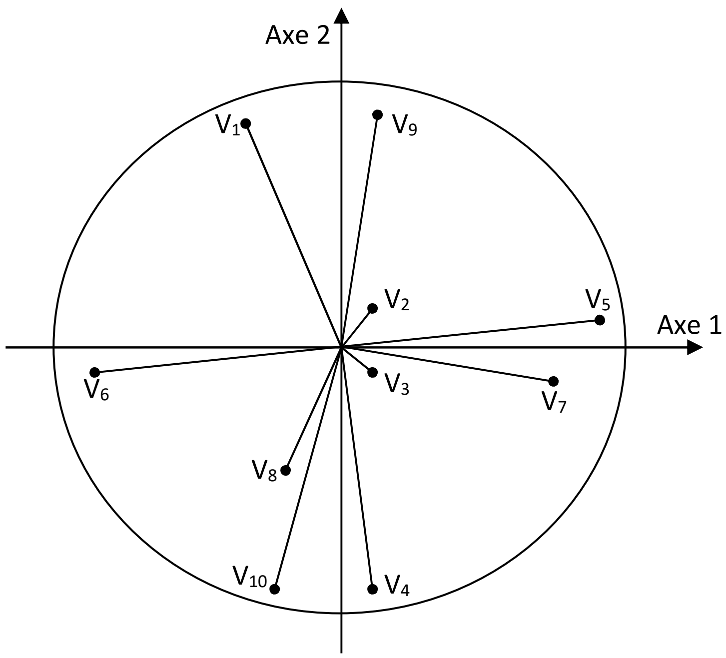
TP : Analyse en composantes principales : principes

Durée : 3h

Ce TD a pour objectif une meilleure compréhension des différents résultats fournis par l'analyse en composantes principales.

Exercice 1 — *Interprétation des axes*

Une analyse en composante principale (ACP normée) a été effectuée sur 50 avions.
On a déterminé, 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.

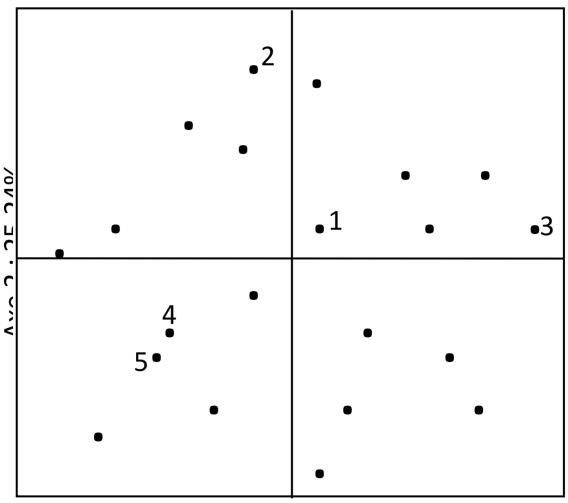


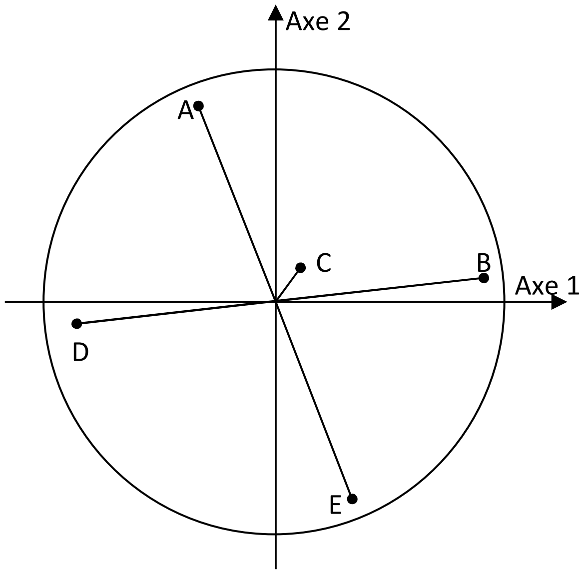
1) Quelles sont les variables qui peuvent aider à donner une signification à l'axe 1 ?
2) Quelles sont les variables qui ne doivent pas être interprétées sur cette figure ?
3) Donner 3 groupes de variables qui, au sein d'un même groupe, sont fortement corrélées positivement entre elles.
4) Citer deux 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 peut-on dire sur la corrélation entre les variables V_3 et V_7 ?
9) Que signifie le coefficient de corrélation entre la première et la deuxième composante principale ?

Exercice 2 — *Interprétation des individus*

A partir des graphiques répondre aux questions suivantes :

1) Que peut-on penser des valeurs prises par l'individu 3 pour les variables B et D?
2) Même question pour les valeurs prises par l'individu 2. Quelles variables permettent de caractériser l'individu 2 ?
3) Peut-on dire que l'individu 1 est proche du centre de gravité ?
4) Peut-on dire que les individus 4 et 5 ont des valeurs similaires pour chacune des variables ?





Exercice 3 — *Vrai - 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 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 R^p, un individu très proche du centre de gravité a des valeurs brutes proches de zéro pour l'ensemble des variables.

Exercice 4 *Etude complète*

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

	MATH	PHYS	FRAN	ANGL
Lucas	6.00	6.00	5.00	5.50
Maxime	8.00	8.00	8.00	8.00
Quentin	6.00	7.00	11.00	9.50
Louna	14.50	14.50	15.50	15.00
Camille	14.00	14.00	12.00	12.50
Pierre	11.00	10.00	5.50	7.00
Emmy	5.50	7.00	14.00	11.50
Remi	13.00	12.50	8.50	9.50
Marie	9.00	9.50	12.50	12.00

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

Résultats préliminaires

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

Coefficient de corrélation

	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

Que remarquez-vous ? Que pouvez-vous en conclure sur les axes de l'ACP ?

Résultats sur l'inertie

Matrice de variance-covariance

	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 et variance expliquée

COMP.	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)×100 = 70%.

a) Quelle est la relation entre λ_i est la variance de C_i ?
b) Comment interprétez-vous la relation suivante qui relie la variance des variables initiales X_i avec celle des composantes principales C_i ?

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

Résultats sur les variables

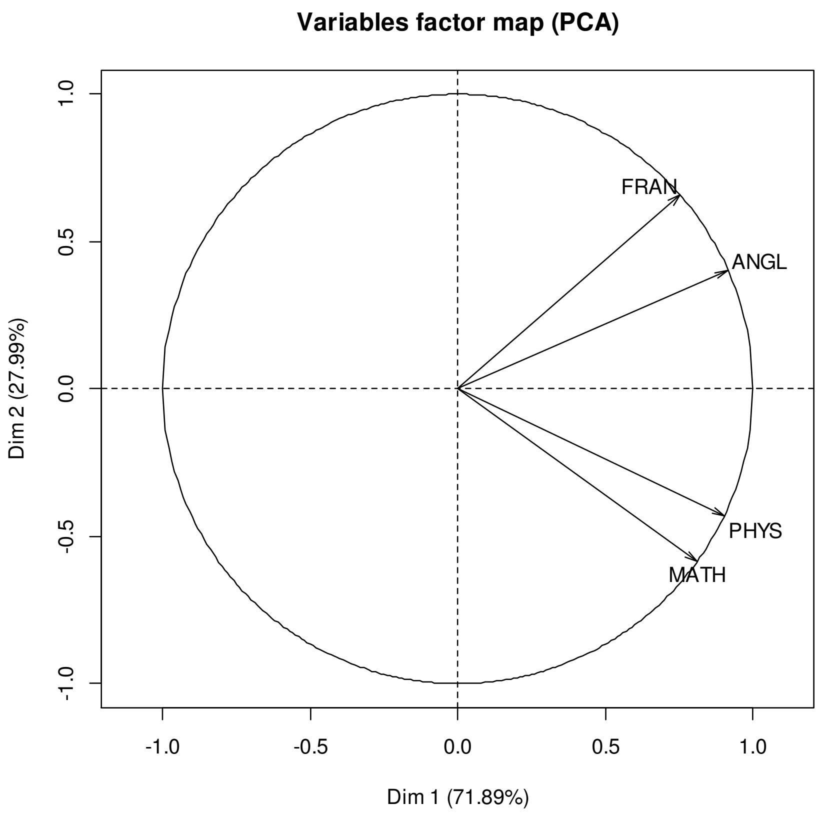
Le résultat fondamental concernant les variables est le tableau des corrélations variables-composantes (tableau des $r(X_j,C_k)$). Il s'agit des coefficients de corrélation linéaire entre les variables initiales et les composantes principales. Ce sont ces corrélations qui vont permettre de donner un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

Corrélations variables – composantes $r(X_j,C_k)$

	C1	C2	C3	C4
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 ci-dessous.

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



Comment interprétez-vous ces résultats ?

Résultats sur les individus

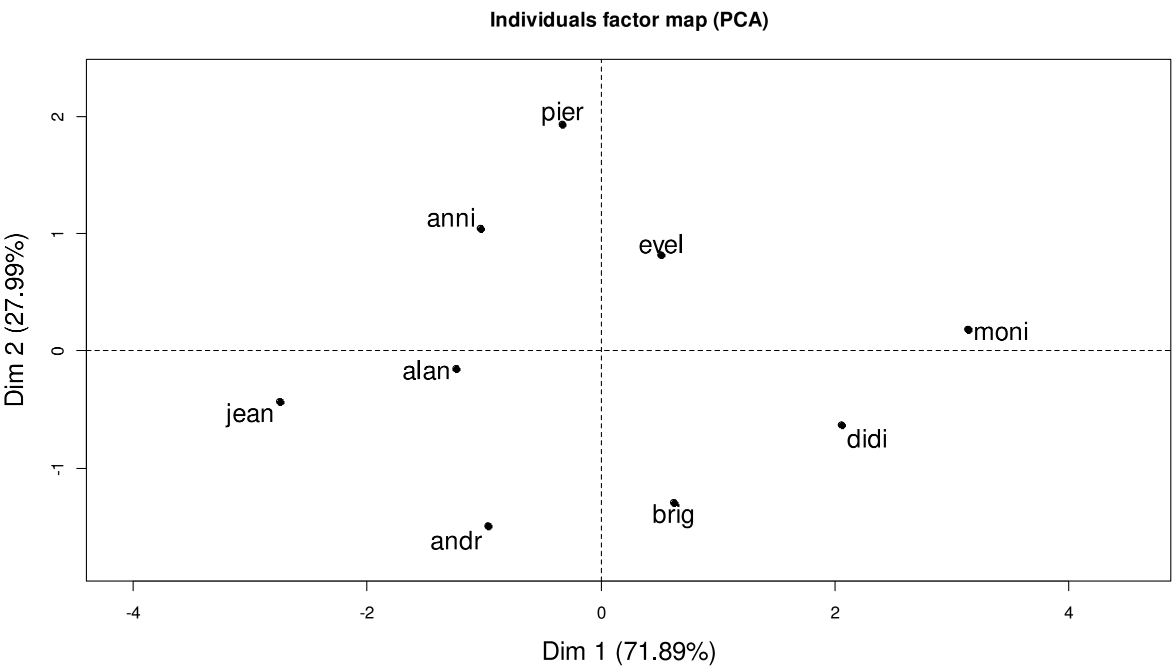
Le tableau donné ci-dessous contient tous les résultats importants de l'A.C.P. sur les individus.

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 ci-dessous. Ce dernier 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.

Comment interprétez-vous les résultats obtenus sur les individus ?



Exercice 5 *Calcul des résultats de l'ACP*

Reprenons l'exemple précédent et essayons de produire nous même les résultats de l'ACP.

Notons n=9 le nombre d'observations, p=4 le nombre de variables et X_1, X_2, X_3, X_4 les variables du tableau.

1) Importer le fichier Notes.txt dans R.
2) Afficher la matrice de corrélation (`cor`). Que remarquez-vous ?
3) Calculer la matrice de variance-covariance (`cov`). Puis en déduire ses valeurs propres (`eigen`).
4) Quelle est l'inertie (variance) du nuage de points ? Par quelle autre méthode pouvez-vous déterminer l'inertie ?
5) Centrer et réduire les variables et noter Z la matrice d'ordre $n \times p$ constituée des données centrées et réduites.

```
n=nrow(Mydata)
moyenne=apply(Mydata,2,mean)
variance=(n-1)*variance/n #A
Z=Mydata
for (i in 1:4)
{
  Z[,i]=(Mydata[,i]-moyenne[i])/sqrt(variance[i])
}
```

   a) Vérifier que la matrice de corrélation est égale à $R = \frac{1}{n} {}^{t}ZZ$.
   b) Que devient alors l'inertie ?
6) Quel pourcentage de l'inertie est représenté sur les deux premiers axes principaux de l'ACP ?
7) Projection des points
   L'ACP consiste à faire un changement de repère pour représenter les points. Les axes de ce nouveau repère sont les vecteurs propres de la matrice de variance-covariance R. Si on note pour k=1,…,p, $C_k \in R^n$ le vecteur des projections des n points sur l'axe de vecteur directeur u_k, alors
   $$C_k = Zu_k.$$
   a) Calculer les coordonnées des élèves sur les deux premiers axes principaux et faites une représentation graphique.
   b) Vérifier que les composantes principales C_1 et C_2 sont centrées et non corrélées.
   c) Vérifier que $var(C_k)=\lambda_k$. (Attention il s'agit de la variance s^{*2}

[A] Il existe deux façons de calculer une variance suivant qu'il s'agit de la variance d'un échantillon $s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ou d'une population $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Vous aurez l'explication au second semestre en statistique inférentielle. La fonction var de R calcule s^2. La relation entre les deux est $s^{*2} = \frac{n-1}{n}s^2$.

8) Corrélation des variables avec les nouveaux axes

Si on transpose le jeu de données alors les variables deviennent des points dans un espace de dimension n.

```
> t(Z)
          jean       alan       anni     moni      didi        andr       pier       brig       evel
MATH -1.086503 -0.4938648 -1.0865026 1.432208 1.2840485  0.39509184 -1.2346620  0.9877296 -0.1975459
PHYS -1.281740 -0.6130060 -0.9473730 1.560379 1.3931955  0.05572782 -0.9473730  0.8916451 -0.1114556
FRAN -1.503663 -0.6398567  0.2239498 1.519660 0.5118854 -1.35969546  1.0877564 -0.4958889  0.6558531
ANGL -1.619403 -0.7307061 -0.1974881 1.757644 0.8689478 -1.08618470  0.5134691 -0.1974881  0.6912084
```

Cet espace est une hyper-sphère de rayon 1 car les variables sont centrées et réduites. On détermine les axes de l'ACP de telle façon que les angles entre les variables projetées sur ces axes soient le moins déformés possible.

Les coordonnées des variables $X_1,\ldots,X_p$ dans l'hypersphère sont données par leur coefficient de corrélation avec les axes,

$$F_k=\mathrm{cor}(Z,C_k),\ k=1,\ldots,p.$$

N.B. Matrice Z et on matrice X car les variables sont centrées-réduites

Représenter le cercle trigonométrique et la projection des variables $X_1,\ldots,X_p$ sur les deux premières composantes principales.