Université Mohamed Khider, Biskra 14 juin 2021
Faculté des Sc. Exactes et Sc. de la Nature et la Vie
Master 1: **Modèle linéaire**

Examen (1h)

Exercice 1 ***(ACP, 10 pts)***
Considérons le tableau de données suivant :

	Poids	***Taille***	***Âge***
Farid	45	150	13
Adel	50	165	15
Tarek	60	160	14
Said	52	152	14

Tab.1

où les lignes représentent les individus (noms des collégiens) et les colonnes les variables (poids, taille, âge). On s'intéresse à une Analyse en Composantes Principales (ACP) basée sur la matrice de corrélation associée au tableau ***Tab.1****, donnée par:*

$$R := \begin{pmatrix} 1 & 0.433 & 0.327 \\ 0.433 & 1 & 0.875 \\ 0.327 & 0.875 & 1 \end{pmatrix}.$$

Les composantes principales associées sont:

C_1	C_2	C_3
2.147	-0.330	-0.283
-1.632	-1.137	-0.059
-1.005	1.247	-0.230
0.489	0.219	0.573

1) Que vaut la moyenne de chaque composante principale C_k, $k = 1, 2, 3$? $(0.5pt)$
2) Calculer les variances de C_1 et C_2 puis "sans calcul" déduire celle de C_3. $(1.5pt)$
3) Déterminer les valeurs propres, $\lambda_1 \geq \lambda_2 \geq \lambda_3$ associées à la matrice R? $(0.5pt)$
4) En utilisant la définition de la corrélation, calculer la corrélation de C_1 avec C_2; puis "sans calcul" donner la valeur de la corrélation de C_1 avec C_3. $(1pt)$
Soient

$$u_1 := (-0.430, 0.896, 0.103)^t ;\ u_2 := (-0.649, -0.227, -0.725)^t ;\ u_3 := (-0.626, -0.379, 0.680)^t ,$$

les vecteurs propres normés, de la matrice R, associés aux valeurs propres $\lambda_1, \lambda_2, \lambda_3$.
5) Déterminer les axes principaux associés à cette ACP?$(1.5pt)$
6) Sans centrer-réduire la matrice des données, déduire les corrélations entres la variable-Poids (centré-réduit) et les trois composantes principales.$(1.5pt)$
7) Quels sont les pourcentages d'inerties expliquées par les deux premiers axes principaux?$(1pt)$
8) Quel est le pourcentage d'inertie expliquée par le premier plan principal?(0.5). Que conclut-on?$(0.5pt)$
9) Représenter "soigneusement" sur le premier plan principal les quatre collégiens? $(1.5pt)$

Exercice 2 ***(AFC, 10 pts)***
1) Soient $u_k \neq g_r$ les vecteurs directeurs des axes principaux. Calculer le produi matriciel $g_r^t M_r u_k$ et déduire que

$$\sum_{j=1}^{q} u_k(j) = 0? \ (2pts)$$

2) Donner la formule de l'inertie d'une modalité des profils-lignes X_r par rapport à g_r? $(2pts)$
3) Soient $E_1, ..., E_p$ les axes principaux associés aux profils-colonnes X_c :

$$E_1^{\perp} =?, \ \ (E_2 \oplus ... \oplus E_p)^{\perp} =? \ (2pts)$$

4) Quelle est la relation entre la statistique χ^2 et les valeurs propres associées à la matrice $A_r := X_r^t X_c^t$? $(2pts)$
5) Justifier: $0 \leq \chi^2 \leq n\,(rg\,(A_r) - 1)$, où n est l'effectif total. $(2pts)$

1 Solution

1.1 Solution de l'Exercice 1

1) Les composantes principales sont centrées, donc évidement la moyenne $m_k = \overline{C}_k$ de chacune est nulle.

2) Variance de C_1 (avec $m_1 = 0$) :

$$\begin{aligned} var\left(C_1\right) &= \frac{1}{4}\sum_{i=1}^{4}\left(C_1\left(i\right) - m_1\right)^2 = \frac{1}{4}\left(\left(2.147\right)^2 + \left(-1.632\right)^2 + \left(-1.005\right)^2 + \left(0.489\right)^2\right). \\ &= 2.\,13, \end{aligned}$$

ainsi $var\left(C_1\right) = 2.\,13$. De la même manière on trouve $var\left(C_2\right) = 0.75$. Nous avons $var\left(C_i\right) = \lambda_i$ et

$$\sum_{i=1}^{3}\lambda_i = \text{Inertie totale} = \text{la trace de } R = 1 + 1 + 1 = 3.$$

Donc $2.\,13 + 0.75 + \lambda_3 = 3 \Longrightarrow \lambda_3 = 3 - \left(2.\,13 + 0.75\right) = 0.12$, ainsi $var\left(C_2\right) = 0.12$. En faisant un calcul direct on trouve

$$var\left(C_3\right) = \frac{1}{4}\left(\left(-0.283\right)^2 + \left(-0.059\right)^2 + \left(-0.230\right)^2 + \left(0.573\right)^2\right) = 0.11620.$$

Comme vous remarquez, nous avons eu deux valeurs différentes mais approximativement les mêmes. Ceci est naturel due aux erreurs d'arrondies dans les calculs, à savoir la recherche des zéros (v.p.) du polynôme caractéristique,... En conclusion les deux résultats sont justes $var\left(C_3\right) = 0.11620$ ou 0.12.

3) Les valeurs propres, $\lambda_1, \lambda_2, \lambda_3$ associées à la matrice R sont, respectivement, les variances de C_1, C_2 et C_3 :

$$\lambda_1 = 2.\,13,\ \lambda_2 = 0.75,\ \lambda_3 = 0.12\ \ (\text{ou } 0.11620)\,.$$

Juste une remarque: si vous calculez les valeurs propres de R pas la méthode du polynôme carateristique, vous allez trouver

$$\lambda_1 = 2.\,131\,0,\ \lambda_2 = 0.751\,83,\ \lambda_3 = 0.117\,18,$$

ce qui correspond à nos résultats.

4) Corrélation de C_1 avec C_2 :

$$cor\left(C_1, C_2\right) = \frac{cov\left(C_1, C_2\right)}{\sqrt{var\left(C_1\right)}\sqrt{var\left(C_2\right)}} = \frac{\frac{1}{4}C_1^t C_2}{\sqrt{2.\,13}\sqrt{0.75}}.$$

On a

$$\begin{aligned} C_1^t C_2 = \sum_{i=1}^{4} C_1\left(i\right) C_2\left(i\right) &= 2.147 \times \left(-0.330\right) + \left(-1.632\right) \times \left(-1.137\right) \\ &\quad + \left(-1.005 \times 1.247\right) + \left(0.489 \times 0.219\right) = 0.000\,93 \end{aligned}$$

Ainsi

$$cor\left(C_1, C_2\right) = \frac{0.000\,93}{\sqrt{2.\,13}\sqrt{0.75}} = 7.\,358 \times 10^{-4}.$$

D'autre part, on sait que les composantes principales sont non-corrélés donc $cor\left(C_1, C_2\right) \simeq 0$.

5) Les valeurs propres sont distinctes de la matrice R (symétrique à coefficients réels) donc, les vecteurs propres (u_1, u_2, u_3) sont deux à deux orthogonaux. De plus ces vecteurs sont normés, donc ils représentent les vecteurs directeurs des axes principaux: $E_k = vect\left(u_k\right)$, $k = 1, 2, 3$ (les sous-espaces engendrés par u_k).

6) On note par Z_1 la variable-Poids (centré-réduit). D'après le cours (d'ACP), nous avons

$$cor(Z_j, C_k) = \sqrt{\lambda_k}u_{kj},\ j = 1, 2, 3, 4;\ k = 1, 2, 3.$$

Donc

$$cor\left(Z_1, C_1\right) = \sqrt{\lambda_1}u_{11} = \sqrt{2.\,13} \times \left(-0.430\right) = -0.627,$$

$$cor\left(Z_1, C_2\right) = \sqrt{\lambda_2} u_{21} = \sqrt{0.75} \times (-0.649) = -0.562,$$

et

$$cor\left(Z_1, C_3\right) = \sqrt{\lambda_3} u_{31} = \sqrt{0.12} \times (-0.626) = -0.216.$$

7) Pourcentages d'inerties expliquées par les premiers axes principaux:

$$\frac{\lambda_1}{I_T} = \frac{\lambda_1}{trace(R)} = \frac{2.13}{3} = 0.71 = 71\%,$$

et

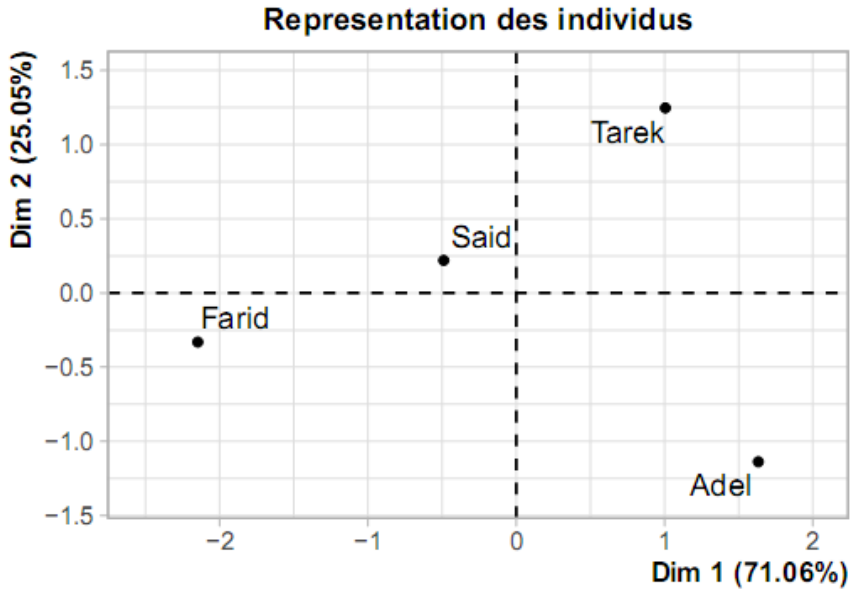
$$\frac{\lambda_2}{I_T} = \frac{\lambda_2}{trace(R)} = \frac{0.75}{3} = 0.25 = 25\%.$$

8) Pourcentage d'inertie expliquée par le premier plan principal:

$$\frac{\lambda_1 + \lambda_2}{I_T} = \frac{\lambda_1 + \lambda_2}{trace(R)} = \frac{2.13 + 0.75}{3} = 0.96 = 96\%.$$

On en déduit que les individus sont parfaitement représentés sur le premier plan principal avec une perte d'information relativement négligeable de 4%.

9) Représentation graphique des individus sur le premier plan principal:



1.2 Solution de l'Exercice 2

1) Les axes principaux u_k sont deux a deux M_r-orthogonaux, donc pour $u_k \neq g_r$, on a $g_r^t M_r u_k = 0$. Nous avons

$$0 = g_r^t M_r u_k = \left(\begin{array}{ccc} f_{\cdot 1}, & \cdots & , f_{\cdot q} \end{array} \right) \left(\begin{array}{ccc} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{array} \right) \left(\begin{array}{c} u_k(1) \\ \vdots \\ u_k(q) \end{array} \right).$$

Il est claire que

$$\left(\begin{array}{ccc} f_{\cdot 1}, & \cdots & , f_{\cdot q} \end{array} \right) \left(\begin{array}{ccc} 1/f_{\cdot 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/f_{\cdot q} \end{array} \right) = \underbrace{\left(\begin{array}{ccc} 1, & \cdots & , 1 \end{array} \right)}_{q \text{ éléments}}.$$

Donc

$$0 = \underbrace{\left(\begin{array}{ccc} 1, & \cdots & , 1 \end{array} \right)}_{q \text{ éléments}} \left(\begin{array}{c} u_k(1) \\ \vdots \\ u_k(q) \end{array} \right) = \sum_{j=1}^{q} u_k(j).$$

2) L'inertie de la modalité i des profils-lignes X_r par rapport à g_r :

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

3) Soient $E_1, ..., E_p$ les axes principaux associés aux profils-colonnes X_c :

$$E_1^{\perp} = E_2 \oplus ... \oplus E_p \text{ et } (E_2 \oplus ... \oplus E_p)^{\perp} = E_1.$$

4) La relation entre la statistique χ^2 et les valeurs propres

$$\frac{\chi^2}{n} = \sum_{k=1}^{rg(A_r)} \lambda_k - 1.$$

5) Nous avons $0 \leq \lambda_k \leq 1$, alors $0 \leq \sum_{k=1}^{\tau} \lambda_k \leq \tau$, où $\tau := rg(V_r M_r) = rg(A_r) - 1$, ainsi

$$0 \leq \chi^2 \leq n\,(rg(A_r) - 1).$$