Analyse Factorielle des Correspondances (AFC)

Abdelhakim Necir

Département de Mathématiques
Université de Biskra

Master 1, 2021-2022

Exemple

Le tableau suivant représente la couleur des cheveux et la couleur des yeux dans un échantillon de **370** individus.

	V_2			
	$\downarrow$			
V_1	1:Brun	2:Châtain	3:Roux	4:Blond
$\downarrow$				
1:Marron	68	119	26	7
2:Noisette	15	54	14	10
3:Vert	4	29	14	10

- V_1 :=(marron,noisette,vert)$\longrightarrow$ variables catégorielle (qualitative)
- V_2 :=(brun,chatain,roux,blond)$\longrightarrow$ variables catégorielle (qualitative)
  Marron $\longrightarrow$ une modalité de V_1 et Blond $\longrightarrow$ une modalité de V_2, ...

Effectif de $(1,1) = 68$ et effectif de $(2,3) = 14$, ...

Tableau de contingence: variables qualitatives, modalités

Notons:

- $V_1 \longrightarrow$ variable qualitative à p modalités
- $V_2 \longrightarrow$ variable qualitative à q modalités

Le tableau de contingence N^* obtenu en croisant les deux variables V_1 et V_2. Plus précisément, on a :

$$N^* = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pq} \end{pmatrix},$$

où

$x_{ij} \longrightarrow$ le nombre d'observations (ou effectif) pour lesquelles

$$V_1 = i \text{ et } V_2 = j.$$

Tableau de contingence: variables qualitatives, modalités

Dans notre exemple (couleur des yeux vs couleur des cheveux), on a

$$N^* = \begin{pmatrix} 68 & 119 & 26 & 7 \\ 15 & 54 & 14 & 10 \\ 4 & 29 & 14 & 10 \end{pmatrix}.$$

Tableaux des fréquences

On définit l'effectif total par

$$n := \sum_{i=1}^{p} \sum_{j=1}^{q} x_{ij}.$$

La fréquence observée du croisement des deux modalités (i, j), est définie par

$$f_{ij} = \frac{x_{ij}}{n} = \mathbf{P}(V_1 = i, V_2 = j).$$

Ainsi, on définit le tableau des **fréquences observées** par

$$N := \begin{pmatrix} f_{11} & \cdots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pq} \end{pmatrix} = \frac{1}{n} N^*.$$

Tableaux des fréquences

On définit, respectivement, les effectifs marginaux des lignes et des colonnes, par

$$x_{i\cdot} := \sum_{j=1}^{q} x_{ij} \text{ et } x_{\cdot j} := \sum_{i=1}^{p} x_{ij}.$$

De même, on définit, respectivement, les fréquences marginales des lignes et des colonnes, par

$$f_{i\cdot} := \sum_{j=1}^{q} f_{ij} = \mathbf{P}\left(V_1 = i\right) \text{ et } f_{\cdot j} := \sum_{i=1}^{p} f_{ij} = \mathbf{P}\left(V_2 = j\right).$$

Il est important de souligner que

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = \sum_{i=1}^{p} f_{i\cdot} = \sum_{j=1}^{q} f_{\cdot j} = 1.$$

Tableaux des fréquences

On définit les fréquences conditionnelles associées aux profils-lignes par

$$f_{j/i} := \mathbf{P}\left(V_2 = j \mid V_1 = i\right) = \frac{\mathbf{P}\left(V_2 = j, V_1 = i\right)}{\mathbf{P}\left(V_1 = i\right)} = \frac{f_{ij}}{f_{i\cdot}}.$$

En outre, on définit les fréquences conditionnelles associées aux profils-colonnes, par

$$f_{i/j} := \mathbf{P}\left(V_1 = i \mid V_2 = j\right) = \frac{\mathbf{P}\left(V_1 = i, V_2 = j\right)}{\mathbf{P}\left(V_2 = j\right)} = \frac{f_{ij}}{f_{\cdot j}}.$$

Tableaux des fréquences

On note aussi que

$$\sum_{j=1}^{q} f_{j/i} = 1, \text{ pour chaque } i = 1, ..., p,$$

et

$$\sum_{i=1}^{p} f_{i/j} = 1, \text{ pour chaque } j = 1, ..., q.$$

Test d'indépendance du Khi-deux

- A priori nous n'avons aucune information sur la dépendance entre les deux variables V_1 et V_2.
- Il se peu que ces deux dérnières sont indépendantes et par conséquent l'AFC est inutile.

On peu traduire ceci, du fait que sous l'hypothèse d'indépendance, H_0, on a

$$\mathbf{P}\left(V_1 = i, V_2 = j\right) = \mathbf{P}\left(V_1 = i\right)\mathbf{P}\left(V_2 = j\right).$$

Contrairement à l'hypothèse de dépendance H_1, on a

$$\mathbf{P}\left(V_1 = i, V_2 = j\right) \neq \mathbf{P}\left(V_1 = i\right)\mathbf{P}\left(V_2 = j\right).$$

En d'autres termes, on affaire à tester l'hypothèse nulle

$$H_0 : \ f_{ij} = f_{i.}f_{.j}$$

contre l'hypothèse alternative

$$H_1 : \ f_{ij} \neq f_{i.}f_{.j}$$

Test d'indépendance du Khi-deux

Donc l'hypothèse d'indépendance, H_0 est une l'hypothèse théorique que nous devons la vérifier par une règle décision. Donc, la quantité

$$\widetilde{f}_{ij} := f_{i}.f_{\cdot j}$$

peut être vue comme **la fréquence théorique**. Ainsi nous définissons le tableau des fréquences théoriques par

$$\widetilde{N} = \begin{pmatrix} f_{1}.f_{\cdot 1} & \cdots & f_{1}.f_{\cdot q} \\ \vdots & \ddots & \vdots \\ f_{p}.f_{\cdot 1} & \cdots & f_{p}.f_{\cdot q} \end{pmatrix}.$$

Ce qui nous permet de définir **l'effectif théorique** par

$$e_{ij} := nf_{i.}f_{.j} = n\frac{x_{i.}}{n}\frac{x_{.j}}{n} = \frac{x_{i.}x_{.j}}{n}.$$

Ansi le tableau les effectifs théoriques par

$$E := \begin{pmatrix} e_{11} & \cdots & e_{1q} \\ \vdots & \ddots & \vdots \\ e_{p1} & \cdots & e_{pq} \end{pmatrix} = n\widetilde{N}.$$

Test d'indépendance du Khi-deux

Notons $\mathbf{f}_{ij}$ la variable aléatoire de la fréquence associée à la fréquence observée f_{ij}.Nous définissons **l'écart à l'indépendance** par

$$\begin{aligned}\Phi^2 &= \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(\mathbf{f}_{ij}-\mathbf{f}_{i}.\mathbf{f}_{.j}\right)^2}{\mathbf{f}_{i}.\mathbf{f}_{.j}} \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(\text{proba. conjointe} - \text{produit proba. marginales}\right)^2}{\text{produit proba. marginales}}.\end{aligned}$$

La statistique du test associée est la **statistique du khi-deux** définie par

$$\chi^2 = n\Phi^2 = n\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(\mathbf{f}_{ij}-\mathbf{f}_{i}.\mathbf{f}_{.j}\right)^2}{\mathbf{f}_{i}.\mathbf{f}_{.j}}.$$

Test d'indépendance du Khi-deux

Nous avons, quand $n \longrightarrow \infty$

$$\chi^2 \xrightarrow{D} \text{ la loi de Khi-deux à } r \text{ degrés de liberté,}$$

où

$$r := (p-1)(q-1).$$

En pratique, nous devons s'assurer que

$$n > 30 \text{ et } e_{ij} \geq 5, \text{ pour tout les } (i,j).$$

Test d'indépendance du Khi-deux

Une fois que les fréquences $\mathbf{f}_{ij}$ sont observées, elles seront notées par f_{ij}, ainsi la statistique du Khi-deux χ^2 sera aussi observée, et on écrit

$$\chi^2_{obs} = n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - f_{i}.f_{.j}\right)^2}{f_{i}.f_{.j}}.$$

Test d'indépendance du Khi-deux

Rappelons que, dans notre exemple, la matrice des données est

$$N^* = \begin{pmatrix} 68 & 119 & 26 & 7 \\ 15 & 54 & 14 & 10 \\ 4 & 29 & 14 & 10 \end{pmatrix}$$

Pour tester l'indépendance entre les deux variables V_1et V_2, il suffit d'utiliser entre autres, la fonction "`chisq.test`" du logiciel R. Voici les commandes à utiliser:

```
tab<-matrix(c(68,119,26,7,15,54,14,10,4,29,14,10)
                                        ,ncol=4,byrow=TRUE)
test=chisq.test(tab)
X-squared =34.114, df=6,p-value=6.394e-06
```

Test d'indépendance du Khi-deux

Comme

$$p - value < \alpha := 0.05,$$

alors on rejette l'hypothèse, nulle, d'association (ou d'indépendance) des deux variables V_1 et V_2.
En conclusion, il y a une liaison entre la couleur des cheveux et la couleur des yeux. L'AFC donc á un sens.

Rappelons que **l'écart à l'indépendance observé** est définit par

$$\Phi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - f_{i.}f_{.j}\right)^2}{f_{i.}f_{.j}} = \frac{\chi^2_{obs}}{n}.$$

Soient les événements suivants

$$A_i := \{V_1 = i\}, \; B_j := \{V_2 = j\}$$

et

$$A_i \cap B_j := \{V_1 = i, V_2 = j\}.$$

Nous avons

$$\mathbf{P}\left(A_i \cap B_j\right) = f_{ij}, \; \mathbf{P}\left(A_i\right) = f_{i\cdot}, \; \mathbf{P}\left(B_j\right) = f_{\cdot j}.$$

Écart à l'indépendance

Ainsi, nous avons

$$\Phi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(\mathbf{P}\left(A_i \cap B_j\right) - \mathbf{P}\left(A_i\right)\mathbf{P}\left(B_j\right)\right)^2}{\mathbf{P}\left(A_i\right)\mathbf{P}\left(B_j\right)}.$$

Rappelons que

$$\begin{aligned} A_i \text{ et } B_j \text{ sont indépendants} \quad &\Leftrightarrow \quad \mathbf{P}\left(A_i \cap B_j\right) = \mathbf{P}\left(A_i\right)\mathbf{P}\left(B_j\right) \\ &\Leftrightarrow \quad \Phi^2_{obs} = 0. \end{aligned}$$

Écart à l'indépendance

Remarquons que l'écart à l'indépendance observé peut être réécrit comme suit

$$\Phi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{i.} \frac{\left(\frac{f_{ij}}{f_{i.}} - f_{.j}\right)^2}{f_{.j}} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{i.} \frac{\underbrace{\left(f_{j/i} - f_{.j}\right)^2}}{f_{.j}}.$$

Autrement dit

$$\Phi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{q} \mathbf{P}\left(A_i\right) \frac{\underbrace{\left(\mathbf{P}\left(B_j/A_i\right) - \mathbf{P}\left(B_j\right)\right)^2}}{\mathbf{P}\left(B_j\right)}.$$

Nous avons

$$\begin{aligned} A_i \text{ et } B_j \text{ sont indépendants} \quad &\Leftrightarrow \quad \mathbf{P}\left(B_j/A_i\right) = \mathbf{P}\left(B_j\right) \\ &\Leftrightarrow \quad \Phi^2_{obs} = 0. \end{aligned}$$

Écart à l'indépendance

Par symétrie nous avons aussi

$$\Phi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{.j} \frac{\left(\dfrac{f_{ij}}{f_{.j}} - f_{i.}\right)^2}{f_{i.}} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{.j} \frac{\underbrace{\left(f_{i/j} - f_{i.}\right)^2}}{f_{i.}}.$$

Autrement dit

$$\Phi^2_{obs} = \sum_{i=1}^{p}\sum_{j=1}^{q} \mathbf{P}\left(B_j\right) \frac{\underbrace{\left(\mathbf{P}\left(A_i/B_j\right) - \mathbf{P}\left(A_i\right)\right)^2}}{\mathbf{P}\left(A_i\right)}.$$

Nous avons

$$\begin{aligned} A_i \text{ et } B_j \text{ sont indépendants} \quad &\Leftrightarrow \quad \mathbf{P}\left(A_i/B_j\right) = \mathbf{P}\left(A_i\right) \\ &\Leftrightarrow \quad \Phi^2_{obs} = 0. \end{aligned}$$

En résumé, nous avons trois variantes de l'écart à l'indépendance:

$$\Phi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - f_{i.}f_{.j}\right)^2}{f_{i.}f_{.j}},$$

$$\Phi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{i.} \frac{\left(\dfrac{f_{ij}}{f_{i.}} - f_{.j}\right)^2}{f_{.j}} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{i.} \frac{\left(f_{j/i} - f_{.j}\right)^2}{f_{.j}}$$

et

$$\Phi^2_{obs} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{.j} \frac{\left(\dfrac{f_{ij}}{f_{.j}} - f_{i.}\right)^2}{f_{i.}} = \sum_{i=1}^{p} \sum_{j=1}^{q} f_{.j} \frac{\left(f_{i/j} - f_{i.}\right)^2}{f_{i.}}$$

Tableaux des profils-lignes et profils-colonnes

On note

$$D_r := \begin{pmatrix} f_{1\cdot} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{p\cdot} \end{pmatrix} \rightarrow D_r^{-1} := \begin{pmatrix} 1/f_{1\cdot} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{p\cdot} \end{pmatrix},$$

$$N := \begin{pmatrix} f_{11} & \cdots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pq} \end{pmatrix} \rightarrow D_r^{-1}N = \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix} =: X_r$$

$X_r \longrightarrow$ le tableau des profils-lignes.

$$D_c := \begin{pmatrix} f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f_{\cdot q} \end{pmatrix} \to D_c^{-1} := \begin{pmatrix} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{pmatrix},$$

$$N^t := \begin{pmatrix} f_{11} & \cdots & f_{1q} \\ \vdots & \ddots & \vdots \\ f_{p1} & \cdots & f_{pq} \end{pmatrix} \to D_c^{-1} N^t = \begin{pmatrix} \frac{f_{11}}{f_{\cdot 1}} & \cdots & \frac{f_{1q}}{f_{\cdot 1}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{\cdot p}} & \cdots & \frac{f_{pq}}{f_{\cdot p}} \end{pmatrix} =: X_c$$

$X_c \longrightarrow$ le tableau des profils-colonnes.

Centres de gravités

On définit, le centre de gravité de ce nuage de points des profils-lignes par

$$\begin{aligned} g_r &= \left(f_{\cdot 1}, ..., f_{\cdot q}\right)^t \\ &= X_r^t D_r \mathbf{1}_p \\ &= \left(D_r^{-1} N\right)^t D_r \mathbf{1}_p = N^t \underbrace{D_r^{-1} D_r}_{} \mathbf{1}_p = N^t \mathbf{1}_p, \end{aligned}$$

où $\mathbf{1}_p = (1, ..., 1)^t$, le vecteur unitaire de $p \times 1$.

Centres de gravités

De même, on définit, leur centre de gravité du nuage des profils-colonnes par

$$\begin{aligned} g_c &= (f_{1\cdot}, ..., f_{p\cdot})^t = X_c^t D_c \mathbf{1}_q \\ &= \left(D_c^{-1} N^t\right)^t D_c \mathbf{1}_q = N \underbrace{D_c^{-1} D_c}_{} \mathbf{1}_p = N \mathbf{1}_q, \end{aligned}$$

où $\mathbf{1}_q := (1, \cdots, 1)^t$, le vecteur unitaire de $q \times 1$.

Exemple

Considérons la matrice des données de l'exemple ci-dessus

$$N^* = \begin{pmatrix} 68 & 119 & 26 & 7 \\ 15 & 54 & 14 & 10 \\ 4 & 29 & 14 & 10 \end{pmatrix}$$

Nous avons ici $p = 3$, $q = 4$ et

$$n = 68 + 15 + 4 + 119 + 54 + 29 + 26 + 14 + 14 + 7 + 10 + 10 = 370.$$

Ainsi

$$N = \begin{pmatrix} 68/370 & 119/370 & 26/370 & 7/370 \\ 15/370 & 54/370 & 14/370 & 10/370 \\ 4/370 & 29/370 & 14/370 & 10/370 \end{pmatrix}.$$

Exemple

Les fréquences marginales-lignes sont

$$
\begin{aligned}
f_{1\cdot} &= \sum_{j=1}^{4} f_{1j} = 68/370 + 119/370 + 26/370 + 7/370 = 220/370 \\
f_{2\cdot} &= \sum_{j=1}^{4} f_{2j} = 15/370 + 54/370 + 14/370 + 10/370 = 93/370 \\
f_{3\cdot} &= \sum_{j=1}^{4} f_{3j} = 4/370 + 29/370 + 14/370 + 10/370 = 57/370.
\end{aligned}
$$

Exemple

Les fréquences marginales-colonnes sont

$$
\begin{aligned}
f_{.1} &= \sum_{i=1}^{3} f_{i1} = 68/370 + 15/370 + 4/370 = 87/370 \\
f_{.2} &= \sum_{i=1}^{3} f_{2j} = 119/370 + 54/370 + 29/370 = 202/370 \\
f_{.3} &= \sum_{i=1}^{3} f_{3j} = 26/370 + 14/370 + 14/370 = 27/185 \\
f_{.4} &= \sum_{i=1}^{3} f_{4j} = 7/370 + 10/370 + 10/370 = 27/370.
\end{aligned}
$$

Exemple

Les centres de gravité des profils-linges et profils-colonnes, respectivement, sont

$$g_r = (87/370, 202/370, 27/185, 27/370)^t \, ,$$

et

$$g_c = (220/370, 93/370, 57/370)^t \, .$$

Exemple

Les matrices diagonales de profils-lignes et profils-colonnes sont respectivement

$$D_r = \begin{pmatrix} 220/370 & 0 & 0 \\ 0 & 93/370 & 0 \\ 0 & 0 & 57/370 \end{pmatrix}$$

et

$$D_c = \begin{pmatrix} 87/370 & 0 & 0 & 0 \\ 0 & 202/370 & 0 & 0 \\ 0 & 0 & 27/185 & 0 \\ 0 & 0 & 0 & 27/370 \end{pmatrix}$$

Exemple

Les matrices profils-lignes et profils-colonnes, respectivement, sont

$$X_r = D_r^{-1}N = \begin{pmatrix} \frac{17}{55} & \frac{119}{220} & \frac{13}{110} & \frac{7}{220} \\ \frac{5}{31} & \frac{18}{31} & \frac{14}{93} & \frac{10}{93} \\ \frac{4}{57} & \frac{29}{57} & \frac{14}{57} & \frac{10}{57} \end{pmatrix}$$

et

$$X_c = D_c^{-1}N^t = \begin{pmatrix} \frac{68}{87} & \frac{5}{29} & \frac{4}{87} \\ \frac{119}{202} & \frac{27}{101} & \frac{29}{202} \\ \frac{13}{27} & \frac{7}{27} & \frac{7}{27} \\ \frac{7}{27} & \frac{10}{27} & \frac{10}{27} \end{pmatrix}.$$

Rappelons que la matrice des profils-lignes est

$$X_r := \begin{array}{c} 1 \longrightarrow \\ \\ i \longrightarrow \\ \\ i' \longrightarrow \\ \\ p \longrightarrow \end{array} \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \cdots & \ddots & \vdots \\ \frac{f_{i1}}{f_{i\cdot}} & \cdots & \frac{f_{iq}}{f_{i\cdot}} \\ \vdots & \ddots & \vdots \\ \frac{f_{i'1}}{f_{i'\cdot}} & \cdots & \frac{f_{i'q}}{f_{i'\cdot}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix}.$$

Métrique du khi-deux

La distance euclidienne entre deux profils-lignes i et i' est définie par

$$d^2\left(i, i'\right) = \sum_{j=1}^{q} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 = \left\| i - i' \right\|^2.$$

La distance de khi-deux entre deux profils-lignes i et i' est

$$d_{\chi^2}^2\left(i, i'\right) = \sum_{j=1}^{q} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2,$$

La pondération par $1/f_{.j}$ à chaque carré de différence revient à donner des importances comparables aux diverses modalités j de la variable V_2. Sans cette pondération, la distance reflète surtout la différence entre les modalitliés de plus grandes fréquences.

Métrique du khi-deux

Autrement dit

$$d^2_{\chi^2}\left(i,i'\right) = \left\| i - i' \right\|^2_{M_r},$$

où

$$M_r := D_c^{-1} = \begin{pmatrix} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{pmatrix}.$$

Celle-ci peut être vu comme la distance euclidienne, entre les deux profils-lignes i et i', pondérée. Plus précisément

$$d^2_{\chi^2}\left(i,i'\right) = \left(i - i'\right)^t M_r \left(i - i'\right).$$

La distance entre un profil-ligne et son centre de gravité g_r est définie par

$$d^2_{\chi^2}\left(i,g_r\right) = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} \right)^2 = \left\| i - g_r \right\|^2_{M_r}.$$

Métrique du khi-deux

La distance entre deux profils-colonnes j et j' est

$$d_{\chi^2}^2\left(j, j'\right) = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - \frac{f_{ij'}}{f_{\cdot j'}} \right)^2 = \left\| j - j' \right\|_{M_c}^2 ,$$

où

$$M_c := D_r^{-1} = \begin{pmatrix} 1/f_{1\cdot} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{p\cdot} \end{pmatrix}.$$

La distance entre un profil-colonne et son centre gravité g_c est définie par

$$d_{\chi^2}^2\left(j, g_c\right) = \sum_{i=1}^{p} \frac{1}{f_{i\cdot}} \left(\frac{f_{ij}}{f_{\cdot j}} - f_{i\cdot} \right)^2 = \left\| j - g_c \right\|_{M_c}^2 .$$

Métrique du khi-deux

La matrice des profils-lignes est

$$X_r = \begin{pmatrix} 68/220 & 119/220 & 26/220 & 7/220 \\ 15/93 & 54/93 & 14/93 & 10/93 \\ 4/57 & 29/57 & 14/57 & 10/57 \end{pmatrix},$$

et les fréquences marginales des profils-colonnes sont

$$f_{\cdot 1} = 87/370,\ f_{\cdot 2} = 202/370,\ f_{\cdot 3} = 27/185,\ f_{\cdot 4} = 27/370.$$

La distance euclidienne entre la première et la deuxième lignes est

$$\begin{aligned} d^2(1,2) &= \left(\frac{68}{220} - \frac{15}{93}\right)^2 + \left(\frac{119}{220} - \frac{54}{93}\right)^2 \\ &\quad + \left(\frac{26}{220} - \frac{14}{93}\right)^2 + \left(\frac{7}{220} - \frac{10}{93}\right)^2 \\ &= 3.0203 \times 10^{-2}. \end{aligned}$$

La distance de chi-deux entre la première et la deuxième lignes est

$$\begin{aligned}
d_{\chi^2}^2(1,2) &= \frac{370}{87}\left(\frac{68}{220}-\frac{15}{93}\right)^2+\frac{370}{202}\left(\frac{119}{220}-\frac{54}{93}\right)^2 \\
&\quad +\frac{370}{27}\left(\frac{26}{220}-\frac{14}{93}\right)^2+\frac{370}{27}\left(\frac{7}{220}-\frac{10}{93}\right)^2 \\
&= 0.188\,69.
\end{aligned}$$

Eneties

Les inerties totales des nuages de points profils-lignes et profils-colonnes par rapport aux centres de gravité correspondants sont définies respectivement par

$$\text{Inertie}\left(X_r / g_r\right) := \sum_{i=1}^{p} f_{i.} d_{\chi^2}^2\left(i, g_r\right),$$

et

$$\text{Inertie}\left(X_c / g_c\right) := \sum_{j=1}^{q} f_{.j} d_{\chi^2}^2\left(j, g_c\right).$$

Observons que

$$
\begin{aligned}
\text{Inertie}\left(X_r/g_r\right) &= \sum_{i=1}^{p} f_{i\cdot}\left\{d_{\chi^2}^2\left(i,g_r\right)\right\}\\
&= \sum_{i=1}^{p} f_{i\cdot}\left\{\sum_{j=1}^{q}\left(\frac{1}{f_{\cdot j}}\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)^2\right)\right\}\\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} f_{i\cdot}\left(\frac{1}{f_{\cdot j}}\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)^2\right)\\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{f_{i\cdot}}{f_{\cdot j}}\left(\frac{f_{ij}}{f_{i\cdot}}-f_{\cdot j}\right)^2\\
&= \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(f_{ij}-f_{i\cdot}f_{\cdot j}\right)^2}{f_{i\cdot}f_{\cdot j}} = \Phi^2 = \frac{\chi^2}{n}.
\end{aligned}
$$

En d'autres termes, étudier l'inertie de X_r revient à étudier l'écart à l'indépendance Φ^2.

On montre aussi que

$$\text{Inertie}\left(X_r\,/\,g_r\right) = \text{Inertie}\left(X_c\,/\,g_c\right) = \Phi^2 = \frac{\chi^2}{n}.$$

Pour notre exemple on a

$$X_r = \begin{pmatrix} 68/220 & 119/220 & 26/220 & 7/220 \\ 15/93 & 54/93 & 14/93 & 10/93 \\ 4/57 & 29/57 & 14/57 & 10/57 \end{pmatrix}.$$

Donc

$$\text{Inertie}\left(X_r/g_r\right) = \sum_{i=1}^{3} f_i.d^2_{\chi^2}\left(i, g_r\right) = \frac{\chi^2}{370} = \frac{34.114}{370} = 0.092\,2 = 9.22\%$$

On en déduit que

$$\text{Inertie}\left(X_c/g_c\right) = 9.22\%.$$

ACP des deux nuages de profils

Rappelons que la matrice des profils-lignes est définie par

$$X_r = D_r^{-1} N = \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} & \cdots & \frac{f_{1q}}{f_{1\cdot}} \\ \vdots & \ddots & \vdots \\ \frac{f_{p1}}{f_{p\cdot}} & \cdots & \frac{f_{pq}}{f_{p\cdot}} \end{pmatrix},$$

et son centre de gravité est

$$g_r = \left(f_{\cdot 1}, ..., f_{\cdot q}\right)^t.$$

On définit le nuage profils-lignes centré par

$$Y_r := X_r - \mathbf{1}_p g_r^t,$$

où

$$\mathbf{1}_p = (1, ..., 1)^t,$$

le vecteur unitaire de $p \times 1$. En d'autres termes

$$Y_r = \begin{pmatrix} \dfrac{f_{11}}{f_{1\cdot}} - f_{\cdot 1} & \cdots & \dfrac{f_{1q}}{f_{1\cdot}} - f_{\cdot q} \\ \vdots & \ddots & \vdots \\ \dfrac{f_{p1}}{f_{p\cdot}} - f_{\cdot 1} & \cdots & \dfrac{f_{pq}}{f_{p\cdot}} - f_{\cdot q} \end{pmatrix}.$$

Les éléments de Y_r sont

$$y_{i,j} = \frac{f_{ij}}{f_{i.}} - f_{.j}, \;\; i = 1, ..., p \text{ et } j = 1, ..., q.$$

On désigne par

$$\mathbf{y}_i := (y_{i,1}, ..., y_{i,q})^t, i = 1, ...p,$$

les vecteurs lignes de la matrice Y_r.

ACP des deux nuages de profils

Rappelons tout d'abord que nous avons muni l'espace $\mathbb{R}^q$ de la métrique de χ^2. Plus précisement

$$d_{\chi^2}^2\left(\mathbf{y}_i, \mathbf{y}_{i'}\right) = \sum_{j=1}^{q} \frac{1}{f_{\cdot j}} \left(y_{ij} - y_{i'j}\right)^2 = \left\| \mathbf{y}_i - \mathbf{y}_{i'} \right\|_{M_r}^2 = \mathbf{y}_i^t M_r \mathbf{y}_{i'},$$

où

$$M_r := D_c^{-1} = \begin{pmatrix} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{pmatrix}.$$

ACP des deux nuages de profils

Nous allons procéder l'ACP sur la matrice des données Y_r.

On note par E l'axe principal de l'ACP et u son vecteur directeur associé de norme 1 par rapport à métrique M_r, c'est à dire

$$\|u\|_{M_r}^2 = \langle u, u \rangle_{M_r} = u^t M_r u = 1.$$

On note

$$\text{la projection du point } \mathbf{y}_i \text{ sur } E^\perp := \mathbf{Proj}_{E^\perp, i}.$$

Nous définissons l'inertie du nuage Y_r par rapport à $E^\perp$ par

$$\text{Inertie}\left(Y_r / E^\perp\right) = \sum_{i=1}^{p} f_{i.} d_{\chi^2}^2 \left(\mathbf{y}_i, \mathbf{Proj}_{E^\perp, i}\right).$$

Rappelons que, d'après la relation de Chasles, on a

$$\frac{\langle \mathbf{y}_i, u \rangle_{M_r} u}{\|u\|_{M_r}^2} + \mathbf{Proj}_{E^\perp,i} = \mathbf{y}_i,$$

ce qui implique que

$$\mathbf{Proj}_{E^\perp,i} - \mathbf{y}_i = \frac{\langle \mathbf{y}_i, u \rangle_{M_r} u}{\|u\|_{M_r}^2}.$$

Alors

$$
\begin{aligned}
d_{\chi^2}^2\left(\mathbf{y}_i, \mathbf{Proj}_{E^\perp,i}\right) &= \left\|\mathbf{y}_i - \mathbf{Proj}_{E^\perp,i}\right\|_{M_r}^2 \\
&= \left\|\frac{\langle \mathbf{y}_i, u\rangle_{M_r} u}{\|u\|_{M_r}^2}\right\|_{M_r}^2 = \frac{\langle \mathbf{y}_i, u\rangle_{M_r}^2 \|u\|_{M_r}^2}{\|u\|_{M_r}^4} \\
&= \langle \mathbf{y}_i, u\rangle_{M_r}^2 \\
&= \left(\mathbf{y}_i^t M_r u\right)^2 = \left(\mathbf{y}_i^t M_r u\right)\left(\mathbf{y}_i^t M_r u\right)^t \\
&= \left(\mathbf{y}_i^t M_r u\right)\left(u^t M_r \mathbf{y}_i\right) \\
&= \left(u^t M_r \mathbf{y}_i\right)\left(\mathbf{y}_i^t M_r u\right) = u^t M_r \mathbf{y}_i \mathbf{y}_i^t M_r u.
\end{aligned}
$$

ACP des deux nuages de profils

Par conséquent

$$\text{Inertie}\left(Y_r / E^{\perp}\right) = u^t M_r \left[\sum_{i=1}^{p} f_{i\cdot}\mathbf{y}_i\mathbf{y}_i^t\right] M_r u.$$

On note que

$$\sum_{i=1}^{p} f_{i\cdot}\mathbf{y}_i\mathbf{y}_i^t = Y_r^t D_r Y_r =: \mathbf{V}_r,$$

comme étant la matrice de variance-covariance associée à la matrice Y_r affectée aux poids D_r. Ainsi

$$\text{Inertie}\left(Y_r / E^{\perp}\right) = u^t M_r \mathbf{V}_r M_r u.$$

Nous allons maintenant chercher le vecteur u maximisant l'inertie$\left(Y_r / E^{\perp}\right)$ sous la contrainte $u^t M_r u = 1$.

ACP des deux nuages de profils

Ce que revient, en utilisant le multiplicateur de Lagrange, à maximiser la fonction

$$u \longrightarrow \eta\left(u\right) := u^t M_r \mathbf{V}_r M_r u - \lambda\left(u^t M_r u - 1\right).$$

Il est clair que la dérivée de cette fonction est

$$\eta'\left(u\right) = 2M_r \mathbf{V}_r M_r u - 2\lambda M_r u.$$

En résolvant l'équation $\eta'(u) = 0$, on trouve

$$M_r \mathbf{V}_r M_r u = \lambda M_r u.$$

Rappelons que la matrice M_r est inversible, alors la dernière équation se réduit à

$$\mathbf{V}_r M_r u = \lambda u.$$

Comme nous l'avons fait au premier chapitre, nous allons appliquer l'ACP à la matrice $\mathbf{V}_r M_r$. En d'autres termes nous cherchons les valeurs propres et les vecteurs propres associés à $\mathbf{V}_r M_r$, définissants les axes principaux.

Rappel:

$$X_r \longrightarrow \text{ matrice des profils-lignes.}$$

$$Y_r := X_r - \mathbf{1}_p g_r^t \longrightarrow \text{ matrice des profils-lignes centrée.}$$

$$\mathbf{V}_r = Y_r^t D_r Y_r = X_r^t D_r X_r - g_r g_r^t \longrightarrow \text{ matrice de Var-Covar.}$$

$$E \longrightarrow \text{l'axe principal et son vecteur directeur } u \text{ avec } \|u\|_{M_r}^2 = 1.$$

$$\text{Inertie}\left(Y_r / E^{\perp}\right) = u^t M_r \mathbf{V}_r M_r u \longrightarrow \text{ inertie totale } (M_r = D_c^{-1}).$$

$$\max_{u, \|u\|_{M_r}^2 = 1} u^t M_r \mathbf{V}_r M_r u \longrightarrow \mathbf{V}_r M_r u = \lambda u \longrightarrow \text{vecteur propre de } \mathbf{V}_r M_r.$$

Remarque. La matrice $\mathbf{V}_r M_r$ est M_r-symétrique, c'est à dire $M_r \mathbf{V}_r M_r$ est symétrique.

En effet, comme $\mathbf{V}_r$ est $M_r = D_c^{-1}$ sont symétriques, il est alors évident que $M_r \mathbf{V}_r M_r$ l'est aussi. En effet,

$$\begin{aligned} (M_r \mathbf{V}_r M_r)^t &= ((M_r \mathbf{V}_r) M_r)^t = M_r^t (M_r \mathbf{V}_r)^t = M_r^t \mathbf{V}_r^t M_r^t \\ &= M_r \mathbf{V}_r M_r, \end{aligned}$$

car $\mathbf{V}_r$ et M_r sont tout les deux symétriques.

Théorème. Le centre de gravité du nuage des profils-lignes, g_r, est M_r-orthogonal au nuage des profils-lignes centré Y_r (de même, g_c est M_c-orthogonal au nuage des profils-colonnes centré Y_c).

Preuve. Tout d'abord observons que

$$M_r g_r = D_c^{-1} g_r = \begin{pmatrix} 1/f_{.1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{.q} \end{pmatrix} \begin{pmatrix} f_{.1} \\ \vdots \\ f_{.q} \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \mathbf{1}_q$$

et

$$\langle g_r, g_r \rangle_{M_r} = \|g_r\|_{M_r}^2 = g_r^t M_r g_r = g_r^t \mathbf{1}_q = f_{.1} + ... + f_{.q} = 1.$$

ACP des deux nuages de profils

Soit $\mathbf{y}_i := r_i - g_r$ la i-ème ligne de Y_r avec

$$r_i := \left(f_{i1}/f_{i\cdot}, ..., f_{iq}/f_{i\cdot}\right)^t$$

la i-ème ligne de X_r. Alors

$$\begin{aligned}
\langle \mathbf{y}_i, g_r \rangle_{M_r} &= \langle r_i - g_r, g_r \rangle_{M_r} = \langle r_i, g_r \rangle_{M_r} - \langle g_r, g_r \rangle_{M_r} \\
&= r_i^t M_r g_r - \langle g_r, g_r \rangle_{M_r} \\
&= r_i^t M_r g_r - \langle g_r, g_r \rangle_{M_r} \\
&= r_i^t M_r g_r - 1 = r_i^t \mathbf{1}_q - 1 \\
&= \left(f_{i1}/f_{i\cdot}, ..., f_{iq}/f_{i\cdot}\right) \mathbf{1}_q - 1 \\
&= f_{i1}/f_{i\cdot} + ... + f_{iq}/f_{i\cdot} - 1 = 1 - 1 = 0.
\end{aligned}$$

ACP des deux nuages de profils

Corollaire. Le centre de gravité g_r est un vecteur propre de $V_r M_r$ associé à la valeur propre $\lambda = 0$.

Preuve. On va démontrer que $V_r M_r g_r = 0 g_r$. Nous avons

$$V_r M_r g_r = V_r \mathbf{1}_q = \begin{pmatrix} \frac{f_{11}}{f_{1\cdot}} - f_{\cdot 1} + \ldots + \frac{f_{1q}}{f_{1\cdot}} - f_{\cdot q} \\ \vdots \\ \frac{f_{p1}}{f_{p\cdot}} - f_{\cdot 1} + \ldots + \frac{f_{pq}}{f_{p\cdot}} - f_{\cdot q} \end{pmatrix}.$$

Observons que

$$\begin{aligned} \frac{f_{11}}{f_{1\cdot}} - f_{\cdot 1} + \ldots + \frac{f_{1q}}{f_{1\cdot}} - f_{\cdot q} &= \left(\frac{f_{11}}{f_{1\cdot}} + \ldots + \frac{f_{1q}}{f_{1\cdot}} \right) - \left(f_{\cdot 1} + \ldots + f_{\cdot q} \right) \\ &= 1 - 1 = 0. \end{aligned}$$

De la même façon on montre aussi que les autres lignes sont nuls. Donc $V_r M_r g_r = 0 g_r$.

ACP des deux nuages de profils

Remarque. On rappel que le déterminant d'une matrice carré égal au produit des valeurs propres.

Comme $\lambda = 0$ est une valeur propre de $V_r M_r$ alors son déterminant est nul. Ce qui implique que cette matrice n'est pas inversible. Comme le rang égal au nombre maximum des valeurs propres non nuls, donc

$$rg\left(V_r M_r\right) \leq q - 1.$$

Avec le même raisonnement, on en conclus que

$$rg\left(V_c M_c\right) \leq p - 1.$$

Théorème. La matrice $V_r M_r$ a les mêmes valeurs propres que

$$X_r^t D_r X_r M_r = N^t D_r^{-1} N D_c^{-1} = X_r^t X_c^t := A_r,$$

sauf g_r qui a une valeur propre $\lambda = 1$.

Preuve. Soit $0 \neq v \neq g_r$ un vecteur propre de matrice $V_r M_r$, i.e. $V_r M_r v = \lambda v$. Ceci implique que

$$\left(X_r^t D_r X_r - g_r g_r^t\right) M_r v = \lambda v,$$

ainsi

$$X_r^t D_r X_r M_r v - g_r g_r^t M_r v = \lambda v.$$

Observons que

$$X_r^t D_r X_r M_r v - g_r \langle g_r, v \rangle_{M_r} = \lambda v.$$

D'un autre coté g_r est v sont deux vecteurs propres, M_r-orthogonaux, de $V_r M_r$, donc

$$X_r^t D_r X_r M_r v = \lambda v,$$

ce qui implique que v est un vecteur propre de $X_r^t D_r X_r M_r$ associé à la même paramètre λ. Par ailleurs nous avons

$$\begin{aligned} X_r^t D_r X_r M_r &= \left(D_r^{-1} N\right)^t D_r D_r^{-1} N D_c^{-1} \\ &= N^t D_r^{-1} N D_c^{-1} = X_r^t X_c^t = A_r. \end{aligned}$$

Corollaire. Il n'est pas nécessaire de centrer le nuage de points des profils-lignes avant de réaliser l'ACP, et on peut travailler directement avec la matrice

$$A_r = X_r^t X_c^t.$$

Dans le cas des profils-colonnes, on s'intéressera à la matrice

$$A_c = X_c^t X_r^t.$$

Lien entre les deux ACP

Rappel:

$$\text{ACP profils-lignes} \rightarrow V_r M_r.$$

$$\text{ACP profils-colonnes} \rightarrow V_c M_c.$$

$$\begin{array}{c} \text{val. pro } V_r M_r \longleftrightarrow (\text{sauf, } g_r, \lambda = 0) \\ \Updownarrow \\ \text{val. pro. de } X_r^t X_c^t \longleftrightarrow (\text{sauf, } g_r, \ \lambda = 1) \end{array}$$

De même

$$\begin{array}{c} \text{val. pro } V_c M_c \longleftrightarrow (\text{sauf, } g_c, \lambda = 0) \\ \Updownarrow \\ \text{val. pro. de } X_c^t X_r^t \longleftrightarrow (\text{sauf, } g_c, \ \lambda = 1) \end{array}$$

Lien entre les deux ACP

Exemple: Supposons que $V_r M_r \in \mathcal{M}(4 \times 4)$ et que ces valeurs propres sont

$$\lambda = \{0.5, 0.3, 0.2, 0\}.$$

Alors les valeurs propres de A_r sont

$$\lambda = \{1, 0.5, 0.3, 0.2\}.$$

Lien entre les deux ACP: formule de transition

Rappel: $M_r = D_c^{-1}$ et $M_c = D_r^{-1}$.

Théorème. Si u, est un vecteur propre, associé à la valeur propre $\lambda \neq 0$, M_r-norme 1, de A_r, alors

$$\widetilde{u} := \frac{1}{\sqrt{\lambda}} N D_c^{-1} u$$

est un vecteur propre, M_c-norme 1, pour A_c, pour la même valeur propre λ.

Lien entre les deux ACP: formule de transition

Inversement, si $\widetilde{u}$ est vecteur propre, associé à la valeur propre $\lambda \neq 0$, M_c-norme 1, de A_c, alors

$$u := \frac{1}{\sqrt{\lambda}} N^t D_r^{-1} \widetilde{u},$$

est un vecteur propre, M_r-norme 1, pour A_r, pour la même valeur propre λ.

Preuve. Notons u un vecteur propre $A_r = N^t D_r^{-1} N D_c^{-1}$, de norme 1 pour la métrique M_r, et associé à la valeur propre λ. C'est à dire $A_r u = \lambda u$. En d'autres termes

$$N^t D_r^{-1} N D_c^{-1} u = \lambda u.$$

En multipliant les deux membres de l'équation par$N D_c^{-1}$, on obient

$$\left(N D_c^{-1}\right) N^t D_r^{-1} N D_c^{-1} u = \lambda \left(N D_c^{-1}\right) u,$$

ainsi

$$\left(N D_c^{-1} N^t D_r^{-1}\right) \left(N D_c^{-1} u\right) = \lambda \left(N D_c^{-1} u\right).$$

Lien entre les deux ACP: formule de transition

Donc $kND_c^{-1}u$ est un vecteur propre de $ND_c^{-1}N^tD_r^{-1} = A_c$ associé à la même valeur propre λ.

On cherche ensuite la valeur de k telle que le vecteur propre soit, de norme 1 pour la métrique M_c.

Lien entre les deux ACP: formule de transition

Autrement dit on cherche k telle que

$$
\begin{aligned}
\left(kND_c^{-1}u\right)^t M_c \left(kND_c^{-1}u\right) &= 1 \Longleftrightarrow k^2 \left(u^t D_c^{-1} N^t\right) M_c N D_c^{-1} u = 1 \\
&\Longleftrightarrow k^2 u^t D_c^{-1} \left(N^t D_r^{-1} N D_c^{-1}\right) u = 1 \\
&\Longleftrightarrow k^2 u^t D_c^{-1} \left(X_r^t X_c^t\right) u = 1 \\
&\Longleftrightarrow k^2 u^t D_c^{-1} \left(A_r u\right) = 1 \\
&\Longleftrightarrow k^2 u^t D_c^{-1} \lambda u = 1
\end{aligned}
$$

Lien entre les deux ACP

$$
\begin{aligned}
&\iff \lambda k^2 \left(u^t M_r u\right) = 1 \\
&\iff \lambda k^2 \left\|u\right\|_{M_r}^2 = 1 \\
&\iff k = \frac{1}{\sqrt{\lambda}}.
\end{aligned}
$$

En résumé, si u est un vecteur propre $A_r = N^t D_r^{-1} N D_c^{-1}$, pour la valeur propre λ, alors

$$\widetilde{u} := \frac{1}{\sqrt{\lambda}} N D_c^{-1} u$$

est un vecteur propre pour $A_c = N D_c^{-1} N^t D_r^{-1}$, pour la même valeur propre λ.

Lien entre les deux ACP

Il existe un lien entre les vecteurs et valeurs propres des deux matrices

$$A_r := X_r^t X_c^t \text{ et } A_c := X_c^t X_r^t.$$

Nous avons

$$A_r \in M\left(q \times q\right) \text{ et } A_c \in M\left(p \times p\right).$$

Les deux matrices A_r et A_c ont les mêmes valeurs propres non nulles.

Corollaire. Les deux matrices A_r et A_c ont les mêmes valeurs propres non nulles. Par conséquent

$$rg\left(A_r\right) = rg\left(A_c\right),$$

ainsi

$$\tau := rg\left(V_r M_r\right) = rg\left(V_c M_c\right).$$

De plus

$$0 < \tau \leq \min\left(p-1, q-1\right).$$

Exemple: Supposons que $V_rM_r \in \mathcal{M}(4 \times 4)$ et $V_cM_c \in \mathcal{M}(5 \times 5)$, et que les valeurs propres de V_rM_r sont

$$\lambda = \{0.5, 0.3, 0.2., 0\} \to rg(V_rM_r) = 3$$

Alors les valeurs propres de $A_r := X_r^t X_c^t$ sont

$$\lambda = \{1, 0.5, 0.3, 0.2\} \to rg(A_r) = 4$$

Ainsi les valeurs propres de $A_c := X_c^t X_r^t$ sont

$$\lambda = \{1, 0.5, 0.3, 0.2, 0\} \to rg(A_c) = 4$$

Alors les valeurs propres de V_cM_c sont

$$\lambda = \{0.5, 0.3, 0.2, 0, 0\} \to rg(V_cM_c) = 3.$$

Rappel: u_i (sont M_r-orthonormés) et $\widetilde{u}_i$ (sont M_c-orthonormés) sont les vecteurs propres de A_r et A_c associés aux meme valeurs propres non-nulles,

Définition. On appel **facteurs principaux** des profils-lignes (resp. profils-colonnes), les vecteurs

$$w_i := M_r u_i,$$

resp.

$$\widetilde{w}_i := M_c \widetilde{u}_i.$$

Proposition. Les facteurs principaux des profils-lignes (resp. profils-colonnes) sont $M_r^{-1}-$orthonormés (resp. $M_c^{-1}-$orthonormé.

Preuve. Soient $u_i, i = 1, ..., p$ les axe principaux des profils-lignes. On sait que les u_i sont M_r-orthonormés, donc

$$w_i^t M_r^{-1} w_j = u_i^t M_r M_r^{-1} M_r u_j = u_i^t M_r u_i = 1 \text{ si } i = j \text{ et } 0 \text{ si } i \neq j.$$

Proposition. Les facteurs principaux des profils-lignes (rep. profils colonnes) sont les vecteurs propres $M_r^{-1}-$orthonormés (resp. $M_c^{-1}-$orthonormés) de la matrice de données $M_r V_r$ (resp. $M_c V_c$).

Preuve.. Soit u un vecteur propre de $V_r M_r$ associé à la valeur propre λ, alors $V_r M_r u = \lambda u$. En multipliant les deux membres de cette équation par M_r, on obtient

$$M_r \left(V_r M_r u \right) = \lambda M_r u.$$

Donc

$$M_r V_r \left(M_r u \right) = \lambda \left(M_r u \right),$$

où $M_r u =: w$ est le facteur principal associé à u. De plus

$$w^t M_r^{-1} w = u^t M_r M_r^{-1} M_r u = u^t M_r u = 1,$$

car u est $M_r -$normé.

Facteurs principaux et composantes principales

Définition. Les composantes principales des profils-lignes sont les M_r-coordonnées des vecteurs colonnes de la matrice de données $Y_r := X_r - \mathbf{1}_p g_r^t$. C'est à dire

$$c_k := Y_r w_k, \ k = 1, ..., \tau,$$

où $w_k = M_r u_k$ sont les facteurs principaux.

Les composantes principales des profils-colonnes sont les M_c-coordonnées des vecteurs colonnes de la matrice de données $Y_c := X_c - \mathbf{1}_q g_c^t$. C'est à dire

$$\widetilde{c}_k := Y_c \widetilde{w}_k, \ k = 1, ..., \tau,$$

où $\widetilde{w}_k = M_c \widetilde{u}_k$ sont les facteurs principaux.

Proposition

Proposition. *Les composantes principales des profils-lignes sont les M_r-coordonnées des vecteurs colonnes de la matrice de données X_r, C'est à dire*

$$c_k := X_r w_k, \;\; k = 1, ..., \tau.$$

Les composantes principales des profils-colonnes sont les M_c-coordonnées des vecteurs colonnes de la matrice de données X_c. C'est à dire

$$\widetilde{c}_k := X_c \widetilde{w}_k, \;\; k = 1, ..., \tau.$$

Nous avons $Y_r := X_r - \mathbf{1}_p g_r^t$ et

$$c_k = Y_r w_k = X_r w_k - \mathbf{1}_p g_r^t w_k = X_r w_k - \mathbf{1}_p g_r^t M_r u_k$$

Nous avons enoncé que les vecteurs propres de $V_r M_r$ sont M_r-orthogonaux, donc $g_r^t M_r u_k = 0$.

Preuve. En effet, nous avons

$$c_k = Y_r w_k = \left(X_r - \mathbf{1}_p g_r^t\right) M_r u_k = X_r M_r u_k - \mathbf{1}_p \left(g_r^t M_r u_k\right).$$

Rappelons que g_r et u_k sont deux vecteurs propres $V_r M_r$ associes aux valeurs propores $\lambda = 0$ et $\lambda_k \neq 0$. Nous avons aussi enoncé aussi que les vecteurs propres de $V_r M_r$ sont M_r-orthogonaux. Donc $g_r^t M_r u_k = 0$, ainsi

$$c_k = X_r w_k.$$

Proposition. Nous avons

$$\frac{1}{p}\sum_{j=1}^{p} c_k\left(j\right) = \frac{1}{q}\sum_{j=1}^{q} \widetilde{c}_k\left(j\right) = 0,\ \ k = 1, ..., \tau,$$

$$\frac{1}{p}\sum_{j=1}^{p} c_k^2\left(j\right) = \frac{1}{q}\sum_{j=1}^{q} \widetilde{c}_k^2\left(j\right) = \lambda_k,\ \ k = 1, ..., \tau,$$

et

$$\begin{cases} \dfrac{1}{p}\sum_{i=1}^{p} c_k\left(i\right) c_\ell\left(i\right) = 0, \text{ pour } k \neq \ell \\ \dfrac{1}{q}\sum_{j=1}^{q} \widetilde{c}_k\left(j\right) \widetilde{c}_\ell\left(j\right) = 0, \text{ pour } k \neq \ell. \end{cases}$$

Proposition. Les facteurs principaux de l'ACP des profils-colonnes, associés aux valeurs propres non nulles, sont, à une constante près, les composantes principales de l'ACP des profils-lignes, et vice-versa. Plus précisement

$$c = \frac{1}{\sqrt{\lambda}} A_c^t \widetilde{w} \text{ et } \widetilde{c} = \frac{1}{\sqrt{\lambda}} A_r^t w.$$

Ce qui equivalent à

$$c = \sqrt{\lambda} \widetilde{w} \text{ et } \widetilde{c} = \sqrt{\lambda} w.$$

Preuve. En effet, soit u un axe principal, des profils-lignes, associé à la valeur propre $\lambda \neq 0$ et $\widetilde{u}$ l'axe principal correspondant des profils-colonnes. Observons maintenant que, la composante principale associée à u est

$$\begin{aligned} c &= X_r w = X_r M_r u = \left(D_r^{-1} N\right) D_c^{-1} u = \frac{1}{\sqrt{\lambda}} D_r^{-1} N D_c^{-1} N^t D_r^{-1} \widetilde{u} \\ &= \frac{1}{\sqrt{\lambda}} X_r X_c \widetilde{w} = \frac{1}{\sqrt{\lambda}} A_c^t \widetilde{w}. \end{aligned}$$

Rappelons que $\widetilde{u} := N D_c^{-1} u / \sqrt{\lambda}$ est un axe principale des profils-colonnes. Donc

$$c = \sqrt{\lambda} D_r^{-1} \widetilde{u} = \sqrt{\lambda} M_c \widetilde{u} = \sqrt{\lambda} \widetilde{w}.$$

Inversement

$$\widetilde{c} = X_c M_c \widetilde{u} = D_c^{-1} \left(N^t D_r^{-1} \widetilde{u}\right) = D_c^{-1} \left(\sqrt{\lambda} u\right) = \sqrt{\lambda} w.$$

Facteurs principaux et composantes principales:

Le résultat précédent conduit aux relations fondamentales de l'AFC reliant les composantes principales entre elles, dites les relations *quasi-barycentriques*:

Proposition. Soit $\lambda_1 > \lambda_2 > ... > \lambda_\tau \neq 0$. Alors, pour tout $k \leq \tau$, on a

$$c_k := \frac{1}{\sqrt{\lambda_k}} X_r \widetilde{c}_k \text{ et } \widetilde{c}_k := \frac{1}{\sqrt{\lambda_k}} X_c c_k.$$

Par conséquent

$$c_k := \frac{1}{\sqrt{\lambda_k}} X_r X_c c_k$$

Preuve. On a

$$\begin{aligned} c_k &= X_r w_k = X_r M_r u_k = X_r D_c^{-1} \left(\frac{1}{\sqrt{\lambda}} N^t D_r^{-1} \widetilde{u} \right) \\ &= \frac{1}{\sqrt{\lambda}} X_r \left(D_c^{-1} N^t \right) \left(D_r^{-1} \widetilde{u} \right) = \frac{1}{\sqrt{\lambda}} X_r \left(X_c \right) \left(M_c \widetilde{u} \right) \\ &= \frac{1}{\sqrt{\lambda}} X_r \left(X_c \widetilde{w} \right) = \frac{1}{\sqrt{\lambda}} X_r \widetilde{c}. \end{aligned}$$

Proposition. On a

$$0 \leq \lambda_k \leq 1, \; k = 1, ..., \tau.$$

Preuve. Soient $\mathbf{A} = (a_{ij}) \in \mathcal{M}(p \times q)$, $\mathbf{x} = (x_1, ..., x_q)^t \in \mathbb{R}^p$ et $\mathbf{y} = (y_1, ..., y_p)^t \in \mathbb{R}^q$. On munit $\mathbb{R}^q$ et $\mathbb{R}^p$ des normes L_1 définis par

$$\|\mathbf{x}\| = \sum_{i=1}^{q} |x_i|, \;\; \|\mathbf{y}\| = \sum_{i=1}^{p} |y_i|.$$

On munit aussi l'espace des matricies $\mathcal{M}(p \times q)$ de la norme

$$\|\mathbf{A}\| = \sup_{\|\mathbf{x}\| \neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{Ax}\| = \sup_{\|\mathbf{x}\| = 1} \|\mathbf{Ax}\|.$$

On montre que

$$\|\mathbf{A}\| = \max_{j=1,\dots,q} \sum_{i=1}^{p} |a_{ij}| \, ,$$

(voir mon site moodel). Il est claire que

$$\frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\| \Longrightarrow \|\mathbf{Ax}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\| \, ,$$

ainsi

$$\|\mathbf{Ax}\| \leq \max_{j=1,\dots,q} \sum_{i=1}^{p} |a_{ij}| \, \|\mathbf{x}\| \, .$$

Application: $\mathbf{A} = X_r \in \mathcal{M}\left(p \times q\right)$. Comme c_k est un vecteur de $\mathbb{R}^p$ alors $\mathbf{x} = X_c c_k$ étant un vecteur de $\mathbb{R}^p$. Donc on a

$$\begin{aligned} \lambda \left\| c_k \right\| &= \left\| \left(X_r\right)\left(X_c c_k\right) \right\| = \left\| \mathbf{Ax} \right\| \leq \left\| \mathbf{A} \right\| \left\| \mathbf{x} \right\| \\ &\leq \max_{j=1,\dots,q} \sum_{i=1}^{p} \left| a_{ij} \right| \left\| \mathbf{x} \right\|. \end{aligned}$$

C'est à dire

$$\lambda \left\| c_k \right\| \leq \max_{j=1,\dots,q} \sum_{i=1}^{p} \left| a_{ij} \right| \left\| X_c c_k \right\|.$$

En appliquant ces dernières inégalites $\mathbf{A} = X_c = (b_{ij}) \in \mathcal{M}(p \times q)$, on ecrit

$$\|X_c c_k\| \leq \|X_c\| \, \|c_k\| = \left\{ \max_{j=1,\ldots,p} \sum_{i=1}^{q} |b_{ij}| \right\} \|c_k\| .$$

Ce qui implique

$$\lambda \|c_k\| = \|X_r X_c c_k\| \leq \left\{ \max_{j=1,\ldots,q} \sum_{i=1}^{p} |a_{ij}| \right\} \left\{ \max_{j=1,\ldots,p} \sum_{i=1}^{q} |b_{ij}| \right\} \|c_k\| .$$

En simplifiant par $\|c_k\|$ (qui est évidement non nulle), on obtien

$$0 < \lambda \leq \left\{ \max_{j=1,\ldots,q} \sum_{i=1}^{p} |a_{ij}| \right\} \left\{ \max_{i=1,\ldots,p} \sum_{j=1}^{q} |b_{ij}| \right\} .$$

$$0 < \lambda \leq \left\{ \max_{i=1,\ldots,p} \sum_{j=1}^{q} \frac{f_{ij}}{f_{i.}} \right\} \left\{ \max_{j=1,\ldots,q} \sum_{i=1}^{p} \frac{f_{ji}}{f_{.j.}} \right\}.$$

Notons que

$$\sum_{j=1}^{q} \frac{f_{ij}}{f_{i.}} = \frac{1}{f_{i.}} \sum_{j=1}^{q} f_{ij} = \frac{f_{i.}}{f_{i.}} = 1$$

et

$$\sum_{i=1}^{p} \frac{f_{ij}}{f_{.j.}} = \frac{1}{f_{.j}} \sum_{i=1}^{p} f_{ij} = \frac{f_{.j}}{f_{.j}} = 1,$$

et parconséquent $0 < \lambda \leq 1$.

Formule de reconstitution des données

Théorème. Nous avons

$$\frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j} = f_{\cdot j} \sum_{k=1}^{\tau} \frac{1}{\sqrt{\lambda_k}} c_k\left(i\right) \widetilde{c}_k\left(j\right).$$

Les composantes principales et les valeurs propres expliquent en quoi les fréquences observées s'écartent des fréquences théoriques.

Formule de reconstitution des données

- Cette formule, appelée formule de reconstitution des données, permet de recalculer les valeurs du tableau initial en fonction des marges et des facteurs.
- Lorsque l'on dépouille les résultats d'une AFC, on limite généralement l'interprétation aux premiers axes principaux. Cela revient à considérer non pas le tableau des données mais son approximation obtenue à l'aide des premiers termes de la somme ci-dessus:

$$f_{ij} - f_{i.}f_{j.} \simeq f_{i.}f_{.j} \sum_{k=1}^{2} \frac{1}{\sqrt{\lambda_k}} c_k(i)\, \widetilde{c}_k(j).$$

En d'autres termes

$$f_{ij} \simeq f_{i.}f_{.j} \left(1 + \sum_{k=1}^{2} \frac{1}{\sqrt{\lambda_k}} c_k(i)\, \widetilde{c}_k(j) \right).$$

Preuve. Nous avons

$$Y_r := X_r - \mathbf{1}_p g_r^t,$$

dont ces éléments sont

$$y_{i,j} = \frac{f_{ij}}{f_{i\cdot}} - f_{\cdot j}, \quad i = 1, ..., p \text{ et } j = 1, ..., q.$$

Rappelons que $\{u_1, ..., u_p\}$ est une base M_r-orthonormée de $\mathbb{R}^p$ et les c_k, $k = 1, ..., q$ sont coordonnées de la i-ième ligne de Y_r dans la base $\{u_1, ..., u_p\}$. En d'autres termes

$$Y_r = \sum_{k=1}^{p} c_k u_k = \sum_{k=1}^{\tau} c_k u_k.$$

Formule de reconstitution des données

Rappelons aussi que

$$u_k = \frac{1}{\sqrt{\lambda_k}} N^t D_r^{-1} \widetilde{u}_k = \frac{1}{\sqrt{\lambda_k}} X_r^t \widetilde{u}_k.$$

Alors

$$\begin{aligned} Y_r &= \sum_{k=1}^{\tau} \frac{1}{\sqrt{\lambda_k}} c_k N^t D_r^{-1} \widetilde{u}_k = \sum_{k=1}^{\tau} \frac{1}{\sqrt{\lambda_k}} c_k D_c \left(D_c^{-1} N^t \right) \left(D_r^{-1} \widetilde{u}_k \right) \\ &= \sum_{k=1}^{\tau} \frac{1}{\sqrt{\lambda_k}} c_k D_c X_c \widetilde{w}_k = \sum_{k=1}^{\tau} \frac{1}{\sqrt{\lambda_k}} c_k D_c \widetilde{c}_k. \end{aligned}$$

Formule de reconstitution des données

En écrivant cette relation coordonnées par coordonnées, on obtient

$$\frac{f_{ij}}{f_{i.}} - f_{.j} = \frac{1}{\sqrt{\lambda_k}} \sum_{k=1}^{\tau} c_k\left(i\right) f_{.j} \widetilde{c}_k\left(j\right).$$

En multipliant les deux membres de cette équation par $f_{.i}$ on obtient

$$f_{ij} - f_{.j} f_{i.} = f_{i.} f_{j.} \sum_{k=1}^{\tau} \frac{1}{\sqrt{\lambda_k}} c_k\left(i\right) \widetilde{c}_k\left(j\right).$$

Aide à la représentation

- Contribution relative d'inertie
- Qualité de la représentation
- Cercle de corrélations

Contributions relatives

- Contribution relative du profil-ligne i de de la matrice Y_r au $k-$ième axe u_k :

$$Ctr\left(i,k\right):=\frac{f_{i.}c_k^2\left(i\right)}{\sum_{i=1}^{p}f_{i.}c_k^2\left(i\right)}=\frac{f_{i.}c_k^2\left(i\right)}{\lambda_k},\ \ i=1,...,p;\ \ k=1,...,\tau.$$

- Contribution relative du profil-colonne j de de la matrice Y_c au $k-$ième axe $\widetilde{u}_k$:

$$\widetilde{Ctr}\left(j,k\right)=\frac{f_{.j}\widetilde{c}_k^2\left(j\right)}{\sum_{j=1}^{q}f_{.j}\widetilde{c}_k^2\left(j\right)}=\frac{f_{.j}\widetilde{c}_k^2\left(j\right)}{\lambda_k},\ \ j=1,...,q;\ \ k=1,...,\tau.$$

Qualité de représentation sur un axe

- Qualité de la représentation du profil-lignes i de la matrice Y_r au $k-$ième axe u_k :

$$\begin{aligned}\cos^2(\widehat{i, u_k}) &= \left(\frac{\langle i, u_k\rangle_{M_r}}{\|i\|_{M_r}\|u_k\|_{M_r}}\right)^2 = \frac{(i^t M_r u_k)^2}{\|i\|^2_{M_r}}, \quad (\|u_k\|_{M_r} = 1)\\ &= \frac{c_k^2(i)}{\sum_{k=1}^{\tau} c_k^2(i)}, \quad i = 1, ..., p.\end{aligned}$$

- Qualité de la représentation du profil-colonnes j de la matrice Y_c au $k-$ième axe $\widetilde{u}$:

$$\begin{aligned}\cos^2(\widehat{j, \widetilde{u}_k}) &= \left(\frac{\langle j, \widetilde{u}_k\rangle_{M_c}}{\|j\|_{M_c}\|\widetilde{u}_k\|_{M_c}}\right)^2 = \frac{\widetilde{c}_k^2(j)}{\|j\|^2_{M_c}}, \quad (\|\widetilde{u}_k\|_{M_c} = 1)\\ &= \frac{\widetilde{c}_k^2(j)}{\sum_{k=1}^{\tau} \widetilde{c}_k^2(j)}, \quad j = 1, ..., q.\end{aligned}$$

Proposition. Pour chaque $j \in \{1, ..., q\}$ et $k \in \{1, ..., \tau\}$, on a

$$\mathbf{Cor}(c_k, Y_r^{(j)}) = \frac{\sqrt{\lambda_k}}{s_j} u_k (j) ,$$

où $s_j = \sqrt{\mathbf{Var}\left(Y_r^{(j)}\right)}$ et $u_k (j)$ désigne la $j - \grave{e}me$ composante du vecteur propre u_k.

Preuve. On a

$$\mathbf{Cor}\left(c_k, Y_r^{(j)}\right) = \frac{\mathbf{Cov}\left(c_k, Y_r^{(j)}\right)}{\sqrt{\mathbf{Var}\left(c_k\right)\mathbf{Var}\left(Y_r^{(j)}\right)}} = \frac{1}{\sqrt{\lambda_k}} \frac{c_k^t D_r Y_r^{(j)}}{s_j}.$$

Remarquons que $Y_r^{(j)} = Y_r \delta$, où $\delta := (0, ..., 1, ..., 0)^t$ est un vecteur à q composantes. Donc

$$\begin{aligned}
\mathbf{Cor}(j, k) &= \frac{1}{\sqrt{\lambda_k}} \frac{\left(Y_r w_k\right)^t D_r Y_r \delta}{s_j} = \frac{1}{\sqrt{\lambda_k}} \frac{\left(Y_r M_r u_k\right)^t D_r Y_r \delta}{s_j} \\
&= \frac{1}{\sqrt{\lambda_k}} \frac{u_k^t M_r Y_r^t D_r Y_r \delta}{s_j} = \frac{1}{\sqrt{\lambda_k}} \frac{u_k^t M_r V_r \delta}{s_j} \\
&= \frac{1}{\sqrt{\lambda_k}} \frac{\left(V_r M_r u_k\right)^t \delta}{s_j} = \frac{1}{\sqrt{\lambda_k}} \frac{\lambda_k u_k^t \delta}{s_j} \\
&= \frac{\sqrt{\lambda_k}}{s_j} u_k^t \delta = \frac{\sqrt{\lambda_k}}{s_j} u_k\left(j\right).
\end{aligned}$$