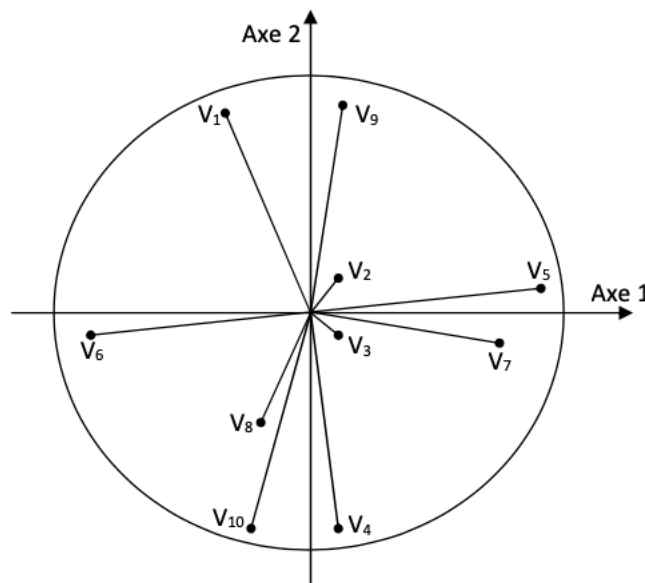


## TD5 : Analyse en composantes principales

**Exercice 1 :** 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 V4.
6. Quel est approximativement le coefficient de corrélation entre la variable V1 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 V3 et V7 ?
9. Que signifie le coefficient de corrélation entre la première et la deuxième composante principale ?

**Exercice 2 :** 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 3 :** 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.

<i>Statistiques élémentaires</i>				
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

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

### 2-Résultats sur l'inertie

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 ?

### ***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	

### ***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

PCT.VAR= pourcentage de variance

PCT.CUM=pourcentage cumulé : exemple  $(28.23/40.3)*100=70\%$  .

a) Quelle est la relation entre  $\lambda_i$  et 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  $C_i$ .

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

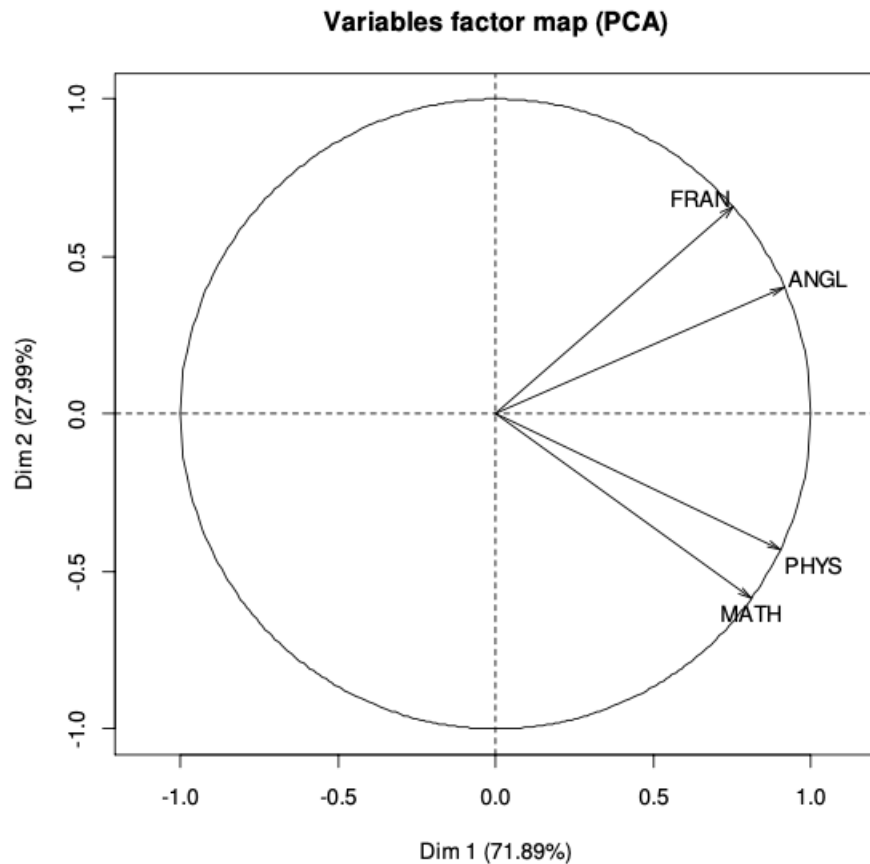
### **3-Résultats sur les variables**

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 ?

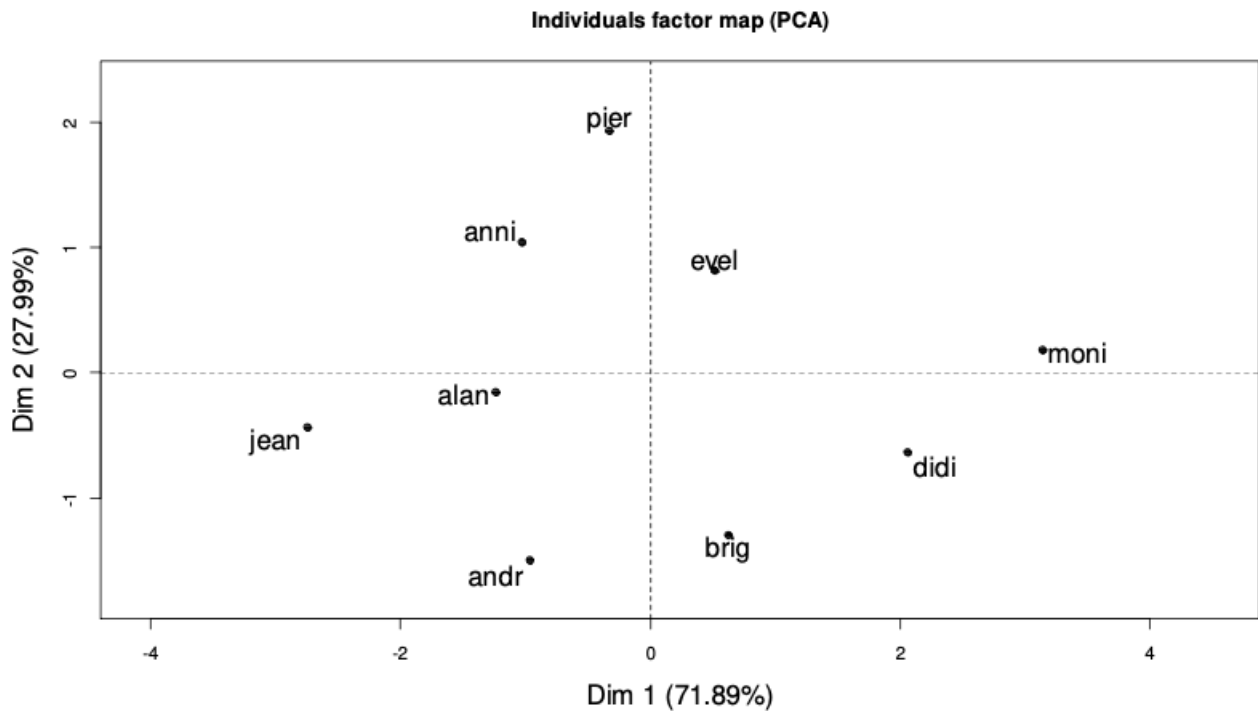
#### 4-Résultats sur les individus

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

	<i>Coordonnées des individus ; contributions ; cosinus carrés</i>							
	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 4 :** L'objectif de cet exercice est d'apprendre à utiliser le package R permettant de faire une ACP : ***FactoMineR***.

1. Installer et charger ce package dans votre session de travail.
2. Lire le jeu de données utilisé en illustration du cours : **EspVieACPData.txt**
3. Préparation des données
  - a- Représenter les nuages de points des données. Y-a-t-il des individus atypiques ? Quelles sont les variables corrélées ?
  - b- Centrer et réduire les variables.
4. Faire une ACP avec FactoMineR
  - a- Afficher l'aide R concernant la fonction PCA
  - b- Faire une ACP avec les variables TNAT, TMORT, EV, T65, NBENF et TCR en en gardant toutes les composantes principales et en affichant les graphiques sur les axes 1 et 2.
  - c- Afficher le diagramme des valeurs propres
  - d- Afficher les résultats concernant les variables
 

Calculer la somme du cos2 de TNAT. Sur Quel(s) axe(s) la variable TMORT est-elle bien représentée ? Quelles variables contribuent à la formation de l'axe 1 ?
  - e- Afficher les résultats sur les individus
 

Quel(s) axe(s) faut-il afficher pour avoir des informations concernant le Bangladesh ?

Quelle est la contribution moyenne d'un pays à la construction des axes ? Y-a-t'il des pays qui

dépassent très largement cette contribution moyenne ? Supprimer le pays ayant la plus grande contribution et regarder si cela change la construction des axes.

f- Ajouter la variable Continent sur le graphique des individus

**Exercice 5 :** Le jeu de données **DecathlonData.xls** de données contient 41 lignes et 13 colonnes.

Les colonnes 1 à 12 sont des variables quantitatives :

Les dix premières colonnes correspondent aux performances des athlètes pour les dix épreuves du décathlon.

Les colonnes 11 et 12 correspondent respectivement au rang et au nombre de points obtenus.

La dernière colonne est une variable qualitative correspondant au nom de la compétition (Jeu Olympiques de 2004 ou Décastar 2004). Les lignes désignent les athlètes.

Nous allons faire une ACP sur les colonnes de 1 à 10.

1. Quel pourcentage de l'inertie totale contiennent les deux premières composantes principales ? Combien faut-il choisir de composantes principales pour avoir plus de 70% de l'inertie totale ?

## 2. Etude des variables.

a. Pourquoi les variables "X100m ", "X400m", "X110m.hurdle " et "X1500m" se trouvent-elles à gauche de l'axe des ordonnées ?

b. Comment interprétez-vous la corrélation entre les variables "X100m " et " long.jump" ?

c. Peut-on distinguer des groupes de variables ? Quelle est la corrélation entre ces groupes ? Comment l'interprétez-vous ?

d. Quelles variables contribuent majoritairement à la première composante principale, à la deuxième composante principale ? Comment pouvez-vous interpréter le plan défini par les deux premières composantes principales ?

## 3. Etude des individus

a. Comment qualifieriez-vous l'athlète Lorenzo ? A votre avis, comment se fait-il qu'il ne soit pas dernier de sa compétition ?

b. Comment qualifieriez-vous les athlètes suivants : Karpov, Sebrle, Casarsa ? Quel est leur classement ?

c. On ajoute les variables supplémentaires "Rank" et "Points". Ces variables n'entrent pas en compte dans le calcul des composantes principales mais aident à une meilleure compréhension des axes. Que pouvez-vous en conclure ?

d. Comparer la position de Karpov, Clay, ... aux jeux olympiques et au décastar. Peut-on en conclure que le niveau des deux compétitions n'est pas le même ? Pour répondre à cette question, on ajoute la variable supplémentaire "Competition". Cette variable est qualitative et est qualifiée de facteur. Deux nouveaux individus représentant un individu moyen pour chaque compétition sont ajoutés au graphique. Que pouvez-vous en conclure ?

**Exercice 6 :** Le jeu de données **EconomieEuropaData.txt** contient 20 lignes et 7 colonnes.

Les lignes désignent des pays.

Les colonnes 1 à 5 sont des données économiques des pays.

La colonne 6 est la population du pays.

La colonne 7 est une variable qualitative désignant l'appartenance du pays à la zone euro.

Nous allons faire une ACP sur les colonnes de 1 à 5. Les colonnes 6 et 7 seront ajoutées comme variable et facteur supplémentaires.

1. Quel pourcentage de l'inertie totale contiennent les deux premières composantes principales ? Combien faut-il choisir de composantes principales pour avoir plus de 70% de l'inertie totale ?
2. Interpréter les axes et la liaison entre les variables.
3. Etude des individus : Que pouvez-vous dire du Luxembourg, de la Grèce et de la Slovaquie ? Peut-on caractériser les pays de la zone euro ?
4. Que pouvez-vous conclure sur la contribution des individus (pays) à la construction des axes ? Refaite une ACP avec le Luxembourg (individu 13) et la Slovaquie (individu 17) en individus supplémentaires. Que se passe-t-il ? Que pouvez-vous dire sur vos conclusions précédentes ?