Prof. R Naciri ✉ *dr.r.naciri@gmail.com*

Travaux digérés : Analyse des données
Semestre 5 : EE
Année universitaire : 2020-2021

Exercice 1:

On considère une matrice de données $\mathcal{A} = (x_{ij})$ de dimensions $r \times s$. Rappelons que x_{ij} est l'observation du i-ème individu à la j-ème variable. Indiquer la ou les réponses exactes.

1. La matrice de corrélation $\mathcal{R}$ s'exprime par :
   ◯ $\mathcal{R} = {}^t\mathcal{A}\mathcal{A}$ ◯ $\mathcal{R} = (\frac{1}{s})\ {}^t\mathcal{A}\mathcal{A}$ ◯ $\mathcal{R} = \mathcal{A}\mathcal{A}$
   ◯ $\mathcal{R} = (\frac{1}{r})\ {}^t\mathcal{A}\mathcal{A}$ ◯ $\mathcal{R} = (\frac{1}{r})\mathcal{A}\ {}^t\mathcal{A}$
2. Calculer les composantes principales à partir de la matrice X revient à :
   ◯ Diagonaliser la matrice de corrélation associée à $\mathcal{A}$.
   ◯ Diagonaliser la matrice de corrélation associée à $\mathcal{A}$.
   ◯ Diagonaliser la matrice $(\frac{1}{n})\mathcal{A}$.
3. La somme des variances de toutes les composantes principales est égale à :
   ◯ $r \times s$. ◯ s. ◯ r. ◯ $\frac{r}{s}$.
4. La première composante principale est :
   ◯ Une combinaison linéaire des variables observées ayant une variance maximale.
   ◯ La variable ayant la plus grande variance
   ◯ Une combinaison linéaire des lignes de la matrice $\mathcal{A}$
   ◯ Une combinaison linéaire des variables observées.
5. La variance de la première composante principale :
   ◯ Est toujours ≥ 1. ◯ Peut être < 1.
6. La qualité de la représentation des variables sur le plan principal 1-2 se mesure en fonction de la proximité du vecteur par rapport : ◯ A la première composante principale.
   ◯ au cercle de corrélation.
   ◯ a l'origine du repère.

Si $\mathcal{R}$ désigne la matrice de corrélation de cette analyse on note par λ_i sa i-éme grande valeur propre.

1. ACP est l'acronyme (abréviation) de :
2. Donner l'expression de la matrice de corrélation $\mathcal{R}$ en fonction de $\mathcal{A}$. Quelle est la dimension de $\mathcal{R}$?
3. Quel le nombre d'axes principaux obtenus à partir de cette ACP.
4. Donner les propriétés du premier axe principale.
5. Quel la valeur de la moyenne et la variance de la première composante principale ?
6. Quelle est la somme des contributions de tout les individus par apport à la première composante principale ?
7. Quelle est la somme des carrés des facteurs de corrélation de tout les variables par apport à la première composante principale ?

Exercice 2:

On a effectué l'ACP de la matrice de corrélations d'un jeu de 10 données avec 3 variables (var_1, var_2, var_3). Les valeurs est les vecteurs propres (normés) de la matrice de corrélation sont :

$$u_1 = \begin{pmatrix} -0.70 \\ -0.15 \\ \mathbf{x} \end{pmatrix} \quad u_2 = \begin{pmatrix} 0.05 \\ -0.98 \\ 0.17 \end{pmatrix} \quad u_3 = \begin{pmatrix} -0.71 \\ 0.08 \\ 0.70 \end{pmatrix}$$

Les composante principale de quelques observations sont données dans le tableau suivant :

	comp_1	comp_2	comp_2
Obs_1	1.48	-1.92	-0.08
Obs_2	0.02	0.44	1.38
Obs_3	-0.04	-0.85	0.02
⋮	⋮	⋮	⋮
Obs_{10}	-0.14	0.73	-0.56

Table 1 – *Composantes principales*

1. Quel est l'enertie totale du nuage.

2. Donner la valeur de x.

3. Compléter le tableau suivant en donnant les relation utilisées.

Valeur propre (λ_i)	Inertie expliqué	Inertie expliqué cumuler
-	-	0,56
-	-	-
0,33	-	-

TABLE 2 – *Inerties expliquées*

4. Calculer la qualité de représentation de la première observation par apport au deuxième axe principale, sachant que sa qualité de représentation par apport au premier axe principale est égale à 0.37.

5. Après avoir donné la formule, calculer la contribution de la première observation pour la deuxième composante principale.

6. (a) Donner l'expression de $comp_1$ en fonction de var_1, var_2 et var_3.

   (b) Monter qu'il existe une forte corrélation entre la var_1 et la première composante principale.

Exercice 3:

partir d'une ACP normé (données centrées réduites) On souhaite étudier la relation entre un ensemble de critères contribuant à la qualité de l'air dans 30 région différentes d'un pays. Les variables prises en considération sont :

- **airq** : Un indicateur de la qualité de l'air (des valeurs faibles de cet indicateur indiquent une bonne qualité de l'air).
- **vala** : La valeur ajoutée des compagnies de la région (exprimée en milliers de dollars)
- **rain** : Quantité de pluie (exprimée en inch)
- **dens** : Densité de la population
- **medi** : Revenu moyen par habitant.

	airq	vala	rain	dens	medi
région$_1$	104	2734.4	12.63	1815.86	4397

TABLE 3 – *Données concernant la première région*

comp$_1$	comp$_2$	comp$_3$	comp$_4$	comp$_5$
2.116	**x**	0.983	0.753	0.104

TABLE 4 – *Variances des composantes principales*

Variables	comp$_1$	comp$_2$
airq	-0.481	-0.492
vala	**y**	-0.009
rain	0.207	**z**
dens	-0.277	0.832
medi	-0.929	-0.09

TABLE 5 – *Corrélations entre les variables avec comp$_1$ et comp$_2$*

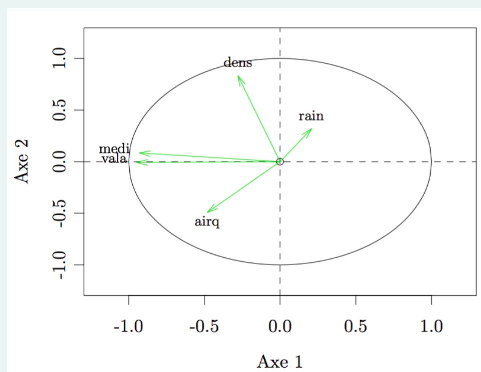

FIGURE 1 – *Cercle de corrélation*

1. Quel est l'inertie totale du nuage ?

2. Calculer la valeur de x dans le tableau 2.

3. Quel est le signe de y et z dans le tableau 3.

4. Calculer la valeur de y et z.

5. Calculer les cordonnées du vecteur propre associé à la première composante principale.

6. Quelle est la valeur de la cordonnée de la région$_1$ sur le premier axe principale.

7. Calculer la qualité de représentation de la région$_1$ par apport au premier axe principale. Cette région est-elle bien représentée ?

8. Après avoir donné la formule, calculer la contribution de la région$_1$ pour la première composante principale.