

$$N(I) = \left\{ \left((X_L)_i, f_{i+} \right), i = 1,2,3 \right\} \subset \Re^2.$$

Où

$$(X_L)_i = \left(\frac{f_{i1}}{f_{i+}}, \frac{f_{i2}}{f_{i+}} \right) \text{ pour tout } i = 1,2,3.$$

Ce qui donne :

$$X_L = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \\ 3/4 & 1/4 \end{pmatrix}.$$

3) La matrice des variances-covariances est associée au nuage des profils transformé centré. Pour ce faire, calculons le vecteur profil moyen qui représente le centre de gravité du nuage et déterminons le nuage des profils transformé.

En effet,

$$g_I = \left(\sqrt{f_{+1}}, \sqrt{f_{+2}} \right) \in \Re^2 \quad \text{c'est-à-dire} \quad g_I = \left(\sqrt{1/2}, \sqrt{1/2} \right) = \left(1/\sqrt{2}, 1/\sqrt{2} \right).$$

Analyse des données

D'autre part, le profil-ligne transformé est défini par :

$$(X'_L)_i = \left(\frac{f_{i1}}{f_{i+} \cdot \sqrt{f_{+1}}}, \frac{f_{i2}}{f_{i+} \cdot \sqrt{f_{+2}}} \right) \text{ pour tout } i = 1,2,3.$$

Ce qui donne,

$$X'_L = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ \sqrt{2}/4 & 3\sqrt{2}/4 \\ 3\sqrt{2}/4 & \sqrt{2}/4 \end{pmatrix}.$$

Par suite, le nuage centré est défini par :

$$(Y_L)_i = \left(\frac{f_{i1}}{f_{i+} \cdot \sqrt{f_{+1}}} - \sqrt{f_{+1}} \ , \ \frac{f_{i2}}{f_{i+} \cdot \sqrt{f_{+2}}} - \sqrt{f_{+2}} \right) \text{ pour tout } i = 1,2,3.$$

D'où

$$Y_L = \begin{pmatrix} 0 & 0 \\ -\sqrt{2}/4 & \sqrt{2}/4 \\ \sqrt{2}/4 & -\sqrt{2}/4 \end{pmatrix}.$$

Sachant que la matrice des poids associés aux profils-lignes est définie par :

$$D_I = \begin{pmatrix} f_{1+} & 0 & 0 \\ 0 & f_{2+} & 0 \\ 0 & 0 & f_{3+} \end{pmatrix} \text{ c'est à dire } D_I = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/3 \end{pmatrix}.$$

Alors, la matrice d'inertie est définie par :

$$\boldsymbol{V_I = Y_L^t \cdot D_I \cdot Y_L.}$$

Nous remarquons que pour cet exercice, les fréquences marginales des profils-lignes c'est-à-dire les poids sont tous identiques et sont égales à 1/3 et par suite la matrice des poids n'est que :

$$D_I = \frac{1}{3} \cdot Id \ .$$

Par conséquent, la matrice d'inertie devient :

$$V_I = \frac{1}{3} \cdot ({}^t Y_L \cdot Y_L).$$

Après calcul, nous obtenons :

$$V_I = \begin{pmatrix} 1/12 & -1/12 \\ -1/12 & 1/12 \end{pmatrix}.$$

Analyse des données

4) Pour ajuster le nuage des points, nous déterminons tout d'abord les valeurs propres qui reflètent la quantité d'information portée par les axes correspondant aux valeurs propres. En effet,

$P_{V_I}(\lambda) = \det(V_I - \lambda \cdot Id) = 0$ est équivalent à

$$\begin{vmatrix} (1/12)-\lambda & -1/12 \\ -1/12 & (1/12)-\lambda \end{vmatrix} = \left(\frac{1}{12}-\lambda\right)^2 - \left(\frac{1}{12}\right)^2 = 0$$

Ce qui donne :

$$\lambda_1 = \frac{1}{12} - \frac{1}{12} = 0 \qquad \text{et} \qquad \lambda_2 = \frac{1}{12} + \frac{1}{12} = \frac{1}{6}.$$

De ces valeurs propres, nous pouvons exprimer le taux d'inertie expliqué par chaque axe :
Comme le 1èr axe correspond toujours à la plus grande valeur propre, alors le taux expliqué par le 1er axe est défini par :

$$T_1 = \frac{MAX|\lambda_i|}{Trace(V_I)} = \frac{1/6}{1/12+1/12} = \frac{1/6}{1/6} = 1.$$

C'est-à-dire

$$\boldsymbol{T_1 = 100\%.}$$

Le taux expliqué par le deuxième axe correspondant à la deuxième valeur propre est donné par :

$$T_2 = \frac{\lambda_2}{Trace(V_I)} = \frac{0}{1/12+1/12} = 0.$$

C'est-à-dire

$$\boldsymbol{T_2 = 0\%.}$$

Donc, toute l'information réside dans le 1er axe. Ce qui explique que nous vous demandons de déterminer le sous espace de dimension 1 ajustant le nuage.
Pour déterminer cet axe, calculons le vecteur propre normé associé à la valeur propre : $\lambda_2 = \frac{1}{6}$.

Soit $u_1 = {}^t(x, y) \in \Re_*^2$ tel que : $V_I \cdot u_1 = \lambda_2 \cdot u_1$, alors :

$$\begin{cases} (1/12)\cdot x - (1/12)\cdot y = (1/6)\cdot x \\ (-1/12)\cdot x + (1/12)\cdot y = (1/6)\cdot y \end{cases} \Leftrightarrow \begin{cases} (-1/12)\cdot x = (1/12)\cdot y \\ (-1/12)\cdot x = (1/12)\cdot y \end{cases}.$$

Ce qui donne :

$$y = -x.$$

Par suite, le vecteur propre s'écrit comme suit :

Analyse des données

$$u_1 = {}^t(x, -x) \text{ avec } x \in \Re^*.$$

Soit alors, $u_1 = {}^t(1, -1)$. Par suite, l'axe représentatif est donné par le vecteur

$$u = {}^t(1/\sqrt{2}, \quad -1/\sqrt{2}).$$

Par conséquent :

C1 = Y*u

5) La proximité entre les deux profils-lignes $(X_L)_1$ et $(X_L)_2$ est donnée moyennant la ***distance khi-deux*** comme suit :

$$d^2_{\chi^2}\left((X_L)_1, (X_L)_2\right) = \sum_{j=1}^{2} \frac{1}{f_{+j}} \left(\frac{f_{1j}}{f_{1+}} - \frac{f_{2j}}{f_{2+}} \right)^2.$$

Ou bien, en utilisant la ***distance Euclidienne*** sur le nuage transformé :

$$d^2\left((X'_L)_1, (X'_L)_2\right) = \left(\frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{4} \right)^2 + \left(\frac{\sqrt{2}}{2} - \frac{3\sqrt{2}}{4} \right)^2 = 1/4.$$

Par suite, la proximité entre les deux premièrs profils-lignes est **1/2**.