ACT6100	**Analyse de données**
H2019	**Série 1**

Rappels d'algèbre linéaire

1. On définit les matrices

$$\mathbf{A} = \begin{bmatrix} 3 & 0 \\ -1 & 2 \\ 1 & 1 \end{bmatrix}$$

et

$$\mathbf{B} = \begin{bmatrix} 1 & 4 & 2 \\ 3 & 1 & 5 \end{bmatrix}.$$

Évaluer $\mathbf{AB}$ à la main et vérifier le résultat à l'aide de R.

2. On définit les matrices

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

et

$$\mathbf{B} = \begin{bmatrix} 7 & 0 \\ 3 & 1 \end{bmatrix}.$$

Évaluer $\mathbf{A}^T + \mathbf{B}$ à la main et vérifier le résultat à l'aide de R.

3. On définit la matrice

$$\mathbf{Z} = \begin{bmatrix} 1 & 2 & 5 \\ 3 & 9 & 6 \\ 1 & 2 & 9 \end{bmatrix}.$$

Calculer à la main et à l'aide de R la trace de cette matrice.

4. On définit la matrice

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}.$$

Déterminer à la main et à l'aide de R les valeurs propres de cette matrice.

5. Quelle est la somme des valeurs propres de la matrice

$$\mathbf{H} = \begin{bmatrix} 2 & 1 & 7 & -2 \\ 0 & 3 & 2 & -5 \\ -1 & -1 & 0 & 8 \\ 2 & 2 & -1 & 11 \end{bmatrix}?$$

6. Est-ce que l'ensemble de vecteurs suivant dans un espace en trois dimensions est linéairement dépendants : $(-3,0,4)$, $(5,-1,2)$ et $(1,1,3)$?

Analyse en composantes principales

1. Soit $\mathbf{X}$, un vecteur aléatoire de longueur p tel que $E[\mathbf{X}] = \boldsymbol{\mu}$ et $\text{Var}[\mathbf{X}] = \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, où $\boldsymbol{\Lambda}$ est une matrice diagonale. Après une analyse en composantes principales, les nouvelles variables sont données par l'expression $\mathbf{Y} = \boldsymbol{\Gamma}^T(\mathbf{X} - \boldsymbol{\mu})$. Déterminer l'espérance et la variance de $\mathbf{Y}$. Interpréter brièvement les résultats.

2. La matrice de variance-covariance d'une base de données est

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \tau \\ \tau & 1 \end{bmatrix},$$

où $0 < \tau < 1$.

(a) Déterminer les valeurs propres de cette matrice.

(b) On considère maintenant un changement d'échelle dans la variable X_1 (par exemple, un changement d'unité), c'est-à-dire que X_1 devient cX_1 avec $c > 1$. En quoi la décomposition spectrale sera-t-elle affectée ? MAT8594

3. Démontrer le résultat de la Proposition 1 des notes de cours. MAT8594

4. Une analyse en composantes principales est réalisée à partir d'une base de données contenant huit variables. La matrice de corrélation est

$$\mathbf{R} = \begin{bmatrix} 1.00 & -0.20 & -0.60 & -0.88 & 0.71 & 0.74 & 0.88 & 0.81 \\ -0.20 & 1.00 & 0.61 & 0.29 & 0.33 & 0.36 & -0.20 & 0.25 \\ -0.60 & 0.61 & 1.00 & 0.76 & -0.27 & -0.27 & -0.45 & -0.38 \\ -0.88 & 0.29 & 0.76 & 1.00 & -0.70 & -0.68 & -0.84 & -0.81 \\ 0.71 & 0.33 & -0.27 & -0.70 & 1.00 & 0.92 & 0.66 & 0.90 \\ 0.74 & 0.36 & -0.27 & -0.68 & 0.92 & 1.00 & 0.68 & 0.93 \\ 0.88 & -0.20 & -0.45 & -0.84 & 0.66 & 0.68 & 1.00 & 0.80 \\ 0.81 & 0.25 & -0.38 & -0.81 & 0.90 & 0.93 & 0.80 & 1.00 \end{bmatrix}.$$

Une décomposition spectrale de cette matrice permet d'obtenir le vecteur de valeurs propres $\boldsymbol{\lambda} = (1.93, 0.13, 0.07, 0.02, 5.28, 0.41, 0.12, K)$, où K est une valeur inconnue.

(a) Déterminer la valeur de K.

(b) Quelle proportion de la variabilité des données initiales sera expliquée par les trois premières composantes principales ?

(c) Soit les vecteurs propres

$$\mathbf{a}_i = (-0.117, 0.697, 0.491, 0.204, 0.284, 0.298, -0.091, 0.200)$$

et

$$\mathbf{a}_j = (0.406, -0.016, -0.256, -0.406, 0.377, 0.381, 0.386, 0.410).$$

Déterminer à quelles valeurs propres ils correspondent.

(d) Définir explicitement l'équation de la première composante principale.

5. On souhaite réaliser une analyse en composantes principales à partir d'une base de données contenant deux variables. La Table 1 présente les valeurs pour les dix individus de l'échantillon. La matrice de variance-covariance est donnée par

$$\mathbf{\Sigma} = \begin{bmatrix} 1.0000 & 0.4297 \\ 0.4297 & 0.5840 \end{bmatrix}.$$

(a) Tracer la carte initiale des individus.

(b) Déterminer la matrice de corrélation.

(c) Calculer les valeurs propres et les vecteurs propres de la matrice de corrélation.

(d) Quelle proportion de la variabilité initiale des données est expliquée par la première composante principale ?

(e) Si on considère qu'une seule composante sera conservée, quelle sera la coordonnée de l'individu 7 sur la carte finale des individus ?

#	X_1	X_2
1	-1.49	-0.61
2	-0.07	-1.13
3	0.57	-0.73
4	1.22	0.91
5	2.44	1.04
6	0.24	1.24
7	0.25	0.16
8	1.12	0.61
9	-0.35	0.19
10	0.94	-0.13

TABLE 1 – Individus

(f) Si on considère que deux composantes sont conservées, quelle est la corrélation entre la variable X_1 et la seconde composante principale ?

6. La base de données *Rank.csv* disponible sur le site *Moodle* du cours contient plusieurs variables élaborées afin de mesurer la qualité scientifique de 50 universités. En particulier, on s'intéresse aux 6 variables suivantes :
   — **Alumni** : score basé sur le nombre d'alumnis ayant obtenu un prix Nobel ou une médaille Field ;
   — **HiCi** : score basé sur le nombre de membres du personnel académique repris dans la liste des *highly cited researchers Sciences* entre 2002 et 2006 ;
   — **Award** : score basé sur le nombre de membres du personnel académique ayant obtenu un prix Nobel ou une médaille Field ;
   — **NS** : score basé sur le nombre d'articles publiés dans *Nature* et *Sciences* entre 2002 et 2006 ;
   — **SCI** : score basé sur le nombre d'articles indexés dans *Science Citation Index-expanded* et *Social Science Citation Index 2006* ; et
   — **Size** : moyenne pondérée des cinq mesures précédentes divisées par le nombre d'équivalents temps-plein du personnel académique de l'institution.

   Réaliser une analyse en composantes principales à partir des six variables disponibles et répondre brièvement aux questions suivantes à l'aide de R.

   (a) Combien de composantes principales doit-on conserver dans cette analyse ?

   (b) Dans cette analyse, quelle proportion de la variabilité totale des données est expliquée par les trois premières composantes ?

   (c) Dans la première composante, quelle variable est la moins bien représentée ?

   (d) L'analyse a permis de passer d'un espace $\mathbb{R}^6$ à un espace $\mathbb{R}^2$. Quelle proportion de l'information a été perdue à la suite de cette procédure ?

   (e) Quelle proportion de la variance initiale de la variable **Award** est expliquée par les deux premières composantes ?

   (f) À partir du cercle des corrélations, quelle variable est la moins bien représentée par les deux premières composantes ?

7. La base de données *carsACP.csv* disponible sur le site *Moodle* du cours contient les résultats d'une étude faite pour différentes marques de voitures. Les variables considérées sont
   — X_1 : économie ;
   — X_2 : service ;
   — X_3 : non dépréciation de la valeur de la voiture ;
   — X_4 : prix ;

— X_5 : design ;
— X_6 : caractère «sport» de la voiture ;
— X_7 : sécurité ; et
— X_8 : facilité de conduite.

Pour chacune des 24 marques considérées, un score est attribué à chacun des ces critères (1 étant le meilleur et 6 étant le moins bon). Le cercle des corrélations obtenu est présenté à la Figure 1 et la carte finale des individus à la Figure 2. Commenter.

8. Une étude a été réalisée afin de comparer les niveaux moyens des salaires pour 6 professions dans 24 villes du monde. Les professions considérées sont
— manoeuvre en bâtiment (**man**),
— chef de service (**chefserv**),
— ingénieur (**inge**),
— employé de banque (**banquier**),
— vendeur (**vendeur**) et
— ouvrier du textile (**ouvrier**).

La variable **indicateur** est une mesure de la qualité de vie globale. Les données sont disponibles dans la base de données *metiers.csv* sur le site *Moodle* du cours. Utiliser R pour répondre aux questions ci-dessous.

   (a) Quelle est la corrélation entre les variables **ouvrier** et **man** ?

   (b) Réaliser une analyse en composantes principales en utilisant toutes les variables et tous les individus disponibles dans la base de données **sauf la variable indicateur** (utiliser la même analyse pour les prochaine sous-question). Si on conserve deux composantes principales, quel pourcentage de l'information est perdu ?

   (c) Quelle variable est la mieux représentée sur le premier plan factoriel ?

   (d) Quelle observation est la moins bien représentée dans le premier plan factoriel ?

9. On considère deux variables aléatoires indépendantes U_1 et U_2 avec distribution Uniforme sur l'intervalle $(0, 1)$. On définit $\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \end{bmatrix}^T$, où $X_1 = U_1$, $X_2 = U_2$, $X_3 = U_1 + U_2$ et $X_4 = U_1 - U_2$. MAT8594

   (a) Calculer la matrice des corrélations de $\mathbf{X}$.

   (b) On sait que que $\mathbf{v}_1 = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 1 & 0 \end{bmatrix}^T$ et $\mathbf{v}_2 = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & 1 \end{bmatrix}^T$ sont les vecteurs propres correspondant à λ_1 et λ_2. Quelle pourcentage de l'information se retrouvera dans chacune des 4 dimensions ?

10. Une ACP a été effectuée sur la matrice de corrélation des variables $X_1, \ldots, X_5$ avec 6 individus. Les données et les deux premières composantes principales sont présentées dans la Table 2 (à gauche), ainsi que les corrélations entre les variables X_j et les deux premières composantes (à droite). La carte des individus et le cercle de corrélation sont donnés dans la Figure 3.

	X_1	X_2	X_3	X_4	X_5	Y_1	Y_2
i_1	0.56	1.19	1.76	-1.58	1.80	0.37	0.11
i_2	-1.22	1.75	4.05	1.93	1.45	2.48	0.76
i_3	3.10	3.14	1.78	-1.92	1.02	0.02	1.58
i_4	-0.47	0.70	0.19	0.34	4.12	1.12	-2.15
i_5	4.86	-1.22	-1.17	-4.23	0.63	-3.16	-0.10
i_6	0.50	-0.38	0.92	-3.05	1.18	-0.83	**?**

	Y_1	Y_2
X_1	-0.87	0.31
X_2	0.66	0.53
X_3	0.83	0.51
X_4	0.95	-0.11
X_5	0.50	-0.80

TABLE 2 – *A gauche :* les variables X_j et les composantes principales Y_k. *A droite :* les corrélations entre les variables X_j et les composantes principales Y_k.

(a) Une valeur de l'individu i_6 pour la composante Y_2 est manquante dans la Table 2 (à gauche). Calculer cette valeur.

(b) Quelle est la valeur principale la plus grande de la matrice de corrélation des variables intiales ($\mathbf{X}$) ?

(c) Commenter la phrase suivante : « Plus de 90% de la première composante Y_1 est déterminée par la variable X_4. »

11. On considère une matrice $\mathbf{X}$ de taille $(m \times d)$ avec $d >> m$ et on définit les matrices $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ et $\mathbf{B} = \mathbf{X}\mathbf{X}^T$. Si $\mathbf{u}$ est un vecteur propre de $\mathbf{B}$, démontrer que $\mathbf{X}^T\mathbf{u}/||\mathbf{X}^T\mathbf{u}||_2$ est un vecteur propre de la matrice $\mathbf{A}$. Ce résultat est à la base de la méthode accélérée utilisée dans l'Application 2 des notes de cours. MAT8594

12. La base de données *brains* contient des observations pour 20 individus sur, entre autres, les 6 variables suivantes :

CCMIDSA superficie du corps calleux [1] (cm^2)

FIQ quotient intellectuel

HC circonférence de la tête (cm)

TOTSA superficie totale (cm^2)

TOTVOL volume total du cerveau (cm^3)

WEIGHT poids du corps (kg).

Les résultats d'une analyse en composantes principales sur la matrice de corrélation sont présentés à la Figure 4 et à la Figure 5.

```
cos2 des variables
          Dim.1         Dim.2       Dim.3
 CCMIDSA 0.6697679994  0.06946701  0.053775764
 FIQ     0.0001476554  0.81340936  0.018223004
 HC      0.6063027377  0.08140132  0.001671312
 TOTSA   0.4624595541  0.27698830  0.013378111
 TOTVOL  0.7734577797  0.01352115  0.002787014
 WEIGHT  0.0898547131  0.01231922  0.881231817

 cos2 des individus 3, 4, 7 et 8
    Dim.1       Dim.2       Dim.3
3  0.54988433  0.37126579  0.01924192
4  0.65602320  0.29425953  0.02856360
7  0.02577022  0.06537907  0.78652437
8  0.04694529  0.00537898  0.84458992
```

FIGURE 1 – Résultats de l'analyse en composantes principales.

(a) Le vecteur propre associé à la valeur propre maximale de la matrice de corrélation des six variables est $(\ldots, -0.0075, 0.4827, 0.4216, 0.5452, 0.1858)$. Calculer à trois décimales près la corrélation de la variable CCMIDSA avec la première composante principale.

(b) Donner une interprétation des trois premières composantes principales. Justifier la réponse à partir des résultats de l'analyse.

(c) Les vingt individus sont en fait des jumeaux monozygotes : (1,2), (3,4), etc. Contraster les positions des jumeaux (3,4) d'une part les jumeaux (7,8) d'autre part.

1. Le corps calleux du cerveau assure le transfert d'informations entre les deux hémisphères et leur coordination.

Réponses Rappels d'algèbre linéaire

1.

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

2.

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

3. 19
4. 2 et 3
5. 16
6. Les vecteurs sont linéairement indépendants.

Analyse en composantes principales

1. $E[\mathbf{Y}] = \mathbf{0}_p$ et $\text{Var}[\mathbf{Y}] = \mathbf{\Lambda}$.
2. (a) $1 \pm \tau$
   (b) -
3. Démonstration.
4. (a) 0.04
   (b) 0.9525
   (c) 1.93 pour $\mathbf{a}_i$ et 5.28 pour $\mathbf{a}_j$
   (d) $y^1 = (0.406)X_1 + (-0.016)X_2 + \ldots + (0.410)X_8$
5. (a) La carte initiale des individus est présentée à la Figure 6.

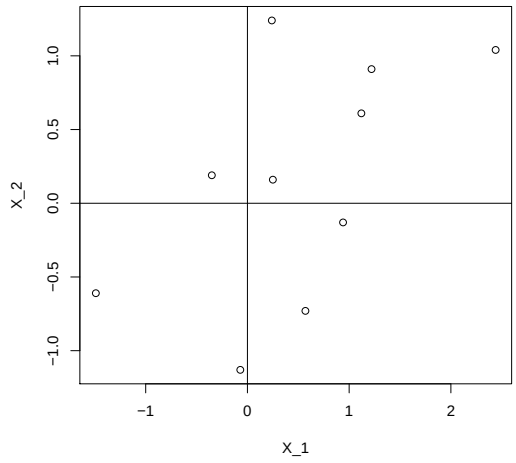

FIGURE 2 – Carte initiale des individus.

(b)

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

(c) $\lambda_1 = 1.5622757$, $\lambda_2 = 0.4377243$, $a_{11} = 0.7071$, $a_{12} = 0.7071$, $a_{21} = -0.7071$ et $a_{22} = 0.7071$ [2]

(d) 0.78114

(e) $(-0.163, 0.172)$

(f) 0.4678226

6. (a) 2

(b) 0.9163

(c) **SCI**

(d) ≈ 0.16

(e) ≈ 0.90

(f) **Alumni**

7. Il est important de noter dans votre analyse que dans cet exercice, 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) 0.87

(b) ≈ 0.0813

(c) **ouvrier**

(d) **Hong Kong**

9. (a)

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

(b) Respectivement 0.5, 0.5, 0.0 et 0.0

10. (a) -0.2

(b) 3.033

(c) L'affirmation est fausse. La variable X_4 compte pour environ 30% de la première composante (contribution).

11. Démonstration.

1. 0.8184
2. -
3. -

2. 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.