27/4/20

Corrigé de la série 2

Exercice 1

1) Le centre de gravité du nuage de points est le point constitué des moyennes arithmétiques des composantes des variables : $g={}^t(\overline{X}^1,\overline{X}^2,\overline{X}^3)={}^t\left(\frac{22}{4},\frac{73}{4},\frac{165}{4}\right)={}^t(5.5,\ 18.25,\ 41.25)$.

2) La matrice des données-centrées n'est que la matrice $Z=(z_{ij})_{i,j}$ où $z_{ij}=y_{ij}/\sigma_i$ avec

$$Y=(y_{ij})_{i,j}=\begin{pmatrix} 5/2 & 47/4 & 55/4 \\ -7/2 & -49/4 & -5/4 \\ -1/2 & -13/4 & -45/4 \\ 3/2 & 15/4 & -5/4 \end{pmatrix}$$ est la matrice centrée des données

et σ_i représente l'écart type de la variable X^i pour $i=1,\ 2,\ 3$.

3) L'expression de la matrice des coefficients de corrélation est donnée par : $R=\frac{1}{4}({}^tZ\cdot Z).$

Elle est de type $(3,3)$.

4) La variance expliquée par les axes factoriels choisis pour le plan principal est exprimée par la valeur propre. Donc pour le premier axe principal, il correspond à la plus grande valeur et nous avons : $Var\left((\Delta u)\right)=2.4598$ et celle du deuxième axe est $Var\left((\Delta v)\right)=0.5368$.

5) Les axes factoriels choisis pour le plan principal sont générés par les deux vecteurs propres normés associés à ces deux valeurs propres citées dans la réponse 4). Nous vérifions tout d'abord que $\|\vec{u}_1\|=1$ et $\|\vec{u}_3\|=1$. Donc, le premier axe principal est (Δu_1) et le second est (Δu_3).

6) L'inertie totale expliquée par le plan factoriel est donnée par $\lambda_1+\lambda_3=2.9966$. En terme de pourcentage, nous pouvons dire que le plan explique $(\lambda_1+\lambda_3)/3=0.99886$ c'est-à-dire plus que 99% de l'inertie totale est expliqué par le plan ajustant le nuage de points.

7) Dans le cas de données centrées réduites, les composantes principales se calculent comme suit :

$C^k=Z\cdot u_k$, pour $k=1,\ 3$.

où $Z=(z_{ij})_{i,j}$ tel que : $z_{ij}=y_{ij}/\sigma_i$ pour tout $i=1,\ 2,\ 3$.

$\sigma_1 = \sqrt{\frac{1}{4}\left(\frac{84}{4}\right)} = \sqrt{5.25}$, $\sigma_2 = \sqrt{\frac{1}{4}\left(\frac{5004}{16}\right)} = \sqrt{78.1875}$,

$\sigma_3 = \sqrt{\frac{1}{4}\left(\frac{5100}{16}\right)} = \sqrt{79.6875}$.

Après calculs, nous trouvons :

$$C^1 = Z \cdot u_1 = {}^t(-2.2536,\ 1.8540,\ 0.9882,\ -0.5886)$$

$$C^3 = Z \cdot u_3 = {}^t(0.4984,\ 0.9137,\ -0.8891,\ -0.5230)$$

Ces deux composantes représentent les deux nouvelles variables.

Exercice 2 La matrice des données X est de type (4, 3).

1) Le centre de gravité est donné par $g = {}^t(\overline{X}^1, \overline{X}^2, \overline{X}^3) = {}^t\left(\frac{8}{4}, \frac{4}{4}, \frac{8}{4}\right) = {}^t(2, 1, 2).$

Alors, la matrice des variances-covariances est donnée par $V = \frac{1}{4}({}^tY \cdot Y)$ où Y est la matrice centrée.

De la matrice X, nous avons :

$$Y = {}^t\begin{pmatrix} -2 & 1 & 1 \\ -2 & -1 & 1 \\ 2 & 1 & -1 \\ 2 & -1 & -1 \end{pmatrix}.$$

Par conséquent :

$$V = \begin{pmatrix} 4 & 0 & -2 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

D'autre part, les écarts types des variables sont $\sigma(X^j) = \sqrt{\frac{1}{4}(\sum_{i=1}^{4} y_{ij})^2}$ pour $j = 1, 2, 3$:

$\sigma(X^1) = \sqrt{\frac{16}{4}} = 2$, $\sigma(X^2) = \sigma(X^3) = \sqrt{\frac{4}{4}} = 1$, et donc la matrice centrée-réduite Z est définie par : $Z = (z_{ij})_{i,j}$ avec $z_{ij} = \frac{y_{ij}}{\sigma(X^j)}$ pour $i = 1, 2, 3$ et $j = 1, 2, 3$.

D'où $$Z = \begin{pmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & -1 \end{pmatrix}.$$

Par conséquent,

la matrice des variances-covariances est donnée par : $$R = \frac{1}{4}({}'Z \cdot Z)$$

$$R = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

2) Les variances des différentes composantes principales sont les valeurs propres de la matrice des corrélations. Ces valeurs propres sont les racines du polynôme caractéristique $P_R(\lambda)$.

$$P_R(\lambda) = \det(R - \lambda.Id) = (1-\lambda)\times(1-\lambda)^2 - (1-\lambda) = (1-\lambda)\times\left[(1-\lambda)^2 - 1\right]$$

$P_R(\lambda) = 0$ donne les valeurs suivantes : $\lambda_1 = 1,\ \lambda_2 = 0$ et $\lambda_3 = 2$.

Var (Ci)=va.p (i)

3) Ces valeurs propres interprètent les inerties portées par chaque axe principal, l'inertie totale étant égale à la somme des valeurs propres :

$$Inertie\ totale = \lambda_1 + \lambda_2 + \lambda_3 = 1 + 0 + 2 = 3 = n$$

Alors, le taux d'inertie projeté sur chaque axe est donné par :

$$T_{\lambda_i} = \lambda_i / \sum_{i=1}^{3} \lambda_i = \lambda_i / 3 \quad , \quad i = 1,2,3 .$$

Donc, $T_{\lambda_1} = \lambda_1 / 3 = 1/3 \approx 0.333, \quad T_{\lambda_2} = \lambda_2 / 3 = 0/3 = 0, \quad T_{\lambda_3} = 2/3 \approx 0.666$

C'est-à-dire

$$T_{\lambda_1} \approx 33.33\%, \quad T_{\lambda_2} = 0\%, \quad T_{\lambda_3} \approx 66.66\% .$$

A partir de ces taux calculés, nous pouvons déduire la dimension du meilleur sous espace ajustant le nuage des individus. Le premier axe principal associé à la plus grande valeur propre recouvre plus que 66.66% de l'inertie totale, alors nous pouvons sélectionner un seul axe. D'autre part, nous remarquons que le taux d'inertie expliqué par les deux premiers axes est égal à $(\lambda_1 + \lambda_3)/3 = 1$ c'est à dire à 100% , et donc ça serait intéressant de le prendre aussi.

4) Pour déterminer les coordonnées des individus le long des axes choisis, il suffit de déterminer les **composantes principales** associées aux axes principaux choisis.

i) Le **1er axe principal** est associé à la **plus grande** valeur propre : $\lambda = 2$.

Soit $v = {}^t(x, y, z) \in \Re^3 - \{0_{\Re^3}\}$ tel que : $R \cdot v = \lambda \cdot v$, alors nous avons :

$$\begin{cases} x - z = 2 \cdot x \\ y = 2 \cdot y \\ -x + z = 2 \cdot z \end{cases} \Leftrightarrow \begin{cases} -x - z = 0 \\ y = 0 \\ -x - z = 0 \end{cases}$$ ce qui donne $z = -x$ et $y = 0$.

Donc $v = {}^t(x, 0, -x) = x \cdot {}^t(1, 0, -1)$ avec $x \in \Re^*$. En particulier, nous prenons $v = {}^t(1, 0, -1)$.

Par conséquent, le 1er axe principal est engendré par le vecteur unitaire $v_1 = v / \|v\|$:

$v_1 = {}^t(\sqrt{2}/2,\ 0,\ -\sqrt{2}/2)$ où $\|v\| = \sqrt{2}$ et $(\Delta v_1) = \langle \overline{v_1} \rangle$.

ii) Le **2ème axe principal** est associé à la valeur propre : $\lambda = 1$.

Nous suivons les mêmes étapes, en effet pour $u = {}^t(x, y, z) \in \Re^3_*$ tel que : $R \cdot u = \lambda \cdot u$, nous avons :

$$\begin{cases} x - z = x \\ y = y \\ -x + z = z \end{cases} \Leftrightarrow \begin{cases} z = 0 \\ y : quelconque \\ x = 0 \end{cases} .$$

ce qui donne $u = {}^t(0, y, 0) = y \cdot {}^t(0, 1, 0)$ pour y quelconques dans $\Re^*$. Donc $u = {}^t(0, 1, 0)$, et le 2ème axe principale est la droite $(\Delta u_2) = \langle \overline{u} \rangle$ car $\|u\| = 1$.

Ainsi, le meilleur plant ajustant le nuage des individus est le plan $\langle \overline{v_1}, \overline{u_2} \rangle$, et les

composantes principales sont : C^1 et C^2 définies comme suit :

$$C^1 = Z \cdot v_1 = {}^t(1, -1, 1, -1) \quad \text{et} \quad C^2 = Z \cdot u_2 = {}^t(-\sqrt{2}, -\sqrt{2}, \sqrt{2}, \sqrt{2}) .$$

Remarque Vous remarquez bien que pour cet exercice, nous avons considéré la matrice de corrélation pour calculer les valeurs propres et donc les espaces factoriels ; Il s'agit donc

d'une ACP normée.

5) Les corrélations entre les variables initiales et les composantes principales s'expriment en fonction de la relation suivante :

$$\varphi^k = \sqrt{\lambda_k} \cdot u_k \quad \text{pour k= 1, 2.}$$

$$\varphi_j^k = r(X^j, C^k) \text{, pour j=1, 2, 3.}$$

Ici dans notre cas, nous avons deux composantes principales associées à deux axes principaux, autrement dit à deux vecteurs propres normés :

$$v_1 = {}^t(\sqrt{2}/2,\ 0,\ -\sqrt{2}/2) \qquad u = {}^t(0,1,0) \ .$$

D'où ,

$$\varphi^1 = \sqrt{2} \cdot u_1 = \sqrt{2} \cdot v_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \quad \text{et} \quad \varphi^2 = \sqrt{1} \cdot u_2 = \sqrt{1} \cdot u = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} .$$

De ces deux vecteurs, nous avons les corrélations suivantes :

$$\varphi_1^1 = r(X^1, C^1) = 1 \quad \text{et} \quad \varphi_1^2 = r(X^1, C^2) = 0$$

$$\varphi_2^1 = r(X^2, C^1) = 0 \quad \text{et} \quad \varphi_2^2 = r(X^2, C^2) = 1$$

$$\varphi_3^1 = r(X^3, C^1) = -1 \quad \text{et} \quad \varphi_3^2 = r(X^3, C^2) = 0 \ .$$

Ce qui nous conduit à associer à chaque point variable un point dans le plan factoriel dont les coordonnées de ce point sont les corrélations entre cette variable et les axes principaux de ce plan.

Ainsi, à :

$$\overrightarrow{X^1} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \overrightarrow{X^2} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ et } \overrightarrow{X^3} \begin{pmatrix} -1 \\ 0 \end{pmatrix} .$$

Variables	**1er axe**	**2ème axe**
X^1	1	0
X^2	0	1
X^3	-1	0

Que nous pouvons présenter par un tableau (voir Tableau 1) et présenter dans le cercle des corrélations (voir figure 1).

Tableau 1. Coordonnées des variables dans le plan principal.

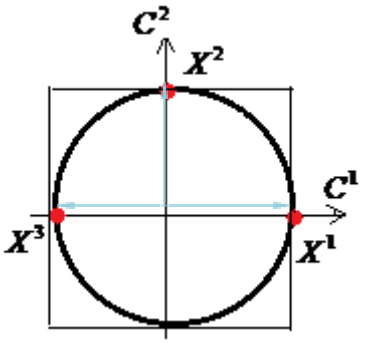

Figure 1. Représentation graphique des variables dans le plan factoriel.

Nous remarquons bien que l'axe 1 qui est la 1ère nouvelle variable C^1 (1ère composante principale) est positivement corrélée avec la variable X^1 et négativement avec la variable X^3 . X^1 et X^3 ont une corrélation égale à -1. D'autre part C^2 est positivement et fortement corrélée avec X^2 .

Exercice 3

1) De la matrice de corrélations donnée, nous tirons les informations suivantes :

- Toutes les variables sont corrélées positivement car les corrélations données sont positives. C'est-à-dire $\rho(X^i, X^j) = r_{ij} \geq 0$ pour tout i, j dans $\{1, 2, ..., 8\}$.

- Les variables X^5 et X^3 , X^7 et X^3 sont fortement corrélées ainsi que les variables X^7 et X^5 . Ceci est vrai si nous fixons le taux acceptable à 68 %. D'autre part, nous remarquons une faible corrélation entre la variable X^1 et les variables X^4 , X^3 , X^6 .

2) Le pourcentage d'inertie expliquée par chaque axe est donné en fonction des valeurs propres déterminés à partir de la matrice de corrélation R . Il est défini par :

$$T_{\lambda_i} = \lambda_i / n = \lambda_i / 8 \quad \text{pour } i = 1, 2, ..., 8 .$$

Ainsi, les différents taux sont donnés dans l'ordre comme suit :

$$T_{\lambda_i} \approx 55.30\%,\ T_{\lambda_2} \approx 14.21\%,\ T_{\lambda_3} \approx 7.35\%,\ T_{\lambda_4} \approx 6.63\%,\ T_{\lambda_5} \approx 5.72\%,$$
$$T_{\lambda_6} \approx 4.70\%,\ T_{\lambda_7} \approx 3.23\% \text{ et } T_{\lambda_8} \approx 2.88\%.$$

3) Le meilleur sous espace principal de dimension 2 ajustant le nuage des individus est le plan engendré par les deux vecteurs propres normés associés aux deux premières plus grandes valeurs propres de la matrice de corrélation :

$$\lambda_1 = 4.4242 \qquad \lambda_2 = 1.1366$$

C'est-à-dire :

$$Plan = (\underset{\sqcup}{u}, \underset{\sqcup}{v})$$

Où les deux vecteurs doivent être normés et vérifient :

$$R \cdot \underset{\sqcup}{u} = \lambda_1 \cdot \underset{\sqcup}{u} \qquad R \cdot \underset{\sqcup}{v} = \lambda_2 \cdot \underset{\sqcup}{v}$$

Le taux d'inertie expliqué par ce plan ajustant est donné par :

$$T_{\lambda_6+\lambda_2} = (\lambda_1 + \lambda_2)/8 \approx 69.51\%$$

4) Les coordonnées des individus dans le plan principal ne sont que les projections des individus sur les axes principaux de ce plan. Autrement dit, elles ne sont que les coordonnées des composantes principales, et donc elles sont données par :

$$C^1 = Z \cdot \underset{\sqcup}{u} \text{ et } C^2 = Z \cdot \underset{\sqcup}{v}.$$

Où C^1 et C^2 sont composées des coordonnées des individus sur le premier axe et le deuxième axe du plan ajustant respectivement.

Les coordonnées des variables dans le plan principal ne sont que les composantes des composantes principales, et donc elles sont données par : $\varphi^k = \sqrt{\lambda_k} \cdot u_k$ pour k= 1, 2.

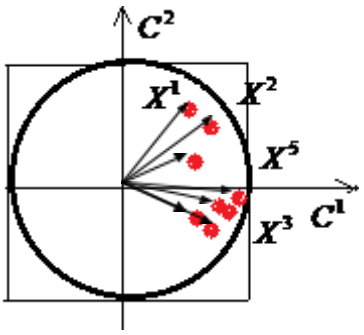

Figure 3. Représentation des variables dans le plan factoriel.

Ce cercle nous permet de résumer les liaisons entre les variables, afin de pouvoir les rassembler dans des groupes séparés. Les petites flèches correspondent aux variables faiblement corrélées aves les deux axes du plan factoriel. Nous ne nous intéressons qu'aux variables qui sont proches du périmètre du cercle.