ACT6100 H2019	Analyse de données Solutions 1

Rappels d'algèbre linéaire

1. Pour deux matrices $\mathbf{A}$ (3×2) et $\mathbf{B}$ (2×3), on a

$$\mathbf{AB} = \begin{bmatrix} 3 & 12 & 6 \\ 5 & -2 & 8 \\ 4 & 5 & 7 \end{bmatrix}.$$

   Par exemple, l'élément en position $(1, 1)$ de cette matrice a été obtenu à l'aide du calcul

$$3 = (3)(1) + (0)(3).$$

   Enfin, on note que de façon générale, $\mathbf{AB} \neq \mathbf{BA}$, mais que $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$ est toujours vérifiée. Le code informatique permettant d'obtenir le résultat est présenté à la Figure 1.

```
A <- matrix(c(3, 0, -1, 2, 1, 1),
            ncol = 2, byrow = TRUE)
B <- matrix(c(1, 4, 2, 3, 1, 5),
            ncol = 3, byrow = TRUE)
A %*% B
```

FIGURE 1 – Code informatique.

2. Pour résoudre l'exercice, on transpose la matrice $\mathbf{A}$

$$\mathbf{A}^T = \begin{bmatrix} 4 & 0 \\ 2 & 9 \end{bmatrix}$$

   et on réalise ensuite la somme pour obtenir la matrice

$$\mathbf{A}^T + \mathbf{B} = \begin{bmatrix} 11 & 0 \\ 5 & 10 \end{bmatrix}.$$

   Le code informatique permettant d'obtenir le résultat est présenté à la Figure 2.

```
A <- matrix(c(4, 2, 0, 9),
            ncol = 2, byrow = TRUE)
B <- matrix(c(7, 0, 3, 1),
            ncol = 2, byrow = TRUE)
t(A) + B
```

FIGURE 2 – Code informatique.

3. La trace d'une matrice $\mathbf{A}$ $(n \times n)$ est simplement la somme des éléments de sa diagonale qui est, dans cet exercice, $\text{tr}(\mathbf{Z}) = 1 + 9 + 9 = 19$.
   Le code informatique permettant d'obtenir le résultat est présenté à la Figure 3.
4. De façon générale, pour une matrice $\mathbf{A}$ $(p \times p)$, les valeurs propres sont les solutions de l'Équation

$$\det(\mathbf{A} - \lambda\mathbf{I}_p) = \mathbf{0}.$$

```
Z <- matrix(c(1, 2, 5, 3, 9, 6, 1, 2, 9),
            ncol = 3, byrow = TRUE)
sum(diag(Z))
```

FIGURE 3 – Code informatique.

Dans cet exercice, on a

$$\begin{aligned}
\det(\mathbf{A} - \lambda \mathbf{I}_p) &= \mathbf{0} \\
\det\left(\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) &= 0 \\
\det\left(\begin{bmatrix} 2-\lambda & 1 \\ 0 & 3-\lambda \end{bmatrix}\right) &= 0 \\
(2-\lambda)(3-\lambda) &= 0 \\
\lambda_1 &= 2 \\
\lambda_2 &= 3.
\end{aligned}$$

Les valeurs propres sont donc 2 et 3.
Le code informatique permettant d'obtenir le résultat est présenté à la Figure 4.

```
A <- matrix(c(2, 1, 0, 3),
            ncol = 2, byrow = TRUE)
eigen(A)
```

FIGURE 4 – Code informatique.

5. Pour une matrice $\mathbf{A}$ $(p \times p)$ dont les valeurs propres sont $\lambda_1, \ldots, \lambda_p$, on a la propriété

$$\sum_{i=1}^{p} \lambda_i = \text{tr}(\mathbf{A}).$$

Dans cet exercice, on a donc

$$\begin{aligned}
\sum_{i=1}^{4} \lambda_i &= \text{tr}(\mathbf{H}) \\
&= 2 + 3 + 0 + 11 \\
&= 16.
\end{aligned}$$

6. Trois vecteurs sont linéairement dépendants si l'un est une combinaison linéaire des deux autres. Si c'est le cas, le déterminant de la matrice formée par les vecteurs sera nul. Ici,

$$\det\begin{bmatrix} -3 & 5 & 1 \\ 0 & -1 & 1 \\ 4 & 2 & 3 \end{bmatrix} = 39 \neq 0$$

ce qui implique que les vecteurs sont linéairement indépendants.

Analyse en composantes principales

1. On a

$$\begin{aligned} E[\mathbf{Y}] &= E\Big[\boldsymbol{\Gamma}^T(\mathbf{X}-\boldsymbol{\mu})\Big] \qquad \text{(par définition)} \\ &= \boldsymbol{\Gamma}^T E[\mathbf{X}-\boldsymbol{\mu}] \\ &= \boldsymbol{\Gamma}^T(E[\mathbf{X}]-\boldsymbol{\mu}) \qquad \text{(puisque } E[\text{cte}]=\text{cte}) \\ &= \boldsymbol{\Gamma}^T(\boldsymbol{\mu}-\boldsymbol{\mu}) \qquad \text{(puisque } E[\mathbf{X}]=\boldsymbol{\mu}) \\ &= \mathbf{0}_p, \end{aligned}$$

et

$$\begin{aligned} \text{Var}[\mathbf{Y}] &= \text{Var}\Big[\boldsymbol{\Gamma}^T(\mathbf{X}-\boldsymbol{\mu})\Big] \qquad \text{(par définition)} \\ &= \boldsymbol{\Gamma}^T\text{Var}[\mathbf{X}-\boldsymbol{\mu}]\,\boldsymbol{\Gamma} \qquad (\text{car Var}[\mathbf{aX}]=\mathbf{a}\text{Var}[\mathbf{X}]\,\mathbf{a}^T) \\ &= \boldsymbol{\Gamma}^T\text{Var}[\mathbf{X}]\,\boldsymbol{\Gamma}+\boldsymbol{\Gamma}^T\text{Var}[\boldsymbol{\mu}]\,\boldsymbol{\Gamma} \qquad (\text{car Var}[-\mathbf{X}]=(-1)^2\text{Var}[\mathbf{X}]=\text{Var}[\mathbf{X}]) \\ &= \boldsymbol{\Gamma}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}+\mathbf{0}_p \qquad (\text{car Var}[\text{cte}]=0) \\ &= \boldsymbol{\Gamma}^T\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} \end{aligned}$$

et puisque $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}=\mathbf{I}_p$,

$$= \boldsymbol{\Lambda}.$$

Ainsi, le vecteur $\mathbf{Y}$ est centré (espérance nulle) et sa matrice de variance-covariance est diagonale.

2. (a) Les valeurs propres de la matrice $\boldsymbol{\Sigma}$ sont les valeurs λ telles que

$$\begin{aligned} \det(\boldsymbol{\Sigma}-\lambda\mathbf{I}_2) &= \mathbf{0} \\ \det\begin{bmatrix} 1-\lambda & \tau \\ \tau & 1-\lambda \end{bmatrix} &= 0 \\ (1-\lambda)^2-\tau^2 &= 0 \\ \lambda &= 1\pm\tau. \end{aligned}$$

(b) La matrice de variance-covariance devient

$$\begin{aligned} \boldsymbol{\Sigma}^* &= \text{Cov}\big[(cX_1, X_2)^T\big] \\ &= \begin{bmatrix} \text{Cov}[cX_1, cX_1] & \text{Cov}[cX_1, X_2] \\ \text{Cov}[cX_1, X_2] & \text{Cov}[X_2, X_2] \end{bmatrix} \\ &= \begin{bmatrix} c^2\text{Cov}[X_1, X_1] & c\text{Cov}[X_1, X_2] \\ c\text{Cov}[X_1, X_2] & \text{Cov}[X_2, X_2] \end{bmatrix} \\ &= \begin{bmatrix} c^2 & c\tau \\ c\tau & 1 \end{bmatrix}. \end{aligned}$$

Les nouvelles valeurs propres sont les valeurs λ solutions de

$$\begin{aligned} \det(\boldsymbol{\Sigma}^*-\lambda\mathbf{I}_2) &= \mathbf{0} \\ \det\begin{bmatrix} c^2-\lambda & c\tau \\ c\tau & 1-\lambda \end{bmatrix} &= 0 \\ \lambda &= 0.5\left(c^2+1\pm\sqrt{(c^2-1)^2+4c^2\tau^2}\right). \end{aligned}$$

Le vecteur propre correspondant à la plus grande valeur propre λ_1 est solution de

$$\begin{aligned} \boldsymbol{\Sigma}^*\mathbf{x} &= \lambda_1\mathbf{x} \\ \begin{bmatrix} c^2 & c\tau \\ c\tau & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \lambda_1\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \end{aligned}$$

On obtient les équations

$$\begin{aligned} c^2 x_1 + c\tau x_2 &= \lambda_1 x_1 \\ c\tau x_1 + x_2 &= \lambda_1 x_2, \end{aligned}$$

ce qui implique que $x_1/x_2 = (\lambda_1 - 1)/c\tau$. On observe que la fonction

$$\begin{aligned} \frac{x_1}{x_2} &= \frac{(\lambda_1 - 1)}{c\tau} \\ &= \frac{0.5\left(c^2 + 1 + \sqrt{(c^2-1)^2 + 4c^2\tau^2}\right) - 1}{c\tau} \end{aligned}$$

est une fonction croissante en c et donc, que le ratio x_1/x_2 est une fonction croissante en c. On peut donc conclure que lorsque c augmente, la première valeur propre λ_1 devient plus grande et la variable cX_1 gagne en importance dans la première composante (uniquement à la suite d'un changement d'échelle). Cela illustre l'importance de travailler avec la matrice des corrélations plutôt qu'avec la matrice de variance-covariance lorsque les différentes variables sont présentées avec des unités différentes.

3. En utilisant les définitions, on a

$$\begin{aligned} \mathcal{I}(\mathbf{X}) &= \sum_{i=1}^{n} w_i d^2_{\mathbf{M}}(\mathbf{x}_i^T, \bar{\mathbf{x}}) \\ &= \sum_{i=1}^{n} w_i ||\mathbf{x}_i^T - \bar{\mathbf{x}}||^2_{\mathbf{M}} \\ &= \sum_{i=1}^{n} w_i < \mathbf{x}_i^T - \bar{\mathbf{x}}, \mathbf{x}_i^T - \bar{\mathbf{x}} >_{\mathbf{M}} \\ &= \sum_{i=1}^{n} w_i \left(\mathbf{x}_i^T - \bar{\mathbf{x}}\right)^T \mathbf{M} \left(\mathbf{x}_i^T - \bar{\mathbf{x}}\right) \end{aligned}$$

qui devient, en prenant une métrique diagonale,

$$= \sum_{i=1}^{n} w_i \left(\mathbf{x}_i^T - \bar{\mathbf{x}}\right)^T \begin{bmatrix} m_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & m_k \end{bmatrix} \left(\mathbf{x}_i^T - \bar{\mathbf{x}}\right)$$

qui devient, en posant $w_i = 1/n$, $\forall i$,

$$\begin{aligned} &= \sum_{i=1}^{n} (1/n) \sum_{j=1}^{k} m_j (x_{ij} - \bar{x}^j)^2 \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} (1/n) m_j (x_{ij} - \bar{x}^j)^2 \\ &= \sum_{j=1}^{k} m_i \mathrm{Var}\left[\mathbf{x}^j\right]. \end{aligned}$$

En posant, $\mathbf{M} = \mathbf{I}_k$, on obtient

$$\begin{aligned} \mathcal{I}(\mathbf{Y}) &= \sum_{j=1}^{k} \mathrm{Var}\left[\mathbf{y}^j\right] \\ \mathcal{I}(\mathbf{Z}) &= \sum_{j=1}^{k} \mathrm{Var}\left[\mathbf{z}^j\right] \\ &= \sum_{j=1}^{k} (1) = k. \end{aligned}$$

4. (a) De façon générale, en travaillant avec la matrice de corrélation, on a $\sum_{j=1}^{p} \lambda_j = p$. Ainsi, $p = 8 = 1.93 + 0.13 + 0.07 + 0.02 + 5.28 + 0.41 + 0.12 + K$, et donc $K = 0.04$.

   (b) La variabilité expliquée par les trois premières composantes principales est donnée respectivement par λ_1, λ_2 et λ_3, où $\lambda_1 \geq \lambda_2 \geq \lambda_3$. La proportion de la variabilité initiale expliquée par les trois premières composantes est donc donnée par

$$\begin{aligned} \frac{\lambda_1 + \lambda_2 + \lambda_3}{\sum_{j=1}^{8} \lambda_j} &= \frac{5.28 + 1.93 + 0.41}{8} \\ &= 0.9525 \end{aligned}$$

   (c) De façon générale, $\mathbf{Ra} = \lambda\mathbf{a}$, et donc, pour a_i, on a

$$\begin{bmatrix} 1.00 & -0.20 & -0.60 & -0.88 & 0.71 & 0.74 & 0.88 & 0.81 \\ -0.20 & 1.00 & 0.61 & 0.29 & 0.33 & 0.36 & -0.20 & 0.25 \\ -0.60 & 0.61 & 1.00 & 0.76 & -0.27 & -0.27 & -0.45 & -0.38 \\ -0.88 & 0.29 & 0.76 & 1.00 & -0.70 & -0.68 & -0.84 & -0.81 \\ 0.71 & 0.33 & -0.27 & -0.70 & 1.00 & 0.92 & 0.66 & 0.90 \\ 0.74 & 0.36 & -0.27 & -0.68 & 0.92 & 1.00 & 0.68 & 0.93 \\ 0.88 & -0.20 & -0.45 & -0.84 & 0.66 & 0.68 & 1.00 & 0.80 \\ 0.81 & 0.25 & -0.38 & -0.81 & 0.90 & 0.93 & 0.80 & 1.00 \end{bmatrix} \begin{bmatrix} -0.117 \\ 0.697 \\ 0.491 \\ 0.204 \\ 0.284 \\ 0.298 \\ -0.091 \\ 0.200 \end{bmatrix}$$

$$= \lambda \begin{bmatrix} -0.117 \\ 0.697 \\ 0.491 \\ 0.204 \\ 0.284 \\ 0.298 \\ -0.091 \\ 0.200 \end{bmatrix}.$$

   En particulier,

$$\begin{aligned} (1.00)(-0.117) + (-0.20)(0.697) + \ldots + (0.81)(0.200) &= \lambda(-0.117) \\ -0.22644 &= \lambda(-0.117) \\ \lambda &= 1.935. \end{aligned}$$

   Ainsi, le vecteur propre $\mathbf{a}_i$ correspond à la valeur propre 1.93 (la petite différence est due aux arrondissements). De la même façon, on trouve que le vecteur propre $\mathbf{a}_j$ correspond à la valeur propre 5.28.

   (d) La i^{e} composante principale est une somme pondérée des variables initiales (éventuellement centrées et standardisées), les poids étant donnés par le vecteur propre correspondant à la valeur propre λ_i. Ainsi, $y^1 = (0.406)X_1 + (-0.016)X_2 + \ldots + (0.410)X_8$.

5. (a) La carte initiale des individus est présentée à la Figure 5.

   (b) Les éléments de la matrice des corrélations sont donnés par $r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$. On obtient alors

$$\mathbf{R} = \begin{bmatrix} 1 & 0.5622757 \\ 0.5622757 & 1 \end{bmatrix}.$$

   (c) Les valeurs propres λ sont données par les solutions de l'équation $\det(\mathbf{R} - \lambda\mathbf{I}_2) = \mathbf{0}$. On a donc

$$\begin{aligned} \det \begin{bmatrix} 1-\lambda & 0.5622757 \\ 0.5622757 & 1-\lambda \end{bmatrix} &= 0 \\ (1-\lambda)^2 - (0.5622757)^2 &= 0 \\ (1-\lambda)^2 &= (0.5622757)^2 \\ \lambda_1 &= 1.5622757 \\ \lambda_2 &= 0.4377243. \end{aligned}$$

   On a bien $\lambda_1 + \lambda_2 = 2$.

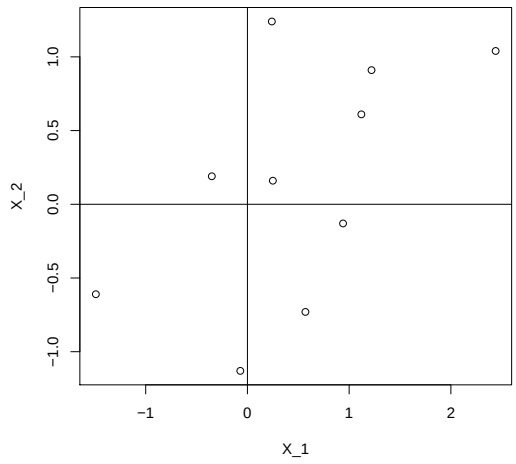

FIGURE 5 – Carte initiale des individus.

Les vecteurs propres $\mathbf{a}_1 = \begin{bmatrix} a_{11} & a_{12} \end{bmatrix}^T$ et $\mathbf{a}_2 = \begin{bmatrix} a_{21} & a_{22} \end{bmatrix}^T$ sont les solutions de $\mathbf{R}\mathbf{a}_i = \lambda_i \mathbf{a}_i$ telles que $\mathbf{a}_i^T \mathbf{a}_i = 1$. On a donc

$$\mathbf{R}\mathbf{a}_1 = \lambda_1 \mathbf{a}_1$$
$$\begin{bmatrix} 1 & 0.5622757 \\ 0.5622757 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = 1.5622757 \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix}.$$

Pour respecter la contrainte de normalisation ($\mathbf{a}^T\mathbf{a} = 1$), on doit avoir $a_{11}^2 + a_{12}^2 = 1$ et donc $a_{12} = \pm\sqrt{1 - a_{11}^2}$. Ainsi,

$$a_{11} = 0.7071$$
$$a_{12} = 0.7071.$$

De la même façon, on trouve que $a_{21} = -0.7071$ et $a_{22} = 0.7071$ [1].

(d) On a

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.5622757}{2}$$
$$= 0.78114.$$

(e) Les valeurs initiales des variables pour l'individu 7 sont 0.25 et 0.16. Il faut premièrement centrer et standardiser ces valeurs. Pour ce faire, on calcule $\bar{X}_1 = 0.487$, $\bar{X}_2 = 0.155$, $s_1 = 1.000040$ et $s_2 = 0.764176$. Pour faire comme R, on définit

$$s_j^2 = \frac{\sum_i^n (X_i - \bar{X}_j)^2}{n}$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_j)^2.$$

On obtient ainsi

$$X_1^{CS} = \frac{0.25 - 0.487}{1.000040}$$
$$= -0.2369$$
$$X_2^{CS} = \frac{0.16 - 0.155}{0.764176}$$
$$= 0.00654.$$

1. Il faut noter que le couple de vecteur propres $-\mathbf{a}_1$ et $-\mathbf{a}_2$ serait également une solution possible puisque les vecteurs propres sont déterminés au signe près.

La première composante principale est donnée par $y^1 = (0.7071)X_1^{CS} + (0.7071)X_2^{CS}$, et donc, la coordonnée de l'individu 7 sur la carte finale des individus (carte unidimensionnelle, c'est-à-dire une droite) est

$$\begin{aligned} y_7^1 &= (0.7071)(-0.2369) + (0.7071)(0.00654) \\ &= -0.1628876. \end{aligned}$$

(f) On a

$$\begin{aligned} \text{Cor}(X_1, y^2) &= \sqrt{\lambda_2} a_{12} \\ &= \sqrt{(0.4377243)}(0.7071) \\ &= 0.4678226. \end{aligned}$$

6. (a) La Figure 6 présente le code permettant d'obtenir les valeurs propres. Ainsi, deux composantes

```
data <- Rank
rownames(data) <- data[,2]
data <- data[,-c(1,2,3,4,5,6,13)]

res.pca <- PCA(data, scale.unit = TRUE, ncp = 5, graph = FALSE)
round(get_eigenvalue(res.pca),2)

      eigenvalue variance.percent cumulative.variance.percent
Dim.1       3.94            65.64                        65.64
Dim.2       1.09            18.13                        83.78
Dim.3       0.47             7.86                        91.63
Dim.4       0.26             4.28                        95.92
Dim.5       0.13             2.13                        98.04
Dim.6       0.12             1.96                       100.00
```

FIGURE 6 – Code informatique.

seront conservées puisque qu'elles sont les seules à avoir des valeurs propres supérieures à 1.

(b) À partir des résultats de la Figure 6, on obtient 91.63 %.

(c) La Figure 7 présente le code permettant d'obtenir les carrés des cosinus des variables. La

```
summary(res.pca)

Variables
          Dim.1    ctr   cos2    Dim.2    ctr   cos2
Alumni |  0.830 17.474  0.688 | -0.087  0.696  0.008 |
Award  |  0.838 17.820  0.702 | -0.441 17.857  0.194 |
HiCi   |  0.863 18.922  0.745 |  0.286  7.509  0.082 |
N.S    |  0.936 22.264  0.877 |  0.062  0.356  0.004 |
SCI    |  0.514  6.697  0.264 |  0.823 62.224  0.677 |
Size   |  0.814 16.822  0.663 | -0.352 11.358  0.124 |
```

FIGURE 7 – Code informatique.

variable **SCI** avec seulement 26.4% de l'information préservée.

(d) À partir des résultats de la Figure 6, on a 100 % − 83.76 % = 16.24 % ≈ 16 %.

(e) À partir des résultats de la Figure 6, on a $(0.8377660)^2 + (-0.44078533)^2 \approx 0.90$ ou $0.702 + 0.194 \approx 0.90$.

(f) Le cercle des corrélations est présenté à la Figure 8. La variable **Alumni** car il s'agit de la variable la plus éloignée du cercle (la flèche est la plus courte).

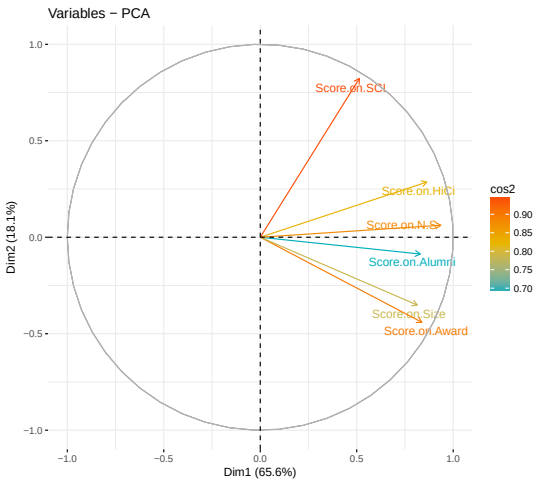

FIGURE 8 – Cercle des corrélations.

7. On remarque que les deux premières dimensions permettent d'expliquer une proportion de $0.6597 + 0.2417 = 0.9014$ de la variabilité initiale des données.

   Toutes les variables sont bien expliquées par les deux premières composantes principales puisqu'elles se retrouvent près du cercle.

   La première composante permet de séparer les véhicules coûteux et ayant des designs intéressants (gauche) et les véhicules à faibles prix et d'allures moins «sportives» (droite). Par exemples, BMW, Mercedes et Ferrari se retrouvent dans la moitié gauche et Lada et Fiat dans la moitié droite.

   La seconde composante permet de séparer les véhicules peu gourmands en essence et simples d'utilisation (bas) et ceux plus gourmands et plus complexes d'utilisation (haut). Par exemples, Volkswagen et Opel se retrouvent dans la moitié inférieure et Ferrari, Wartburg et Jaguar se retrouvent dans la moitié supérieure.

   Il est à noter dans cet exercice que les graphiques obtenus sont inversés (gauche-droite, haut-bas) en raison du type d'encodage des réponses (les petites valeurs étant les meilleures et les grandes valeurs étant les moins bonnes).

8. (a) La Figure 9 présente le code permettant d'obtenir la matrice des corrélations. Ainsi, la corré-

```
round(cor(data), 2)

            man chefserv inge banquier vendeur ouvrier indicateur
man        1.00     0.39 0.56     0.45    0.71    0.87       0.65
chefserv   0.39     1.00 0.89     0.87    0.70    0.46       0.47
inge       0.56     0.89 1.00     0.86    0.81    0.70       0.50
banquier   0.45     0.87 0.86     1.00    0.81    0.58       0.43
vendeur    0.71     0.70 0.81     0.81    1.00    0.80       0.43
ouvrier    0.87     0.46 0.70     0.58    0.80    1.00       0.61
indicateur 0.65     0.47 0.50     0.43    0.43    0.61       1.00
```

FIGURE 9 – Code informatique.

   lation demandée est 0.87.

   (b) La Figure 10 présente le code permettant d'obtenir les valeurs propres. On obtient ainsi une perte d'information de $1 - 0.752556519 - 0.166131552 = 0.08131193$, c'est-à-dire environ 8.13%.

   (c) La Figure 11 présente le code permettant d'obtenir la qualité de la représentation des variables dans le premier plan factoriel (un plan = 2 dimensions !). Si on utilise l'option 1, il faut mesurer la longueur des flèches sur le graphique et choisir la variable dont la flèche est la plus longue (voir Figure 12). Si on utilise l'option 2, il faut additionner les carrés des cosinus des deux

```
### Première option
data <- data[,-7]
eigen(cor(data))$values/6

[1] 0.752556519 0.166131552 0.033411409 0.024035948 0.016368766 0.007495805

### Deuxième option
data <- data[,-7]
res.pca <- PCA(data, scale.unit = TRUE, ncp = 5, graph = FALSE)
get_eigenvalue(res.pca)

      eigenvalue variance.percent cumulative.variance.percent
Dim.1 4.51533912       75.2556519                    75.25565
Dim.2 0.99678931       16.6131552                    91.86881
Dim.3 0.20046846        3.3411409                    95.20995
Dim.4 0.14421569        2.4035948                    97.61354
Dim.5 0.09821260        1.6368766                    99.25042
Dim.6 0.04497483        0.7495805                   100.00000
```

FIGURE 10 – Code informatique.

```
### Option 1
fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)

### Option 2
summary(res.pca)

Variables (j'ai gardé uniquement les colonnes cos2 des
deux premières dimensions)

               Dim.1    Dim.2
man         |  0.571 |  0.354 |
chefserv    |  0.702 |  0.234 |
inge        |  0.871 |  0.049 |
banquier    |  0.789 |  0.127 |
vendeur     |  0.871 |  0.007 |
ouvrier     |  0.711 |  0.226 |
```

FIGURE 11 – Code informatique.

premières dimensions. On obtient ainsi que la variable **ouvrier** est la mieux représentée avec 93.7% de l'information préservée.

(d) La Figure 13 présente le code permettant d'obtenir la qualité de la représentation des observations dans le premier plan factoriel. Si on utilise l'option 1, il faut rechercher l'observation dont le cercle est plus petit (voir Figure 14). Si on utilise l'option 2, il faut additionner les carrés des cosinus des deux premières dimensions. On obtient ainsi que l'observation **Hong Kong** est la moins bien représentée.

9. (a) Pour les variables aléatoires U_1 et U_2, on a

$$E[U_1] = E[U_2] = 0.5$$
$$\text{Var}[U_1] = \text{Var}[U_2] = 1/12.$$

En utilisant l'indépendances entre les variables U_1 et U_2 (uniquement pour le calcul des va-

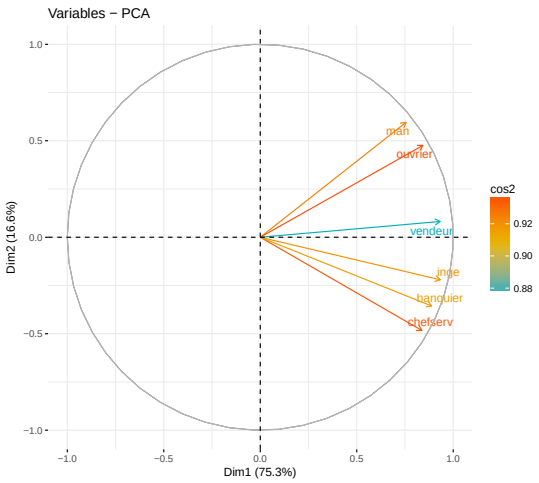

FIGURE 12 – Cercle des corrélations.

```
### Option 1
fviz_pca_ind (res.pca, pointsize = "cos2",
              pointshape = 21, fill = "#E7B800",
              repel = TRUE)

### Option 2
summary(res.pca)
```

FIGURE 13 – Code informatique.

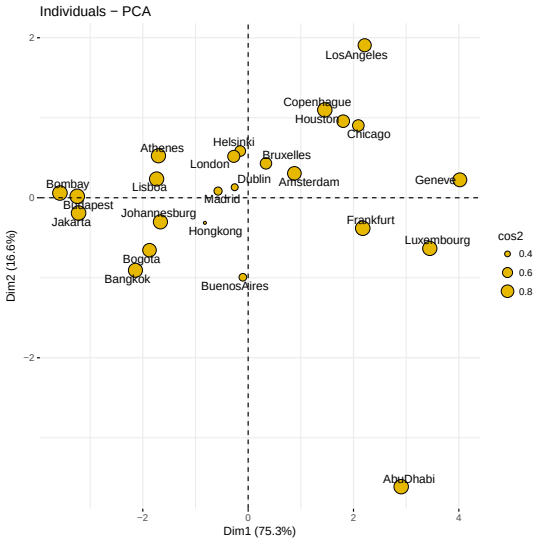

FIGURE 14 – Carte des observations.

riances et des covariances), on obtient

$$E[X_3] = E[U_1] + E[U_2] = 1$$
$$E[X_4] = E[U_1] - E[U_2] = 0$$
$$\mathrm{Var}[X_3] = \mathrm{Var}[X_4] = \mathrm{Var}[U_1] + \mathrm{Var}[U_2] = 1/6.$$

Pour les covariances, on a (je ne fais pas tous les calculs)

$$\begin{aligned}\text{Cov}[X_1, X_3] &= \text{Cov}[U_1, U_1 + U_2] = \text{Var}[U_1] + 0 = 1/12\\ \text{Cov}[X_3, X_4] &= \text{Cov}[U_1 + U_2, U_1 - U_2]\\ &= \text{Var}[U_1] - \text{Var}[U_2] = 0.\end{aligned}$$

Ainsi, on obtient la matrice de variance-covariance

$$\text{Var}[\mathbf{X}] = \left(\frac{1}{12}\right)\begin{bmatrix}1 & 0 & 1 & 1\\ 0 & 1 & 1 & -1\\ 1 & 1 & 2 & 0\\ 1 & -1 & 0 & 2\end{bmatrix}$$

que l'on transforme en une matrice des corrélations

$$\mathbf{R} = \begin{bmatrix}1 & 0 & 1/\sqrt{2} & 1/\sqrt{2}\\ 0 & 1 & 1/\sqrt{2} & -1/\sqrt{2}\\ 1/\sqrt{2} & 1/\sqrt{2} & 1 & 0\\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 1\end{bmatrix}.$$

(b) On sait que $\mathbf{R}\mathbf{v}_i = \lambda_i \mathbf{v}_i$ pour $i = 1, 2$. On a donc

$$\mathbf{R}\mathbf{v}_1 = \begin{bmatrix}1 & 0 & 1/\sqrt{2} & 1/\sqrt{2}\\ 0 & 1 & 1/\sqrt{2} & -1/\sqrt{2}\\ 1/\sqrt{2} & 1/\sqrt{2} & 1 & 0\\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 1\end{bmatrix}\begin{bmatrix}1/\sqrt{2}\\ 1/\sqrt{2}\\ 1\\ 0\end{bmatrix} = \begin{bmatrix}\sqrt{2}\\ \sqrt{2}\\ 2\\ 0\end{bmatrix} = \lambda_1 \begin{bmatrix}1/\sqrt{2}\\ 1/\sqrt{2}\\ 1\\ 0\end{bmatrix}.$$

On obtient ainsi que $\lambda_1 = 2$. De la même façon, on obtient que $\lambda_2 = 2$. La somme des valeurs propres est égale à la trace de la matrice $\mathbf{R}$. Ainsi, $\lambda_3 = \lambda_4 = 0$. On aura donc 50% de l'information dans chacune des deux premières dimensions et 0% dans les 3$^{\text{e}}$ et 4$^{\text{e}}$ dimensions.

10. (a) **Solution simple.** Les composantes principales sont centrées, c'est-à-dire $\bar{Y} = 0$. On a donc $\sum_{i=1}^{6} Y_{i2} = 0$ et on trouve que $Y_{62} = -0.20$. **Solution plus complexe.** Comme l'ACP a été réalisée à partir de la matrice des corrélations, on sait que la somme des valeurs propres sera égale au nombre de variables, c'est-à-dire 5. À partir d'un des deux graphiques (peu importe lequel), on trouve les deux premières valeurs propres :

$$\begin{aligned}0.6066 &= \frac{\lambda_1}{5} \rightarrow \lambda_1 = 3.033\\ 0.2586 &= \frac{\lambda_2}{5} \rightarrow \lambda_2 = 1.293.\end{aligned}$$

Pour pouvoir calculer la valeur manquante, on aura besoin du vecteur propre associé à λ_2. On peut l'obtenir à partir du tableau des corrélations en se rappelant que la corrélation entre la variable i et la composante j est

$$\begin{aligned}\text{Cor}[i, j] &= \sqrt{\lambda_j} a_{ij}\\ \rightarrow 0.31 &= \sqrt{\lambda_2} a_{12}\\ &= \sqrt{1.293} a_{12}\\ \rightarrow a_{12} &= 0.272623\\ a_{22} &= 0.4660973\\ a_{32} &= 0.4485087\\ a_{42} &= -0.09673718\\ a_{52} &= -0.7035431.\end{aligned}$$

Avant de calculer Y_{26}, il faut centrer et réduire les variables. On obtient alors

$$\mathbf{Z} = \begin{bmatrix}-0.315 & 0.231 & 0.315 & -0.079 & 0.088\\ -1.161 & 0.627 & 1.741 & 1.635 & -0.219\\ 0.893 & 1.611 & 0.327 & -0.245 & -0.596\\ -0.804 & -0.116 & -0.663 & 0.859 & 2.122\\ 1.730 & -1.474 & -1.510 & -1.373 & -0.938\\ -0.343 & -0.880 & -0.209 & -0.797 & -0.456\end{bmatrix}.$$

Ainsi, la valeur recherchée est

$$Y_{26} = (0.2726)(-0.343) + \ldots + (-0.7035)(-0.456) \approx -0.2.$$

(b) 3.033

(c) Comme on l'a fait à la sous-question (a), on cherche le vecteur propre associé à la plus grande des valeurs propres (en fait, on aura juste besoin de la quatrième valeur du vecteur)

$$\mathbf{a}_1 \approx \begin{bmatrix} 0.4975536 & -0.3808431 & -0.4747977 & -0.5482675 & -0.2852523 \end{bmatrix}.$$

On sait ensuite que la contribution de la variable i à la composante principale j est

$$\begin{aligned} C(i,j) &= \frac{\mathbf{t}_{ij}^2}{1} \\ &= (-0.5482675)^2 \approx 0.3. \end{aligned}$$

11. Soit $\mathbf{u}$ un vecteur propre de la matrice $\mathbf{B}$. On a alors

$$\mathbf{Bu} = \lambda \mathbf{u}$$

pour $\lambda \in \mathbb{R}$. En multipliant par $\mathbf{X}^T$ de chaque côté du signe d'égalité, on obtient

$$\begin{aligned} \mathbf{X}^T\mathbf{Bu} &= \mathbf{X}^T\lambda\mathbf{u} \\ \mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{u} &= \lambda\mathbf{X}^T\mathbf{u} \\ \mathbf{A}\left(\mathbf{X}^T\mathbf{u}\right) &= \lambda\left(\mathbf{X}^T\mathbf{u}\right). \end{aligned}$$

Ainsi, $\mathbf{v} = \mathbf{X}^T\mathbf{u}/||\mathbf{X}^T\mathbf{u}||_2$ est un vecteur propre de la matrice $\mathbf{A}$. On divise par la norme afin de s'assurer que la norme de $\mathbf{v}$ est égale à 1.

12. (a) *Résolution 1 :* la corrélation ρ est égale au cosinus de l'angle θ de la variable avec la première composante principale. Comme la corrélation est bien positive (à voir dans le cercle de corrélations), on a $\rho = \sqrt{0.6698} = 0.8184$. *Résolution 2 :* On a $\rho = \sqrt{\lambda_1}a_{11}$ et $\lambda_1 = 0.4337 \times 6 = 2.6022$ et

$$\begin{aligned} a_{11} &= \pm\sqrt{1 - (-0.0075)^2 - (0.4827)^2 - (0.4216)^2 - (0.5452)^2 - (0.1858)^2} \\ &= \pm\sqrt{0.2574332}. \end{aligned}$$

On voit sur le cercle que $\rho > 0$, et donc $\rho = \sqrt{2.6022 \times 0.2574332} = 0.8184697$.

(b) Premier axe : propriétés physiques des cerveaux. Deuxième axe : intelligence. Troisième : poids du corps. Justifications : corrélations des variables avec les axes et la qualité de représentation exprimée par les cosinus carré des variables.

(c) Les jumeaux (3,4) : bien représentés dans le premier plan, tailles des cerveaux et intelligences en-dessous de la moyenne. Les jumeaux (7,8) : ,mal représentés dans le premier plan factoriel, bien représentés par le troisième axe, poids au-dessous de la moyenne.