Correction des exercices sur l'ACP

Exercice 1. Une analyse en composante principale (ACP normée) a été effectuée sur 50 avions. On a détermine, pour chacun d'eux, la valeur de 10 variables (vitesse de croisière, rayon d'action, consommation, nombre de places, coût de revient du transport par passager et par kilomètre, etc).

On considère la représentation de ces variables dans le cercle de corrélation ci-dessous.

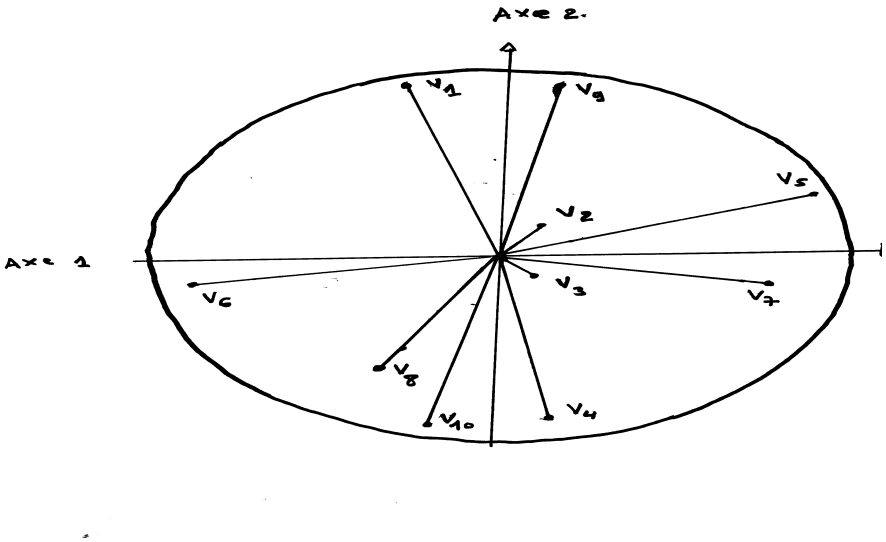

FIGURE 1 –

1) Quelles sont les variables qui peuvent aider à donner une signification à l'axe 1 ?

2) Quelles sont les variables qui ne dvoient pas être interprétées sur cette figure ?

3) Donner 3 groupes de variables qui, au sein d'un même groupe, sont fortement corrélées positivement entre elles.

4) Citer deux variables qui sont peu corrélées entre elles.

5) Citer deux variables qui sont fortement corrélées négativement avec la variable V_4.

6) Quel est approximativement le coefficient de corrélation entre la variable V_1 et la première composante principale ?

7) Citer une variable dont le coefficient de corrélation avec la deuxième composante principale vaut presque 1.

8) Que veut le coefficient de corrélation entre la première et la deuxième composante principale.

Correction de l'exercice 1.

1) Ce sont les variables représentées par des points proches du cercle des corrélations et proches de l'axe 1. Ici : V_5, V_6 et V_7

2) Ce sont les variables représentées par des points trop éloignés du cercle des corrélations (proches de 0). Ici : V_2 et V_3.

3) Les variables représentées par des points proches du cercle des corrélations et proches entre elles sont fortement corrélées positivement.

On distingue 3 groupes :

groupe 1 : V_5 et V_7 ;
groupe 2 : V_1 et V_9 ;
groupe 3 : V_4, V_8 et V_{10}.

4) Deux variables représentées par des points proches du cercle des corrélations et formant avec 0 un angle droit (ou presque droit) ne sont pas corrélées entre elles (ou sont peu corrélées entre elles).
On peut citer ici :
V_7 et V_8, V_7 et V_{10} ; V_7 et V_4 ; V_7 et V_9 ; V_5 et V_9 ; V_6 et V_1 ; V_6 et V_{10}, etc.

5) Deux variables représentées par des points proches du cercle des corrélations et formant avec 0 un angle plat (ou presque plat) sont fortement corrélées négativement entre elles.
on observe ici que les variables fortement corrélées négativement avec V_4 sont V_1 et V_9.

6) Ce coefficient est égal à l'abscisse (coordonnée sur l'axe 1) du point représentant V_1 . Il vaut environ $-0,33$.

7) La variable V_9 convient puisqu'elle est représentée par un point dont la coordonnée sur l'axe 2 vaut presque 1.

8) On sait que les composantes principales sont toutes non corrélées deux à deux, le coefficient de corrélation entre la première et la deuxième est donc nul.

Exercice 2. A partir des graphiques de la figure 2 répondre aux questions suivantes :
1. Que peut-on penser des valeurs prises par l'individu 3 pour les variables B et D ? Même question pour les valeurs prises par l'individu 2.
2. Que peut-on dire des corrélations entre les variables A et B, C et D, A et E ?
3. Vrai, Faux ou ne sait pas ?
- L'individu 1 prend des valeurs élevées pour la variable C.
- Pour chacune des variables, les individus 1 et 2 ont des valeurs similaires.

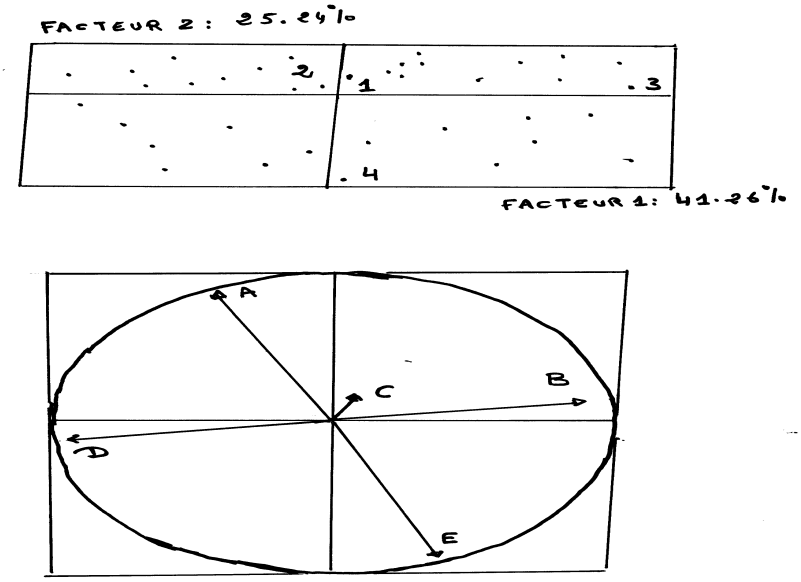

FIGURE 2 –

Correction de l'exercice 2.

- 1.
  L'individu 3 a vraisemblablement des valeurs fortes pour la variable B et des valeurs faibles pour la variable D.
  L'individu 2 a vraisemblablement des valeurs proches de la moyenne de l'ensemble des individus pour les variables B et D.
  Comme toute interprétation d'un point particulier, ces deux interprétations doivent être contrôlées sur les données.
- 2.
  Les variables A et B forment un angle proche de $\pi/2$ dans le plan ; comme elles sont bien projetées, l'angle dans l'espace est proche de $\pi/2$, son cosinus est donc proche de 0 : ces deux variables ne sont pas corrélées ;
  La variable D est dans le plan de projection tandis que la variable C est orthogonale à ce plan. Ces deux variables sont proches de l'orthogonalité : leur corrélation est proche de 0.
  Les variables A et E sont fortement corrélées, négativement.
- 3
  On ne peut pas savoir si l'individu 1 prend des valeurs élevées ou non pour la variable C car cette variable est très mal projetée.
  Pour les variables A, B, D et E, les individus 1 et 2 ont vraisemblablement des valeurs proches de la moyenne de l'ensemble des individus mais pour la variable C, on ne peut rien dire (voir question précédente).

Exercice 3. Vrai ou Faux.

On considère une ACP normée dans laquelle le poids des individus est le même. Répondre par vrai ou faux en justifiant la réponse.

1. Plus les variables sont corrélées entre elles plus le pourcentage d'inertie porté par les premiers axes de l'ACP est grand.

2. Dans l'espace des individus (espace $\mathbb{R}^p$), les individus éloignés du centre de gravité du nuage jouent un rôle important dans l'analyse.

3. La variance des coordonnées des individus sur le premier axe factoriel est plus élevée que la variance des coordonnées sur le second axe.

4. Des variables superposées sur le graphe des corrélations sont nécessairement très corrélées.

5. Dans $\mathbb{R}^p$, un individu très proche du centre de gravité a des valeurs brutes proches de zéro pour l'ensemble des variables.

Corrigé de l'exercice 3

- 1. Vrai.
  Plus les variables sont corrélées entre elles, plus il est facile de les résumer par un petit nombre de variables synthétiques (les composantes principales) et donc plus le pourcentage d'inertie porté par les premières composantes principales est important. Ce pourcentage est la somme des carrés des coefficients de corrélation entre la composante et les variables initiales.
- 2. Vrai.
  Les individus ayant les mêmes poids, les individus très éloignés du centre de gravité du nuage contribuent à une part importante de la variabilité (=de l'inertie). Ils "attirent" donc les axes puisque ces derniers ont pour propriété de représenter au mieux l'inertie du nuage.
- 3. Vrai.
  La variance des coordonnées correspond à l'inertie, et les axes sont classés par inertie décroissante ; la variance des coordonnées sur le premier axe factoriel est donc plus élevée que la variance des coordonnées sur le second axe.

– 4. Faux.
Il faut que les deux variables soient superposées mais aussi qu'elles soient bien projetées (pointes des flèches proches du cercle des corrélations) pour qu'en puisse en déduire qu'elles sont corrélées entre elles (notons que deux variables peuvent être mal projetées et étroitement corrélées).
– 5. Faux.
Un individu très proche du centre de gravité a des valeurs proches de la moyenne pour chacune des variables.

Exercice 4.

Considérons les notes (de 0 à 20) obtenues par 9 élèves dans 4 disciplines (mathématiques, physique, français, anglais) :

	MATH	PHYS	FRAN	ANGL
jean	6.00	6.00	5.00	5.50
alan	8.00	8.00	8.00	8.00
anni	6.00	7.00	11.00	9.50
moni	14.50	14.50	15.50	15.00
didi	14.00	14.00	12.00	12.50
andr	11.00	10.00	5.50	7.00
pier	5.50	7.00	14.00	11.50
brig	13.00	12.50	8.50	9.50
evel	9.00	9.50	12.50	12.00

Nous présentons ci-dessous quelques résultats de l'A.C.P.

1. Résultats préliminaires

Le logiciel fournit tout d'abord la moyenne (mean), l'écart-type (standard deviation), le minimum et le maximum de chaque variable. Il s'agit donc, pour l'instant, d'études univariées.

Statistiques élémentaires

Variable	Moyenne	Ecart-type	Minimum	Maximum
MATH	9.67	3.37	5.50	14.50
PHYS	9.83	2.99	6.00	14.50
FRAN	10.22	3.47	5.00	15.50
ANGL	10.06	2.81	5.50	15.00

1) Que remarquez-vous ?
Réponse. Grande homogénéité des 4 variables : même ordre de grandeur pour la moyenne, les écarts-types, les minima et les maxima

Le tableau suivant donne la matrice des corrélations. Il donne les coefficients de corrélation linéaire des variables prises deux à deux.

Coefficients de correlation

	MATH	PHYS	FRAN	ANGL
MATH	1.00	0.98	0.23	0.51
PHYS	0.98	1.00	0.40	0.65
FRAN	0.23	0.40	1.00	0.95
ANGL	0.51	0.65	0.95	1.00

2) Que remarquez-vous ?
Réponse. Toutes les corrélations linéaires sont positives, ce qui signifie que toutes les variables varient (en moyenne) dans le même sens

2. Résultats généraux

Matrice des variances-covariances

	MATH	PHYS	FRAN	ANGL
MATH	11.39	9.92	2.66	4.82
PHYS	9.92	8.94	4.12	5.48
FRAN	2.66	4.12	12.06	9.29
ANGL	4.82	5.48	9.29	7.91

Valeurs propres ; variances expliquées

FACTEUR	VAL. PR.	PCT. VAR.	PCT. CUM.
1	28.23	0.70	0.70
2	12.03	0.30	1.00
3	0.03	0.00	1.00
4	0.01	0.00	1.00
	− − −	− − −	
	40.30	1.00	

Ici :
PCT=pourcentage de variance
PCT= pourcentage cumulé : exemple $(28,23/40.30) \times 100 = 70\%$.
Facteur i (ou composante principale C_i).

3) Quelle est la relation entre λ_i est la variance de C_i ?
Réponse :

$$Var(C_i) = \lambda_i, \quad \text{où} \quad \lambda_i \quad \text{Val.Pr de la Matrice Variance-Covariance.}$$

4) Comment interpréter la relation suivante qui relié la variance des variables initiales X_i avec celle des composantes principales C_i ?

$$\sum_{i=1}^{4} Var(X_i) = \sum_{i=1}^{4} Var(C_i).$$

Réponse :
Le nuage de points en dimension 4 est toujours le même et sa dispersion globale n'a pas changée. C'est la répartition de cette dispersion, selon les nouvelles variables que sont les facteurs, ou composantes principales, qui se trouve modifiée :

3. Résultats sur les variables

Le résultat fondamental concernant les variables est le tableau des corrélations variables-facteurs (tableau des $r(X_j, C_k)$). Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les facteurs. Ce sont ces corrélations qui vont permettre de donner un sens aux facteurs (de les interpréter).

Corrélations variables-facteurs : $r(X_j, C_k)$

FACTEURS	F1	F2	F3	F4
MATH	0.81	−0.58	0.01	−0.02
PHYS	0.90	−0.43	−0.03	0.02
FRAN	0.75	0.66	−0.02	−0.01
ANGL	0.91	0.40	0.05	0.01

Les deux premières colonnes de ce tableau permettent, tout d'abord, de réaliser le graphique des variables donné par la Fig. 3

Mais, ces deux colonnes permettent également de donner une signification aux facteurs (donc aux axes des graphiques).

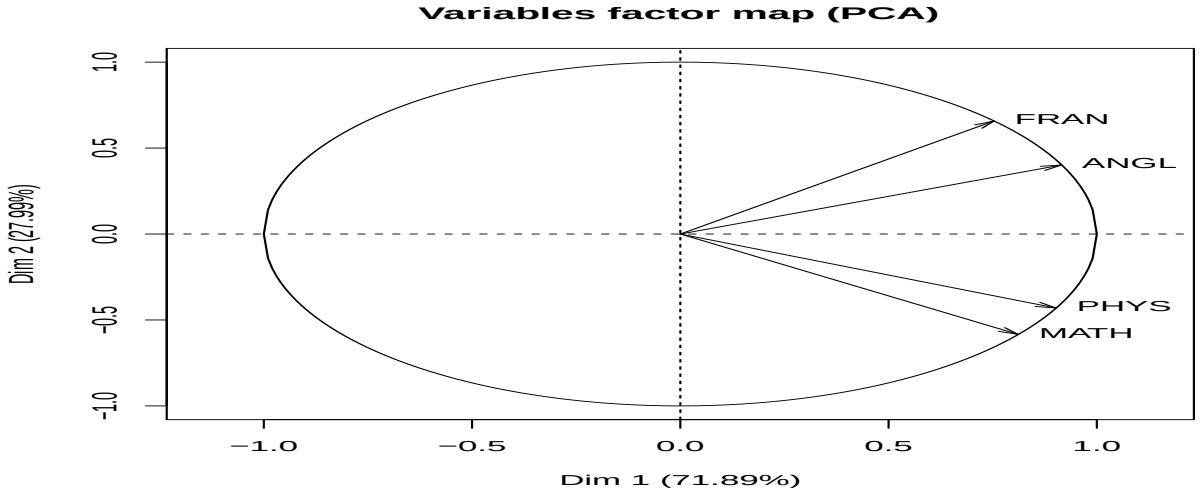

FIGURE 3 –

5) Comment interprétez-vous ces résultats ?

Réponse

On notera que les deux dernières colonnes ne seront pas utilisées puisqu'on ne retient que deux dimensions pour interpréter l'analyse

Interprétation.

On voit que le premier facteur est corrélé positivement, et assez fortement, avec chacune des 4 variables initiales : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score élevé sur l'axe 1 ; réciproquement, plus ses notes sont mauvaises, plus son score est négatif.

- L'axe 1 représente donc, en quelques sortes, le résultat global (dans l'ensemble des 4 disciplines considérées) des élèves.

- L'axe 2, il oppose, d'une part, le français et l'anglais (corrélations positives), d'autre part, les mathématiques et la physique (corrélations négatives). Il s'agit donc d'un axe d'opposition entre disciplines littéraires et disciplines scientifiques, surtout marqué par l'opposition entre le français et les mathématiques.

Cette interprétation peut être précisée avec graphiques et tableaux relatifs aux individus. Ce que nous donnons ci-dessous

4. Résultats sur les individus

Le tableau donné ci-dessous contient tous les résultats importants de l'A.C.P. sur les individus

Coordonnées des individus ; contributions ; cosinus carées

	POIDS	FACT1	FACT2	CONTG	CONT1	CONT2	COSCA1	COSCA2
jean	0.11	−8.61	−1.41	20.99	29.19	1.83	0.97	0.03
alan	0.11	−3.88	−0.50	4.22	5.92	0.23	0.98	0.02
anni	0.11	−3.21	3.47	6.17	4.06	11.11	0.46	0.54
moni	0.11	9.85	0.60	26.86	38.19	0.33	1.00	0.00
didi	0.11	6.41	−2.05	12.48	16.15	3.87	0.91	0.09
andr	0.11	−3.03	−4.92	9.22	3.62	22.37	0.28	0.72
pier	0.11	−1.03	6.38	11.51	0.41	37.56	0.03	0.97
brig	0.11	1.95	−4.20	5.93	1.50	16.29	0.18	0.82
evel	0.11	1.55	2.63	2.63	0.95	6.41	0.25	0.73

On notera que chaque individu représente 1 élément sur 9, d'où un poids (une pondération) de $1/9 = 0.11$*, ce qui est fourni par la première colonne du tableau. Les 2 colonnes suivantes fournissent les coordonnées des individus (les élèves) sur les deux premiers axes (les facteurs) et ont donc permis de réaliser le graphique des individus. Ce dernier (Fig. 4) permet de préciser la signification des axes, donc des facteurs.*

La signification et l'utilisation des dernières colonnes du tableau seront explicitées un peu plus loin.

6) Interpréter les résultats obtenu sur les individus.

Interprétation.

On confirme que : **l'axe 1 représente le résultat d'ensemble des élèves** :

- si on prend leur score - ou coordonnée- sur l'axe 1, on obtient le même classement que si on prend leur moyenne générale.
- L'élève "le plus haut" sur le graphique, celui qui a la coordonnée la plus élevée sur l'axe 2, est Pierre dont les résultats sont les plus contrastés en faveur des disciplines littéraires (14 et 11.5 contre 7 et 5.5). C'est exactement le contraire pour André qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disciplines littéraires (7 et 5.5).

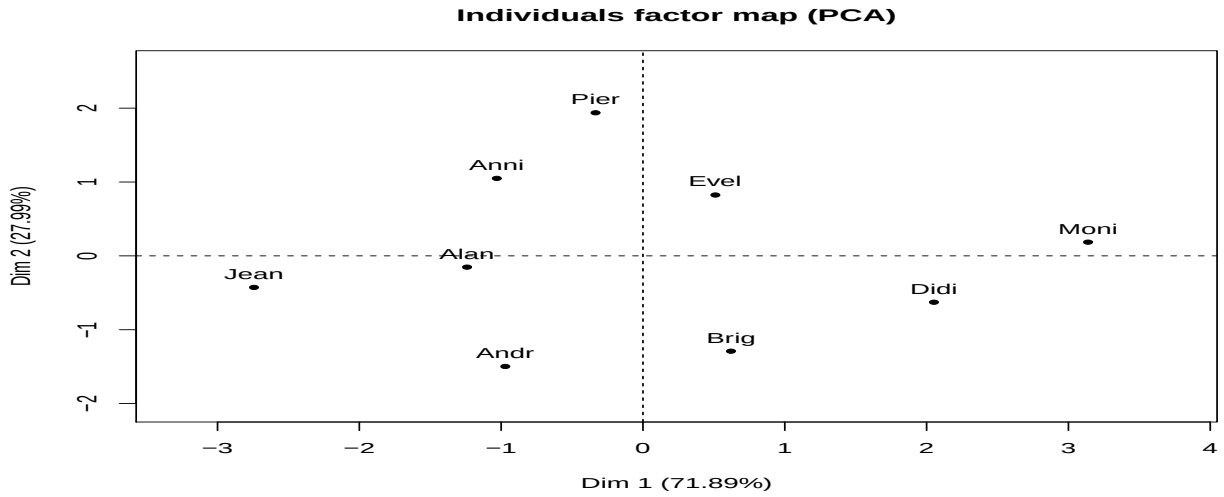

FIGURE 4 –

– Monique et Alain ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines (mais à des niveaux très distincts, ce qu'a déjà révélé l'axe 1).

L'axe 2 oppose bien les "littéraires" (en haut) aux "scientifiques" (en bas).

– Les 3 colonnes du tableau ci-dessus fournissent des contributions des individus à diverses dispersions :
– **cont1 et cont2** donnent les contributions (en pourcentages) des individus à la variance selon les axes 1 et 2 (rappelons que l'on utilise ici la variance pour mesurer la dispersion) ;
– **Contg** donne les contributions générales, c'est-à-dire à la dispersion en dimension 4 (il s'agit de ce que l'on appelle l'inertie du nuage des élèves.
Ces contributions sont fournies en pourcentages (chaque colonne somme à 100) et permettent de repérer les individus les plus importants au niveau de chaque axe. Elles servent en général à affiner l'interprétation des résultats de l'analyse.

Ainsi, par exemple, la variance de l'axe 1 vaut 28.23 (première valeur propre). On peut la retrouver en utilisant la formule de définition de la variance :

$$Var(C_1) = \frac{1}{9}\sum_{i=1}^{9}(c_{1i})^2$$

La coordonnée de Jean (le premier individu du fichier) sur l'axe 1 vaut $c_{11} = -8.61$; sa contribution est donc :

$$\frac{\frac{1}{9}(-8.61)^2}{28.23} \times 100 = 29.19\%$$

A lui seul, cet individu représente près de 30% de la variance : il est prépondérant (au même titre que Monique) dans la définition de l'axe 1 ; cela provient du fait qu'il a le résultat le plus faible, Monique ayant, à l'opposé, le résultat le meilleur.
– Les 2 dernières colonnes du tableau sont des **cosinus carrés** qui fournissent la qualité de la représentation de chaque individu sur chaque axe. Ces quantités s'additionnent axe par axe, de sorte que, en dimension 2, Evelyne est représentée à 98% $(0.25 + 0.73)$, tandis que les 8 autres individus le sont à 100%.