ECOLE MAROCAINE DES SCIENCES DE L'INGENIEUR
Membre de HONORIS UNITED UNIVERSITIES

Prof. Karima EL Qate

Année universitaire 2022-2023

Correction de TD N° 2
Analyses de données : Analyse en Composantes Principales

Solution :

1. Les moyennes des quatre variables :

$$\text{Moyenne de } D_1 : \quad \bar{x_1} = \frac{-11-12-15-14-14.5-13}{6} = -13.25$$

$$\text{Moyenne de } D_2 : \quad \bar{x_2} = \frac{-60-62-80-75-82-72}{6} = -71.833$$

$$\text{Moyenne de } D_3 : \quad \bar{x_3} = \frac{110+93+113+94+100+102}{6} = 102$$

$$\text{Moyenne de } D_4 : \quad \bar{x_4} = \frac{40+25+39+25+30+32}{6} = 31.833$$

	D_1	D_2	D_3	D_4
S_1	-11	-60	110	40
S_2	-12	-62	93	25
S_3	-15	-80	113	39
S_4	-14	-75	94	25
S_5	-14,5	-82	100	30
S_6	-13	-72	102	32
Moyenne : $\bar{x_j}$	-13,25	-71.83	102	31.833
Écart-type : σ_{D_j}	1,407	8,36	7,461	5,986

Le centre de gravité des individus : est le vecteur contenant les moyennes des variables D_1, D_2, D_3, D_4;

$$g = (-13{,}25, \ -71.833, \ 102, \ 31.833)^T$$

L'inertie totale des individus :

$$\begin{aligned} I_{tot}(g) &= \sum_{j=1}^{4} var(D_j) = D_1 + D_2 + D_3 + D_4 \\ &= (1{,}407)^2 + (8{,}36)^2 + (7{,}461)^2 + (5{,}986)^2 = 163{,}50 \end{aligned}$$

2. On doit calculer les distances entre le sujet S_1 et tous les autres sujets :

$$\begin{aligned} d(S_1,S_2) &= \sqrt{(-11+12)^2+(-60+62)^2+(110-93)^2+(40-25)^2} = 22.78 \\ d(S_1,S_3) &= \sqrt{(-11+15)^2+(-60+80)^2+(110-113)^2+(40-39)^2} = 20{,}64 \\ d(S_1,S_4) &= \sqrt{(-11+14)^2+(-60+75)^2+(110-94)^2+(40-25)^2} = 26.74 \\ d(S_1,S_5) &= \sqrt{(-11+14.5)^2+(-60+82)^2+(110-100)^2+(40-30)^2} = 26.39 \\ d(S_1,S_6) &= \sqrt{(-11+13)^2+(-60+72)^2+(110-102)^2+(40-32)^2} = 16.61 \end{aligned}$$

La plus petite des distances entre S_1 et les autres sujets est $d(S_1,S_6)$, donc S_6 est le sujet le plus proche de S_1.

3. **La matrice Z des données centrées réduites.**
On prend chaque variable et on soustrait de ses valeurs sa moyenne et on divise par son écart-type,

$$X_j^* = \frac{X_j - \bar{x_j}}{\sigma(X_j)}$$

Donc la matrice Z est :

$$Z = \begin{pmatrix} 1{,}6 & 1.419 & 1.072 & 1.364 \\ 0.888 & 1.179 & -1.206 & -1.141 \\ -1.244 & -0.979 & 1.474 & 1.197 \\ -0.533 & -0.380 & -1.072 & -1.141 \\ -0.888 & -1.219 & -0.268 & -0.306 \\ 0.178 & -0.020 & 0 & 0.028 \end{pmatrix}$$

Indication : (centrage et réduction de D_1)

Sujet	D_1	$D_1^* = \dfrac{D_1 - \bar{x_1}}{\sigma(D_1)}$
S_1	-11	$\dfrac{-11 + 13{,}25}{1{,}407} = 1{,}599$
S_2	-12	0.888
S_3	-15	-1.244
S_4	-14	-0.533
S_5	-14.5	-0.888
S_6	-13	0.178
Moyenne	-13,25	0
Écart-type	1,407	1

4. La matrice R de corrélation :

$$R = \frac{1}{6} Z^t.Z = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{pmatrix}$$

On sait que $r_{12} = r_{21}$, $r_{13} = r_{31}$, $r_{14} = r_{41}$, $r_{23} = r_{32}$, $r_{24} = r_{42}$, et $r_{34} = r_{43}$

d'où

$$R = \begin{pmatrix} 1 & 0.821 & -0.064 & 0.590 \\ 0.821 & 1 & -0.102 & 0.056 \\ -0.064 & -0.102 & 1 & 0.397 \\ 0.590 & 0.056 & 0.397 & 1 \end{pmatrix}$$

5. Interprétation des résultats obtenus.

	X_1	X_2	X_3	X_4
X_1	1	0.821	-0.064	0.590
X_2	0.821	1	-0.102	0.056
X_3	-0.064	-0.102	1	0.397
X_4	0.590	0.056	0.397	1

La matrice de corrélation montre que la variable 3 est faiblement corrélée avec la variable 1,2 et la variable 4.

6. Sachant que les valeurs propres et les vecteurs propres de R sont

$$\lambda_1 = 2.0011, \quad \lambda_2 = 1.967, \quad \lambda_3 = 0.032, \quad \lambda_4 = 0.0003$$

et

$$v_1 = \begin{pmatrix} 0.441933 \\ 0{,}467633 \\ -0{,}57415 \\ -0{,}50633 \end{pmatrix}, v_2 = \begin{pmatrix} -0{,}55021 \\ -0{,}52745 \\ -0{,}41487 \\ -0{,}49694 \end{pmatrix},$$

$$v_3 = \begin{pmatrix} 0{,}655642 \\ -0{,}68925 \\ -0{,}23389 \\ 0{,}200895 \end{pmatrix}, v_4 = \begin{pmatrix} -0{,}2685 \\ 0{,}167476 \\ -0{,}66598 \\ 0{,}675519 \end{pmatrix}$$

7. **Calcul des Composantes principales : La première composante principale** Y_1 (définie pour tous les individus) est égale au produit entre la matrice des données centrées réduites (la matrice $\mathbf{Z}$) et le premier vecteur propre (les valeurs de v_1). Ainsi, nous avons :

$$Y_1 = \mathbf{Z}v_1 = \begin{pmatrix} 1{,}6 & 1.419 & 1.072 & 1.364 \\ 0.888 & 1.179 & -1.206 & -1.141 \\ -1.244 & -0.979 & 1.474 & 1.197 \\ -0.533 & -0.380 & -1.072 & -1.141 \\ -0.888 & -1.219 & -0.268 & -0.306 \\ 0.178 & -0.020 & 0 & 0.028 \end{pmatrix} \begin{pmatrix} 0{,}441933 \\ 0{,}467633 \\ -0{,}57415 \\ -0{,}50633 \end{pmatrix} = \begin{pmatrix} 0.0641 \\ 2{,}2139 \\ -1{,}2478 \\ 0{,}7800 \\ -0.6537 \\ 0.0551 \end{pmatrix}$$

La deuxième composante principale Y_2 :

$$Y_2 = \mathbf{Z}v_2 = \begin{pmatrix} 1{,}6 & 1.419 & 1.072 & 1.364 \\ 0.888 & 1.179 & -1.206 & -1.141 \\ -1.244 & -0.979 & 1.474 & 1.197 \\ -0.533 & -0.380 & -1.072 & -1.141 \\ -0.888 & -1.219 & -0.268 & -0.306 \\ 0.178 & -0.020 & 0 & 0.028 \end{pmatrix} \begin{pmatrix} -0{,}55021 \\ -0{,}52745 \\ -0{,}41487 \\ -0{,}49694 \end{pmatrix} = \begin{pmatrix} -2{,}7508 \\ -0{,}0431 \\ 1{,}1842 \\ 1.5054 \\ 1.3948 \\ -0.1013 \end{pmatrix}$$

La troisième composante principale :

$$Y_3 = \mathbf{Z}v_3 = \begin{pmatrix} 1{,}6 & 1.419 & 1.072 & 1.364 \\ 0.888 & 1.179 & -1.206 & -1.141 \\ -1.244 & -0.979 & 1.474 & 1.197 \\ -0.533 & -0.380 & -1.072 & -1.141 \\ -0.888 & -1.219 & -0.268 & -0.306 \\ 0.178 & -0.020 & 0 & 0.028 \end{pmatrix} \begin{pmatrix} 0{,}655642 \\ -0{,}68925 \\ -0{,}23389 \\ 0{,}200895 \end{pmatrix} = \begin{pmatrix} 0{,}0936 \\ -0{,}1776 \\ -0{,}7261 \\ -0.0660 \\ 0{,}2592 \\ 0.1361 \end{pmatrix}$$

La quatrième composante principale :

$$Y_4 = \mathbf{Z}v_4 = \begin{pmatrix} 1{,}6 & 1.419 & 1.072 & 1.364 \\ 0.888 & 1.179 & -1.206 & -1.141 \\ -1.244 & -0.979 & 1.474 & 1.197 \\ -0.533 & -0.380 & -1.072 & -1.141 \\ -0.888 & -1.219 & -0.268 & -0.306 \\ 0.178 & -0.020 & 0 & 0.028 \end{pmatrix} \begin{pmatrix} -0{,}2685 \\ 0{,}167476 \\ -0{,}66598 \\ 0{,}675519 \end{pmatrix} = \begin{pmatrix} 0{,}0158 \\ -0{,}0086 \\ -1{,}6202 \\ 0.0226 \\ 0{,}0060 \\ -0.0322 \end{pmatrix}$$

8. **Choix des composantes principales :**

On doit déterminer le nombre de composantes principales à retenir : la règle recommande d'extraire des facteurs de façon à expliquer au moins 80% de la variance totale :

Ainsi, dans notre exemple où nous avons $\lambda_1 = 2.001$ et $I_{tot}(g = 0) = p = 4$, le taux d'inertie expliquée par la première composante principale (facteur) est égal à

$$\frac{\lambda_1}{I_{tot}(g=0)} \times 100 = \frac{2.001}{4} \times 100 = 50.025$$

En répétant ce processus pour les composantes principales (facteurs) suivantes, nous obtenons le tableau ci-dessous

Valeurs propores	**Inertie expliquée (%)**	**Inertie cumulée (%)**
$\lambda_1 = 2.0011$	$\frac{\lambda_1}{p} \times 100 = 50.025\%$	50.025%
$\lambda_2 = 1.967$	$\frac{\lambda_2}{p} \times 100$=49.175%	99.2%
$\lambda_3 = 0.032$	$\frac{\lambda_p}{p} \times 100$=0.8%	99.999...%
$\lambda_4 = 0.0003$	$\frac{\lambda_p}{p} \times 100$=0.0075%	100%

9. Le nombre de composantes principales à retenir est égale : 2
10. On utilise la formule suivante pour le calculé les coefficient de corrélation linéaire entre Y_j et X_l, :

$$r(Y_j,X_l) = \text{élement } l \text{ du vecteur propre } v_j * \sqrt{\lambda_j}$$

Alors, les corrélations entre les variables initiales et les composantes principales retenues sont réumées dans le tableau suivant :

	Y_1	Y_2
X_1	0.625	-0.771
X_2	0.661	-0.739
X_3	-0.812	-0.581
X_4	-0.716	-0.696