Analyses descriptives multivariées : ACP, ACM.

Master 2 Recherche SES-IES Analyse de données

Ana Karina Fermin

Université Paris-Ouest-Nanterre-La Défense

http://fermin.perso.math.cnrs.fr/

Données

En général, un tableau de données avec

- en lignes : les individus
- en colonnes : les variables (quantitatives ou qualitatives)

Motivation

- L'Analyse en Composantes Principales (ACP) :
  tableau croisant des individus et des variables numériques
- L'Analyse des Correspondances Multiples (ACM) :
  tableaux croisant des individus et des variables qualitatives
- L'Analyse Factorielle des Correspondances (AFC) :
  tableaux de fréquence

Exemple 1 : Notes obtenues par 12 élèves dans 4 disciplines ($p = 4$, $n = 12$).
Cet exemple jouet est une pure fiction. Toute ressemblance avec des personnages existants ou ayant existé serait fortuite et indépendante de la volonté de l'enseignante ...

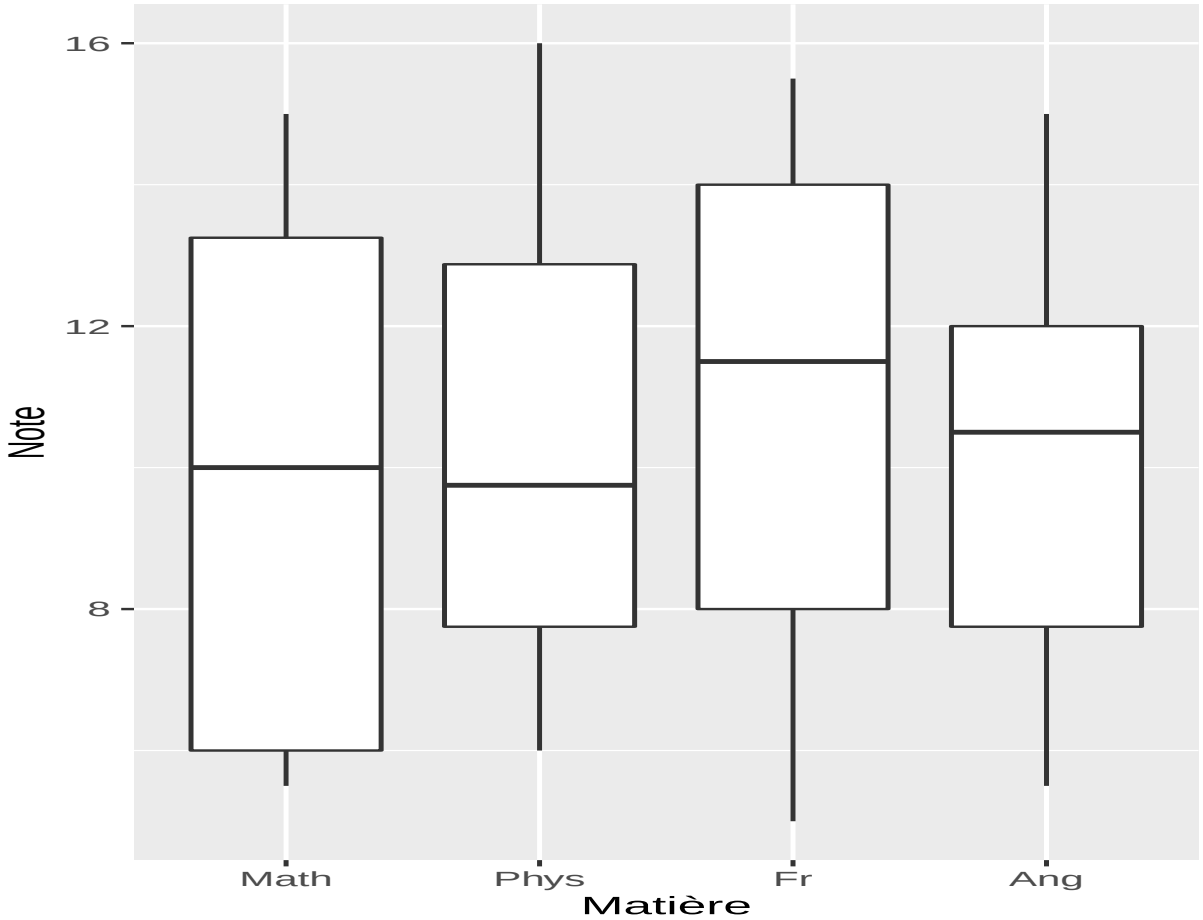
	Math	Phys	Fr	Ang
Rémi	6.0	6.0	5.0	5.5
Thomas	8.0	8.0	8.0	8.0
Gaëtan	6.0	7.0	11.0	9.5
Lenny	14.5	14.5	15.5	15.0
Louis.E	14.0	14.0	12.0	12.5
Louis.H	11.0	10.0	5.5	7.0
Antoine	5.5	7.0	14.0	11.5
Raphaël	13.0	12.5	8.5	9.5
Jean	9.0	9.5	12.5	12.0
Abdou	12.0	11.5	14.0	12.0
Matthieu	6.0	8.0	8.0	7.0
Sophie	15.0	16.0	14.0	12.0

Exemple 2 : Données décathlon

- $p = 10$ variables quantitatives continues (mesurées pour n = 41 athlètes)
- 100m, longueur, poids, hauteur, 400m, 110m haies, disque, perche, javelot, 1500m

A vous de jouer ! A faire dans l'atelier R

Décrire le jeu de données de l'exemple 1.



Statistiques Descriptives des 4 variables :

	Matiere	Note.min	Note.max	Note.moy	Note.var
	(fctr)	(dbl)	(dbl)	(dbl)	(dbl)
1	Math	5.5	15.0	10.00000	12.375000
2	Phys	6.0	16.0	10.33333	10.138889
3	Fr	5.0	15.5	10.66667	11.638889
4	Ang	5.5	15.0	10.12500	7.338542

Matrice de variance-covariances

	Math	Phys	Fr	Ang
Math	12.375000	10.895833	4.937500	5.770833
Phys	10.895833	10.138889	5.673611	5.812500
Fr	4.937500	5.673611	11.638889	8.729167
Ang	5.770833	5.812500	8.729167	7.338542

La somme des variances des 4 variables initiales vaut 41.49,
c'est-à-dire, la dispersion totale des individus considérés vaut 41.49.

Matrice de corrélation

	Math	Phys	Fr	Ang
Math	1.0000000	0.9727308	0.4114137	0.6055654
Phys	0.9727308	1.0000000	0.5222863	0.6738495
Fr	0.4114137	0.5222863	1.0000000	0.9445221
Ang	0.6055654	0.6738495	0.9445221	1.0000000

Toutes les corrélations linéaires sont positives, certaines étant très fortes (0.97 et 0.94) et d'autres moyennes (0.67 et 0.41).

- Quels sont les individus qui se ressemblent (ou non) ?
- Quelles sont les relations entre les variables quantitatives (celles qui se ressemblent ou non) ? Analyse des corrélations ? Mais alors analyse des corrélations 2 à 2, comment faire pour obtenir une information plus globale ?

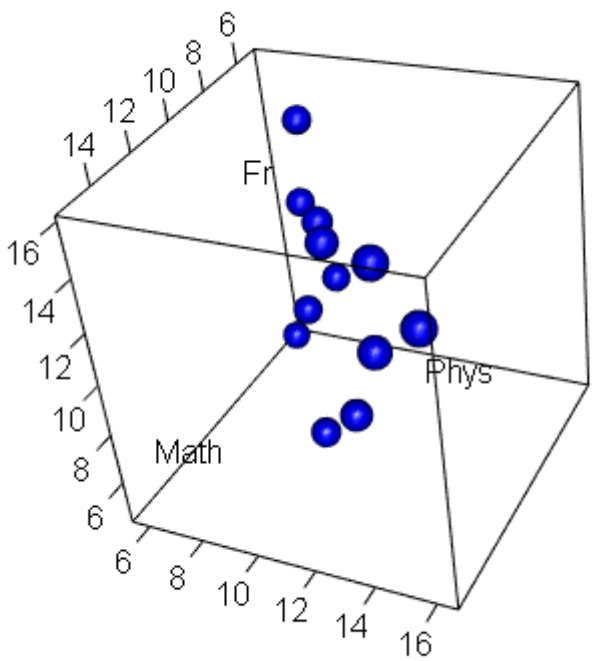
ACP ou PCA en anglais pour Principal Component Analysis

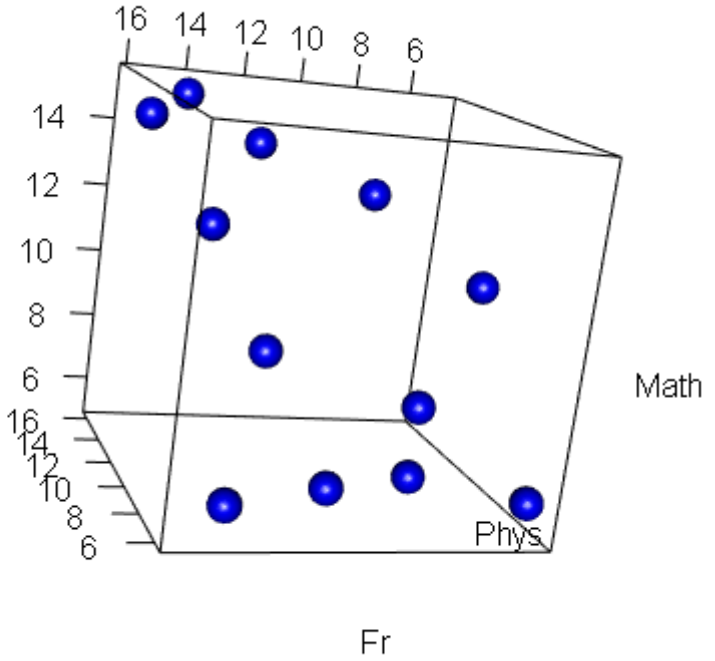
- Lorsqu'on étudie simultanément un nombre p important de variables quantitatives (par exemple $p = 4$), comment en faire un graphique global ?

- L'objectif de l'ACP est la représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace de dimension réduite q avec $(q < p)$. Le but est de déformer le moins possible la réalité en explicitant au "mieux" les liaisons initiales entre ces variables.

- Mathématiquement, l'ACP est une méthode simple qui consiste à faire un changement de base (une rotation) de sorte que l'image soit la plus claire possible.

L'ACP est une méthode simple qui consiste à faire une rotation de sorte que l'image soit la plus claire possible.





But : Séparer le plus possible les points !

Vecteurs propres

ACP

Mathématiquement, l'idée est de passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les **vecteurs propres** de la matrice de corrélations. .

- Soit $\mathbf{X}$ la matrice des données initiales.
- On veut passer la matrice des données $\mathbf{X}$ à une nouvelle matrice $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{XR}$$

où $\mathbf{R}$ est una matrice de base "bien choisie".

$$\mathbf{Y} = \mathbf{XR}$$

où $\mathbf{R}$ est choisie de telle sorte que :

1. la variance des variables aille en ordre décroissante

$$\mathrm{var}(\mathrm{Y}_{(1)}) \geqslant \mathrm{var}(\mathrm{Y}_{(2)}) \geqslant \mathrm{var}(\mathrm{Y}_{(\mathrm{q})}).$$

Il est important de remarquer que c'est la a première "nouvelle variable" $Y_{(1)}$ qui contient le plus d'information.

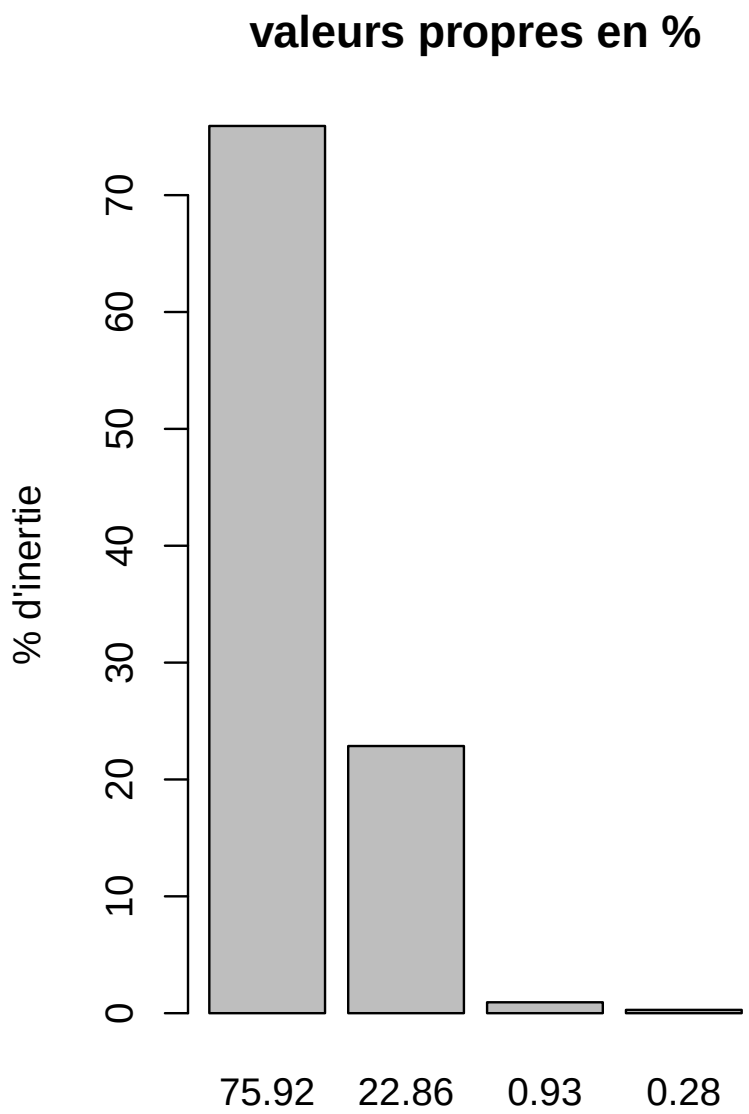
2. les variables $Y_{(i)}$ ne sont plus corrélées.
Cette condition est là pour s'assure que toute la variabilité de la seconde $Y_{(2)}$ est indépendante de celle de la première. Ainsi, on a des nouvelles variables $Y_{(i)}$ qui portent chacune de l'information (les premières en portent le plus)

Chaque ligne du tableau ci-dessus correspond à une variable virtuelle (composantes) dont la colonne eigenvalue (valeur propre) fournit la variance.

	eigenvalue	% of variance	cumulative % of variance
comp 1	31.50	75.92	75.92
comp 2	9.49	22.86	98.78
comp 3	0.39	0.93	99.72
comp 4	0.12	0.28	100.00

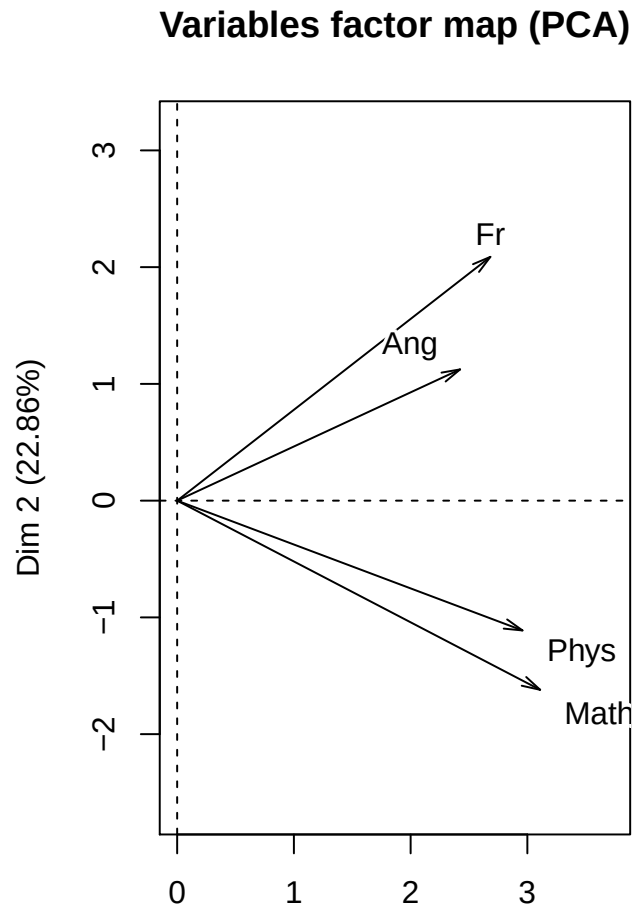
La somme des 4 valeurs propres obtenues vaut 41.49.

Analyse de la décroissance des valeurs propres



- Combien d'axes retiendriez-vous pour l'ACP de ce jeu de données ? Pas de méthode pour choisir le nombre d'axes, critères empiriques. On va choisir ici $q = 2$.
- quelle est la part d'inertie expliquée par le plan principal ?
- visualiser les coordonnées des individus dans le plan factoriel
- commenter les contributions des élèves aux variables principales.

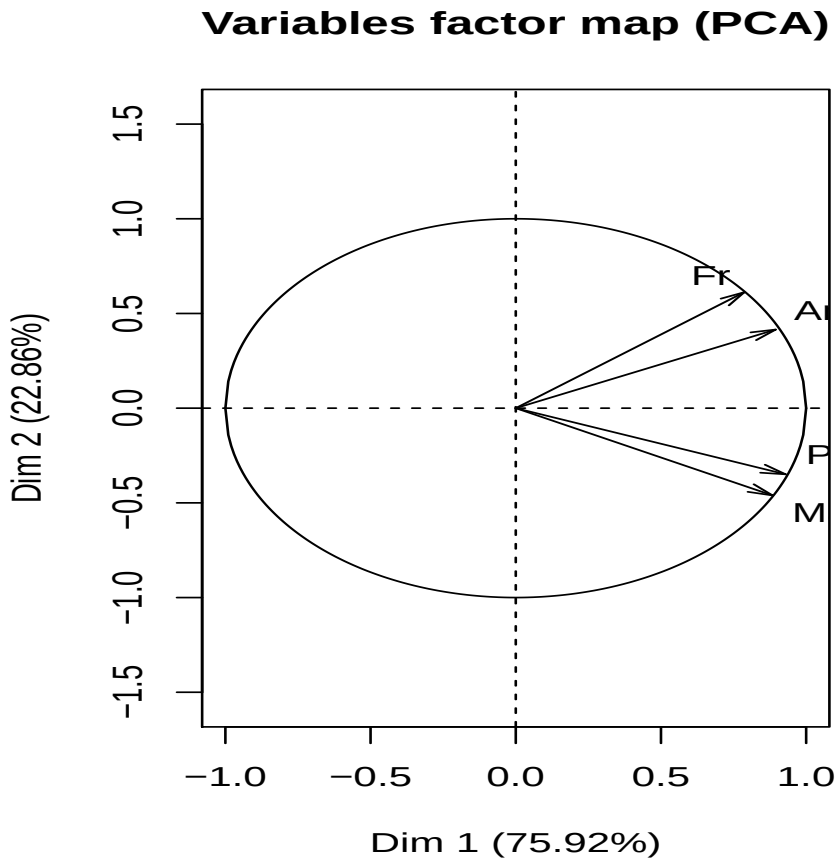
	Dim.1	Dim.2	Dim.3	Dim.4
Math	3.109711	-1.620314	0.2084252	-0.1893098
Phys	2.958512	-1.111935	-0.3389558	0.1865700
Fr	2.683554	2.088091	-0.2414257	-0.1378911
Ang	2.423972	1.124136	0.4135945	0.1678104



```
Notes.PCA$var$cor
         Dim.1      Dim.2       Dim.3       Dim.4
Math 0.8839902 -0.4606028  0.05924854 -0.05381464
Phys 0.9291334 -0.3492080 -0.10645053  0.05859312
Fr   0.7866012  0.6120596 -0.07076651 -0.04041854
Ang  0.8947932  0.4149675  0.15267568  0.06194609
```

- 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.
- Le deuxième axe corresponds à deux type de profil : les littéraires (corrélations positives) en haut et les scientifiques en bas (corrélations négatives).

En regardant le cercle de corrélation on voit bien que les deux axes capturent presque toute l'information.

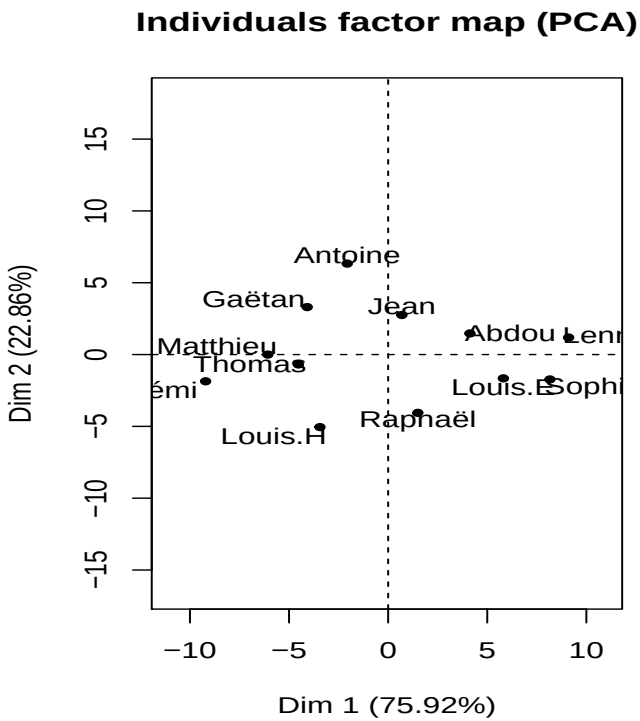


Coordonnés des individus

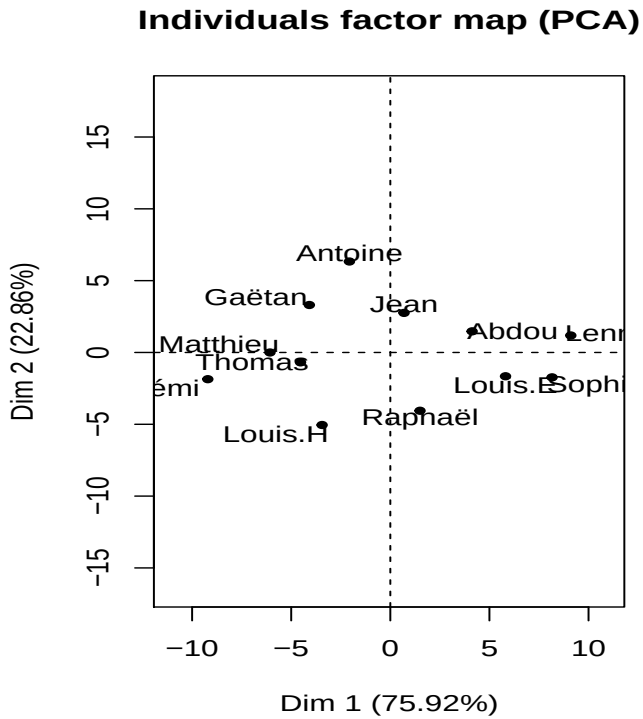
	Dim.1	Dim.2	Dim.3	Dim.4
Rémi	-9.2074308	-1.861117273	0.14504724	-0.133927598
Thomas	-4.5309014	-0.688953550	0.22317918	-0.132841116
Gaëtan	-4.0839197	3.305703670	-0.06877213	-0.045174385
Lenny	9.1061334	1.184506316	0.60224332	0.224566941
Louis.E	5.8123304	-1.657342237	0.40352309	0.412368582
Louis.H	-3.4416729	-5.049254014	0.44374695	-0.184907474
Antoine	-2.0627627	6.332679179	-0.07086136	0.003193134
Raphaël	1.4984208	-4.057606580	0.24968588	0.087948541
Jean	0.6930346	2.754284456	0.65353112	0.278743323
Abdou	4.1267051	1.470895954	-0.01262482	-0.891323810
Matthieu	-6.0709261	-0.001752782	-1.11057190	0.481314549
Sophie	8.1609894	-1.732043139	-1.45812657	-0.099960687

Coordonnés des individus

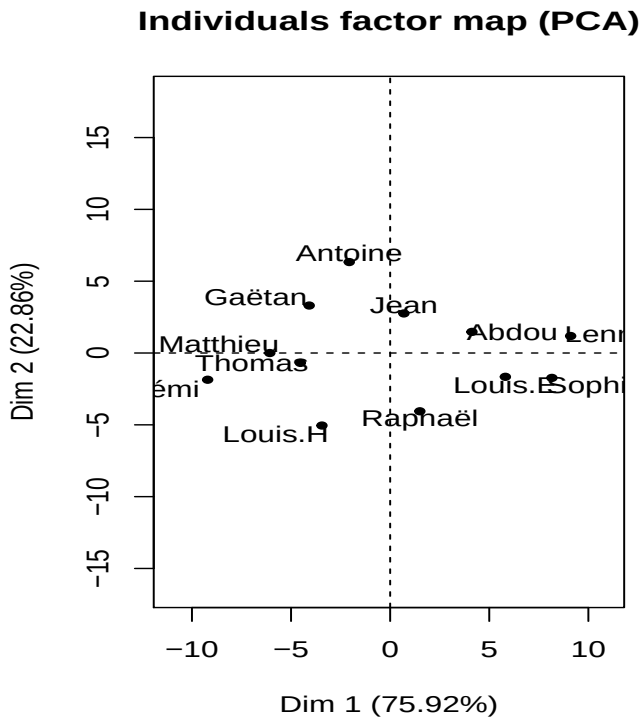
Par exemple , Lenny : (9.1061334 , 1.184506316)



- Dans la représentation d'individus, l'axe 1 capture presque 76% de l'énergie totale. Il correspond, comme on l'a vu précédemment, au niveau global de l'élève. Par exemple, Lenny apparait comme un meilleur élève que Rémi.



L'axe 2 capture presque 23% de l'énergie. Cet axe correspond à la différence entre littéraire et scientifique. Par exemple, l'élève "le plus haut" sur le graphique est Antoine 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 Louis.H qui obtient la moyenne dans les disciplines scientifiques (11 et 10) mais des résultats très faibles dans les disciplines littéraires (7 et 5.5). On notera que Matthieu et Thomas qui ont un score voisin de 0 sur l'axe 2 car ils ont des résultats très homogènes dans les 4 disciplines.



Quand les variables sont mesurées sur différentes échelles

Centrage et réduction

- Centrage : écart à la moyenne
- Réduction : toutes les variables ont même unité de variation

Fait par le logiciel R avec la fonction scale. Fait automatiquement avec la library FactorMineR (voir argument scale).
A faire dans l'atelier avec les données de décathlon. Que feriez vous si vous aviez une variable qualitative ?

ACM