Analyse en Composantes Principales et application avec R

S. Jaubert

23 mai 2020

Elaborée par Karl Pearson en 1901 l'analyse en composantes principales (ACP) nous permet de résumer et de visualiser les informations dans un ensemble de données contenant des individus (ou observations) décrits par de multiples **variables quantitatives**. Harold Hotelling dans les années 30 formalisa la description mathématiques de ces méthodes mais celles ci nécessitant des calculs très importants il fallu attendre l'avènement des ordinateurs pour réellement les mettre en pratique.

Figure 1: Karl Pearson à son bureau 1910

Chaque variable peut être considérée comme une dimension différente et au delà de 3 variables il est très difficile de visualiser un tel hyperespace.

Voici par exemple un tableau (une matrice) de 27 individus décrits par 11 variables (données issues du Package **factoextra** version 1.0.7 de R) :

```
##              100m long. poids haut.  400m  110m Disque Perche  Jav.  1500m  Pts
## SEBRLE      11.04  7.58 14.83  2.07 49.81 14.69  43.75   5.02 63.19 291.70 8217
## CLAY        10.76  7.40 14.26  1.86 49.37 14.05  50.72   4.92 60.15 301.50 8122
## BERNARD     11.02  7.23 14.25  1.92 48.93 14.99  40.87   5.32 62.77 280.10 8067
## YURKOV      11.34  7.09 15.19  2.10 50.42 15.31  46.26   4.72 63.44 276.40 8036
## ZSIVOCZKY   11.13  7.30 13.48  2.01 48.62 14.17  45.67   4.42 55.37 268.00 8004
## McMULLEN    10.83  7.31 13.76  2.13 49.91 14.38  44.41   4.42 56.37 285.10 7995
## MARTINEAU   11.64  6.81 14.57  1.95 50.14 14.93  47.60   4.92 52.33 262.10 7802
## HERNU       11.37  7.56 14.41  1.86 51.10 15.06  44.99   4.82 57.19 285.10 7733
## BARRAS      11.33  6.97 14.09  1.95 49.48 14.48  42.10   4.72 55.40 282.00 7708
```

```
## NOOL        11.33  7.27 12.68  1.98 49.20 15.29  37.92   4.62 57.44 266.60 7651
## BOURGUIGNON 11.36  6.80 13.46  1.86 51.16 15.67  40.49   5.02 54.68 291.70 7313
## Sebrle      10.85  7.84 16.36  2.12 48.36 14.05  48.72   5.00 70.52 280.01 8893
## Clay        10.44  7.96 15.23  2.06 49.19 14.13  50.11   4.90 69.71 282.00 8820
## Karpov      10.50  7.81 15.93  2.09 46.81 13.97  51.65   4.60 55.54 278.11 8725
## Macey       10.89  7.47 15.73  2.15 48.97 14.56  48.34   4.40 58.46 265.42 8414
## Warners     10.62  7.74 14.48  1.97 47.97 14.01  43.73   4.90 55.39 278.05 8343
## Zsivoczky   10.91  7.14 15.31  2.12 49.40 14.95  45.62   4.70 63.45 269.54 8287
## Hernu       10.97  7.19 14.65  2.03 48.73 14.25  44.72   4.80 57.76 264.35 8237
## Bernard     10.69  7.48 14.80  2.12 49.13 14.17  44.75   4.40 55.27 276.31 8225
## Schwarzl    10.98  7.49 14.01  1.94 49.76 14.25  42.43   5.10 56.32 273.56 8102
## Pogorelov   10.95  7.31 15.10  2.06 50.79 14.21  44.60   5.00 53.45 287.63 8084
## Schoenbeck  10.90  7.30 14.77  1.88 50.30 14.34  44.41   5.00 60.89 278.82 8077
## Barras      11.14  6.99 14.91  1.94 49.41 14.37  44.83   4.60 64.55 267.09 8067
## KARPOV      11.02  7.30 14.77  2.04 48.37 14.09  48.95   4.92 50.31 300.20 8099
## WARNERS     11.11  7.60 14.31  1.98 48.68 14.23  41.10   4.92 51.77 278.10 8030
## Nool        10.80  7.53 14.26  1.88 48.81 14.80  42.05   5.40 61.33 276.33 8235
## Drews       10.87  7.38 13.07  1.88 48.51 14.01  40.11   5.00 51.53 274.21 7926
```

Les lignes représentent les individus, les colonnes les variables.

Les individus CLAY (2ème ligne) et BERNARD (19ème ligne) peuvent-être considérés comme deux vecteurs à 11 composantes :

$CLAY = (11.04, 7.58, 14.83, 2.07, 49.81, 14.69, 43.75, 5.02, 63.19, 291.70, 8217)^T$

$BERNARD = (10.69, 7.48, 14.80, 2.12, 49.13, 14.17, 44.75, 4.40, 55.27, 276.31, 8225)^T$

et les variables comme des vecteurs à 27 composantes, par exemple :

$100m = (11.04, 10.76, 11.02, 11.34, 11.13, 10.83, 11.64, 11.37, 11.33, 11.33, 11.36, 10.85, 10.44,$
$10.50, 10.89, 10.62, 10.91, 10.97, 10.69, 10.98, 10.95, 10.90, 11.14, 11.02, 11.11, 10.80, 10.87)^T$

et

$Perche = (5.02, 4.92, 5.32, 4.72, 4.42, 4.42, 4.92, 4.82, 4.72, 4.62, 5.02, 5.00, 4.90, 4.60, 4.40,$
$4.90, 4.70, 4.80, 4.40, 5.10, 5.00, 5.00, 4.60, 4.92, 4.92, 5.40, 5.00)^T$

Comparer ces deux athlètes CLAY et BERNARD, voir quelles sont leurs similarités, n'est pas toujours très facile en grande dimension, de même savoir si les performances obtenues sur 100m sont corrélées à celles obtenues à la perche n'a rien d'évident.

L'ACP a pour objectif de synthétiser ces grandes quantités de données, de les résumer en considérant les ressemblances entre individus et en cherchant les liaisons éventuelles entre les variables.

En d'autres termes, l'ACP réduit la dimensionnalité d'une donnée multivariée à deux (voire trois) composantes principales, appelées aussi axes factoriels qui peuvent être visualisées graphiquement par les axes d'un ellipsoïde, avec une perte minimale d'informations.

L'analyse en composantes principales sera donc utilisée pour extraire les informations importantes d'un tableau de données multivarié et exprimera ces informations sous la forme d'un ensemble de nouvelles variables appelées composantes principales (ou facteurs). Ces nouvelles variables seront exprimées comme une combinaison linéaire des originales.

Notre nuage de données sera projeté selon des axes qui conserveront le maximum d'informations, notre nuage devra être le moins déformé possible après projection. Sur ces axes les écarts entre les données devront donc être le plus importants.

Exemple : Si on considère cette image prise sous deux angles différents, la deuxième où les point sont les plus dispersés nous donnera évidemment le plus d'information.

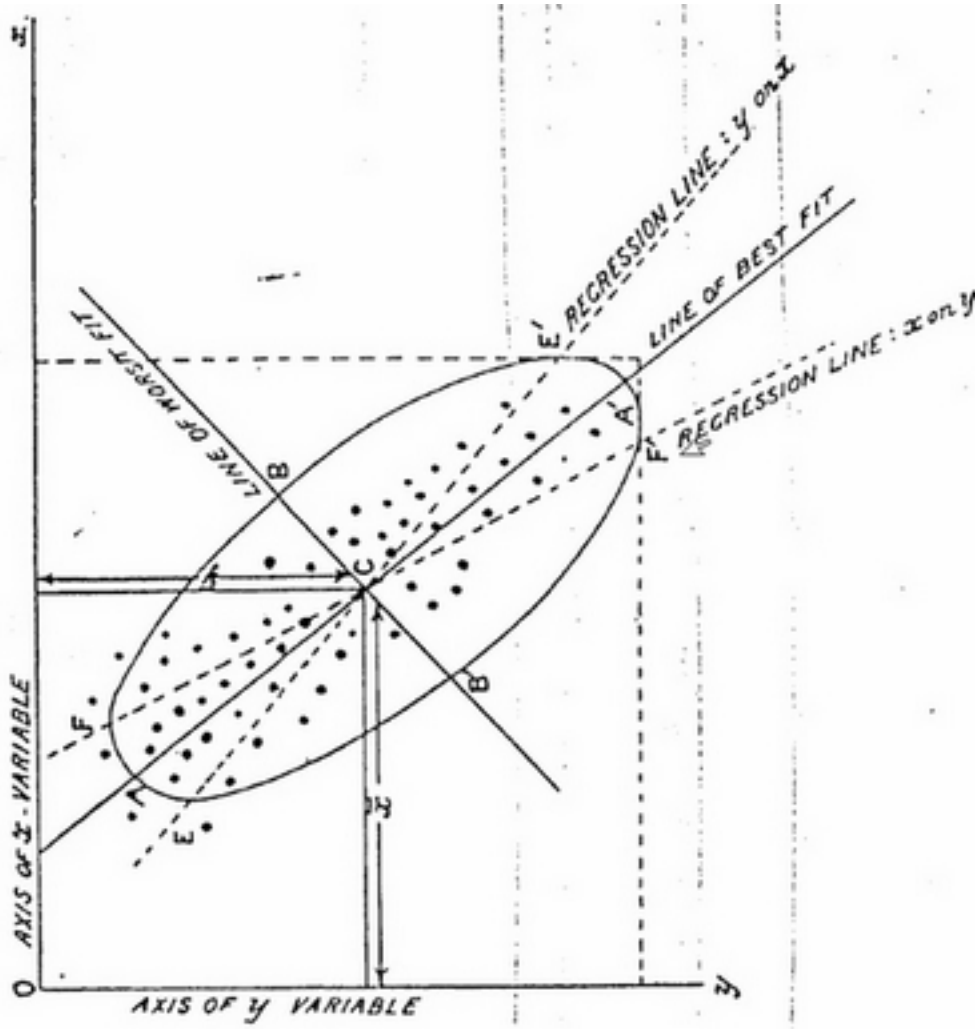


Figure 2: Extrait de l'article de Pearson de 1901 : la recherche de la « droite du meilleur ajustement ».

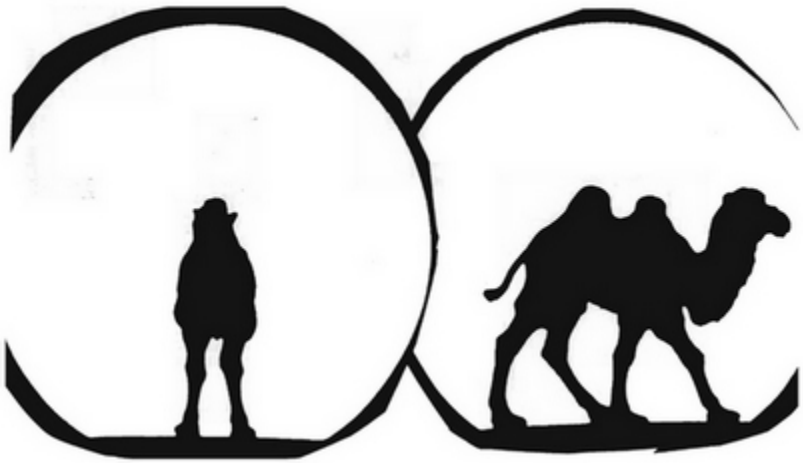


Figure 3: Source J.P FENELON "Qu'est ce que l'analyse des donnees ?"

Principe général de l'ACP

L'information contenue dans un ensemble de données correspond à la variation totale qu'il contient. On va chercher à identifier les directions (ou composantes principales) le long desquelles la variation des données est maximale.

Afin que l'on puisse garder l'idée générale en tête et ne pas se perdre dans des développements techniques, voir en Annexe les justifications mathématiques.

Figure 4: Via Tony Armstrong

Pour simplifier, on considérera que le repère du nuage est centré sur le centre de gravité, de plus pour éviter des problèmes d'échelles ou d'unités, nous réduirons nos variables en les divisant par leur écart-type.

L'idée générale est de déterminer le plan dans lequel la projection du nuage de points conservera le plus possible sa forme originale. Ce plan s'appelle le plan factoriel, défini par deux axes F_1 et F_2 (dits axes factoriels).

Pour trouver F_1, on cherche le vecteur (unitaire), qui passe par le centre de gravité du nuage, tel que la variance des points du nuage projetés orthogonalement sur ce vecteur soit maximale (ça revient à minimiser l'inertie du nuage qui tourne autour de F_1).

Comme on a perdu de l'information (le nuage se résume le long d'une droite) on détermine un second axe F_2 qui passe par le centre de gravité du nuage, orthogonale à F_1 et qui apporte le plus d'inertie (par rapport au centre de gravité). On appelle également F_1 et F_2 les axes principaux d'inertie. Les points du nuage seront exprimés dans le nouveau repère (F_1, F_2) (mais rien ne nous empêche de continuer sur le même principe est de chercher un 3ème axe F_3 qui passe par le centre de gravité perpendiculairement au plan $(F_1, F_2)\ldots$)

On démontre que F_1 n'est rien d'autre que le vecteur propre, de la matrice de corrélation associée à notre nuage de point dont la valeur propre λ_1 associée est maximale. Cette valeur propre représente l'inertie du nuage portée par l'axe F_1 ; F_2 est le deuxième vecteur propre de valeur propre $\lambda_2 \leq \lambda_1$ (voir Annexe).

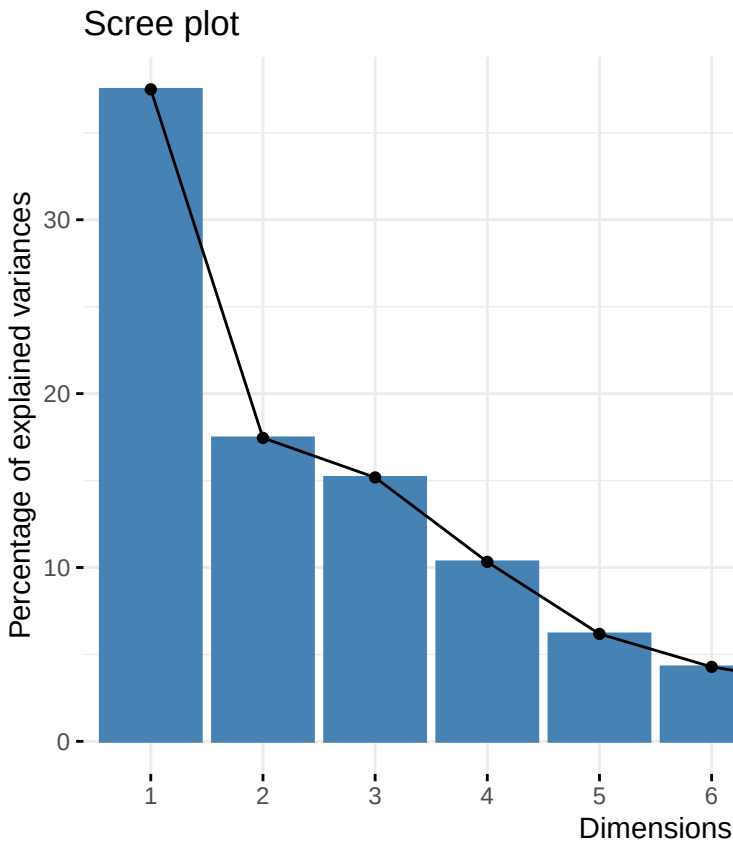
La qualité globale de représentation de nos données sur les k premières composantes principales $(F_1, F_2, .., F_k)$ (en générale k=2 ou 3) est mesurée comme le pourcentage de variance expliquée :

$$\frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{Trace}$$

Avec nos décathloniens nous avons pour valeurs propres :

```
##         eigenvalue percentage of variance cumulative percentage of variance
## comp 1   3.7499727              37.499727                           37.49973
## comp 2   1.7451681              17.451681                           54.95141
## comp 3   1.5178280              15.178280                           70.12969
## comp 4   1.0322001              10.322001                           80.45169
## comp 5   0.6178387               6.178387                           86.63008
## comp 6   0.4282908               4.282908                           90.91298
## comp 7   0.3259103               3.259103                           94.17209
## comp 8   0.2793827               2.793827                           96.96591
## comp 9   0.1911128               1.911128                           98.87704
## comp 10  0.1122959               1.122959                          100.00000
```

la proportion de chaque valeur propre est donnée par la deuxième colonne, le pourcentage cumulé par la 3ème colonne.



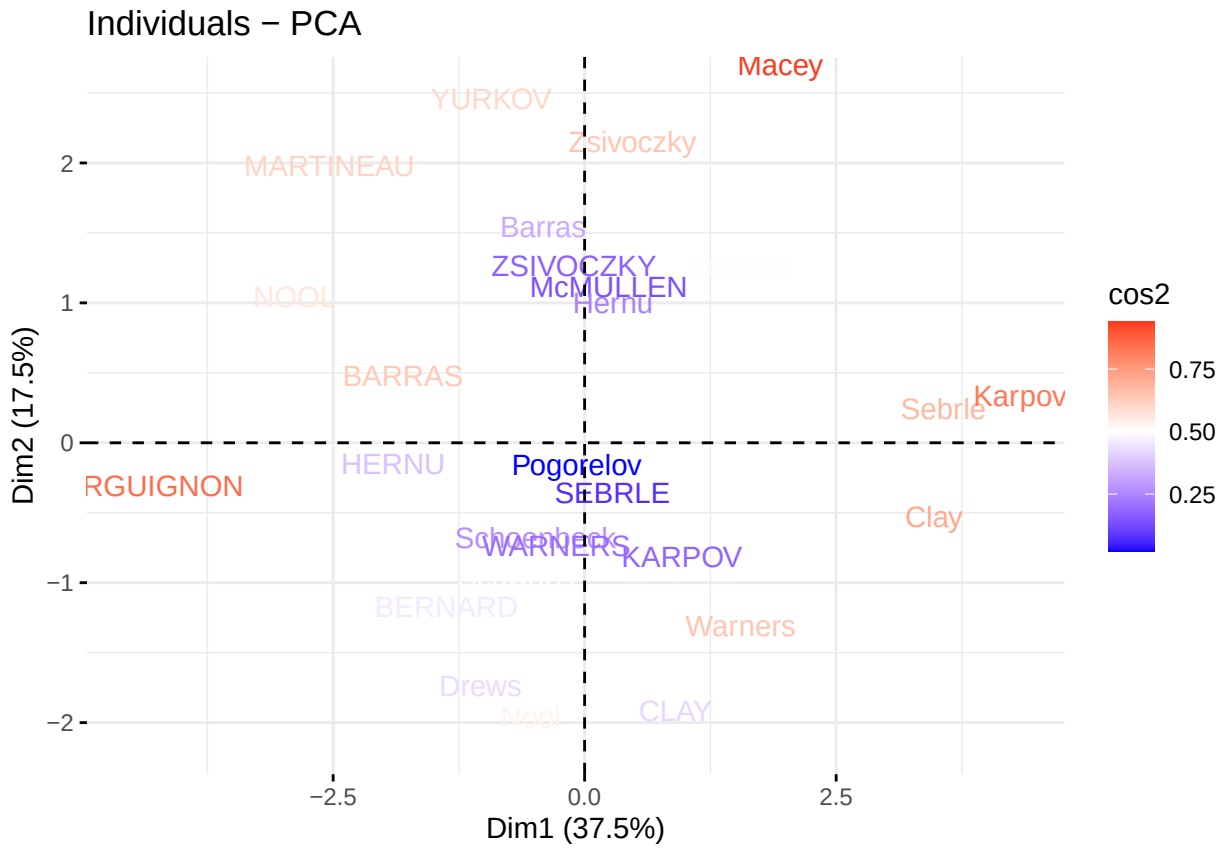
leur contribution peut aussi être représentée graphiquement :

Nous voyons que plus de 70% de la variation est expliquée par les trois premières composantes.

Dans cette nouvelle base (F_1, F_2), l'individu e_i s'écrira : $e_i = (f_{i,1}, f_{i,2})_{(F_1,F_2)}$ (en général on désire une représentation plane des individus)

Ici les coordonnées par exemple de CLAY et Karpov dans le plan factoriel sont :

```
##            Dim.1      Dim.2
## CLAY   0.9048536 -2.0942803
## Karpov 4.3287609  0.1607886
```



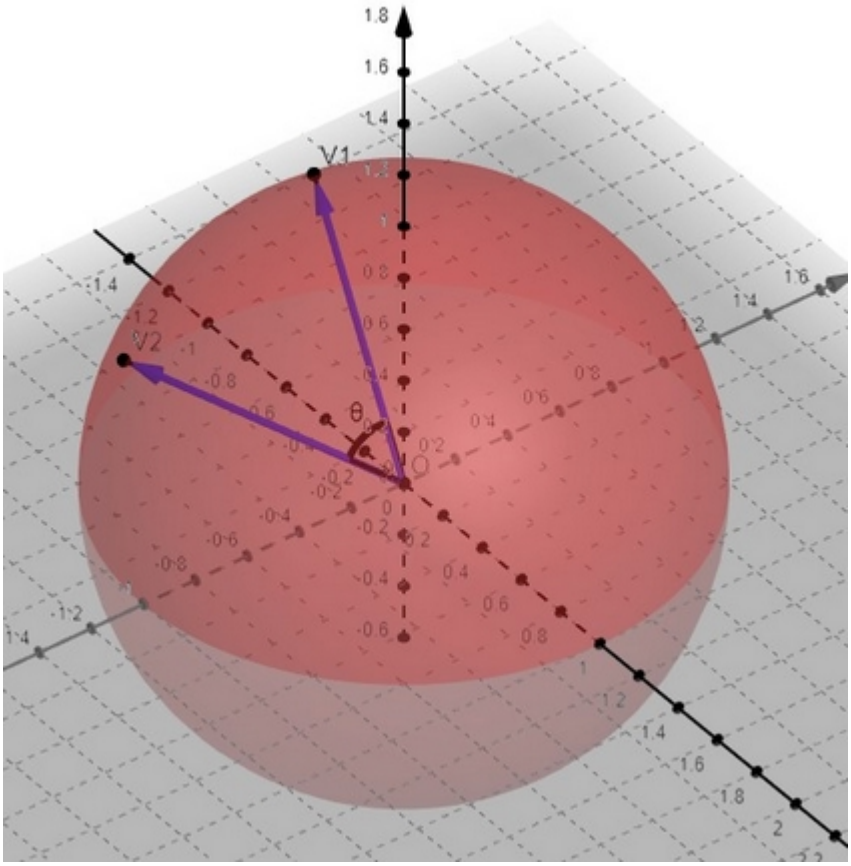
Nous pouvons obtenir la qualité de représentation des individus avec les deux premières composantes (les $\cos^2$) :

```
##            Dim.1       Dim.2
## CLAY   0.06557423 0.351274143
## Karpov 0.82674063 0.001140651
```

Représentation des variables

Deux variables particulières (par exemple le temps réalisé sur 100m et la distance du lancer du poids) seront très proches pour tous les individus si elles sont liées par leur coefficient de corrélation linéaire ; on le comprend bien dans le cas de la régression linéaire simple en dimension 2 ; si les points sont bien corrélés (donc sensiblement alignés) la structure de nos données devient unidimensionnelle, si une variable est connue l'autre aussi !

C'est donc la liaison entre les variables qui nous intéresse. Nos variables v_j (où $j = 1, .., p$) sont décrites dans un espace à n dimensions (n individus) Comme nos données sont centrées réduites, chaque vecteur v_j a pour norme 1 (c'est son écart-type). Ces vecteurs sont les rayons d'une hypersphère, le coefficient de corrélation entre les variables v_j et $v_{j'}$ n'est autre que le cosinus de l'angle formé par ces deux vecteurs (voir Annexe)



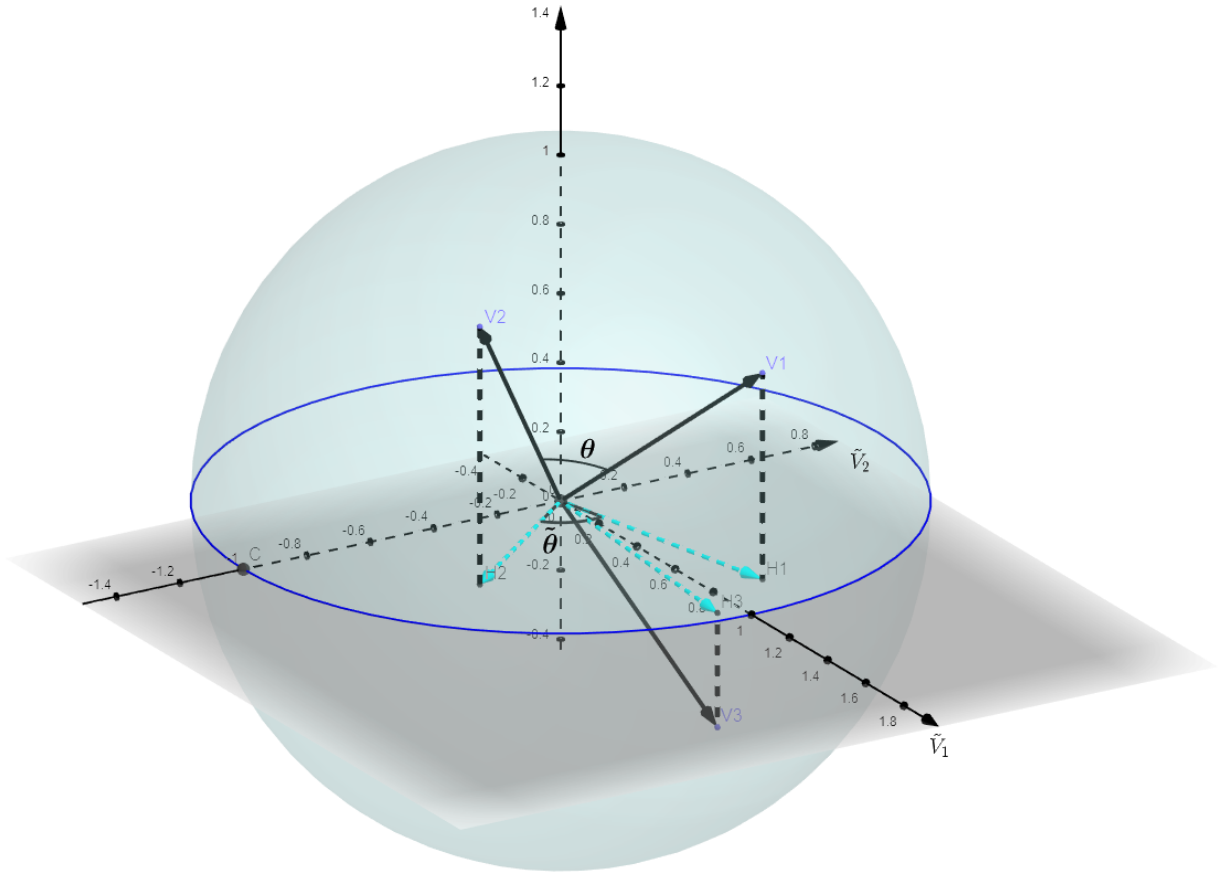
Nous avons donc un critère qui permet de mesurer le degré de liaison entre les variables initiales. Ce critère va nous permettre de regrouper celles qui sont fortement liées.

Nous procéderons de la même façon que pour le nuage des individus, hormis que la "distance" entre les variables sera ici mesurée par le carré du coefficient de corrélation linéaire.

Le premier axe factoriel $\tilde{V}_1$ sera tel qu'il maximisera la somme des carrés des coefficients de corrélations linéaires :

$$\tilde{V}_1 = \underset{\tilde{V} \in \mathbb{R}^n}{Arg\ Max} \quad \sum_{k=1}^{p} r(\tilde{V}, v_k)^2$$

Puis on continue en cherchant un axe $\tilde{V}_2$ orthogonal à $\tilde{V}_1$ qui maximise cette somme des carrés des coefficients de corrélations. Ces deux axes forment donc un plan dans lequel les variables initiales sont projetées.



La projection de l'hypersphère sur le plan factoriel donne :

Si v_1 et v_2 sont bien représentées on a $cos(\tilde{\theta}) \simeq cos(\theta)$

Le cercle des corrélations nous permet d'apprécier visuellement très facilement l'intensité des liaisons entre les variables et les axes factoriels, de plus les variables les mieux représentées seront celles proches du cercle (on conserve ainsi le maximum d'information).

Avec nos décathloniens nous obtenons :

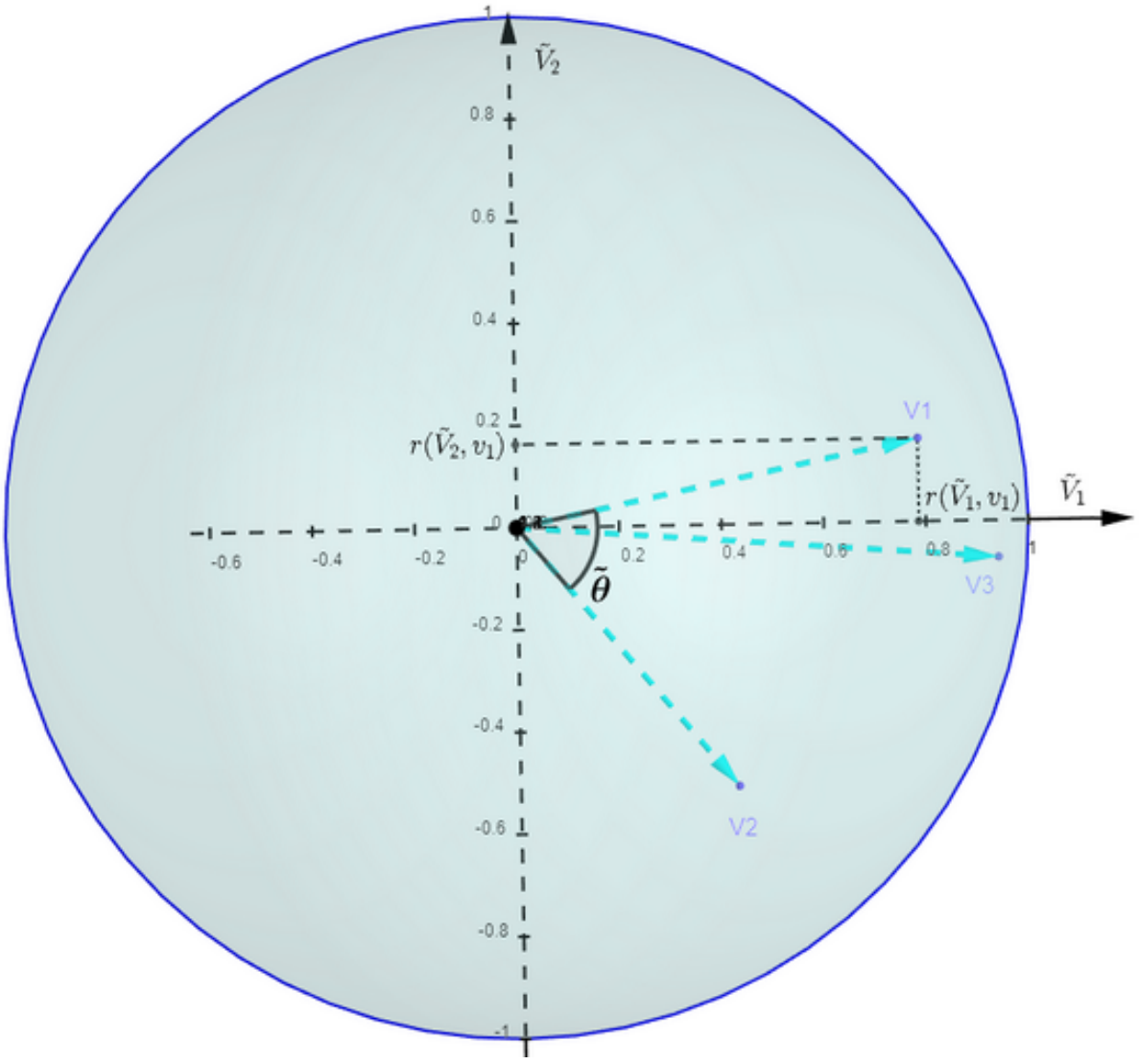
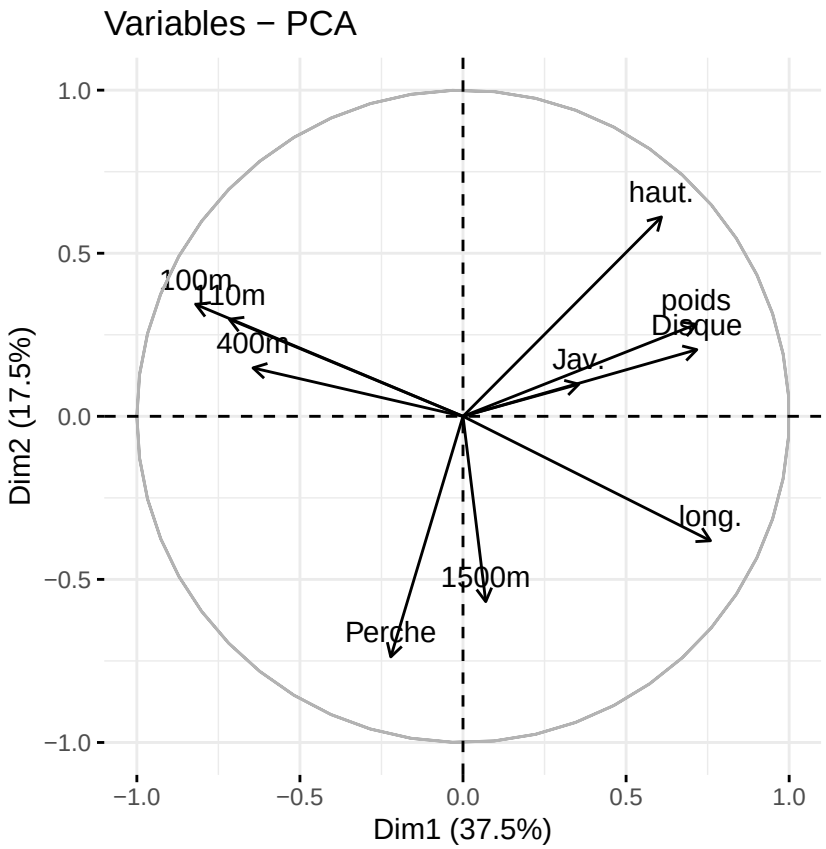
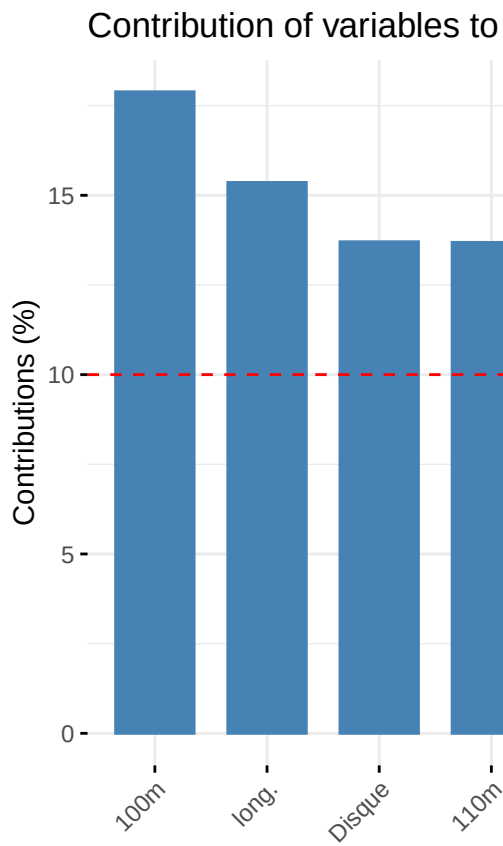


Figure 5: Cercle des corrélations

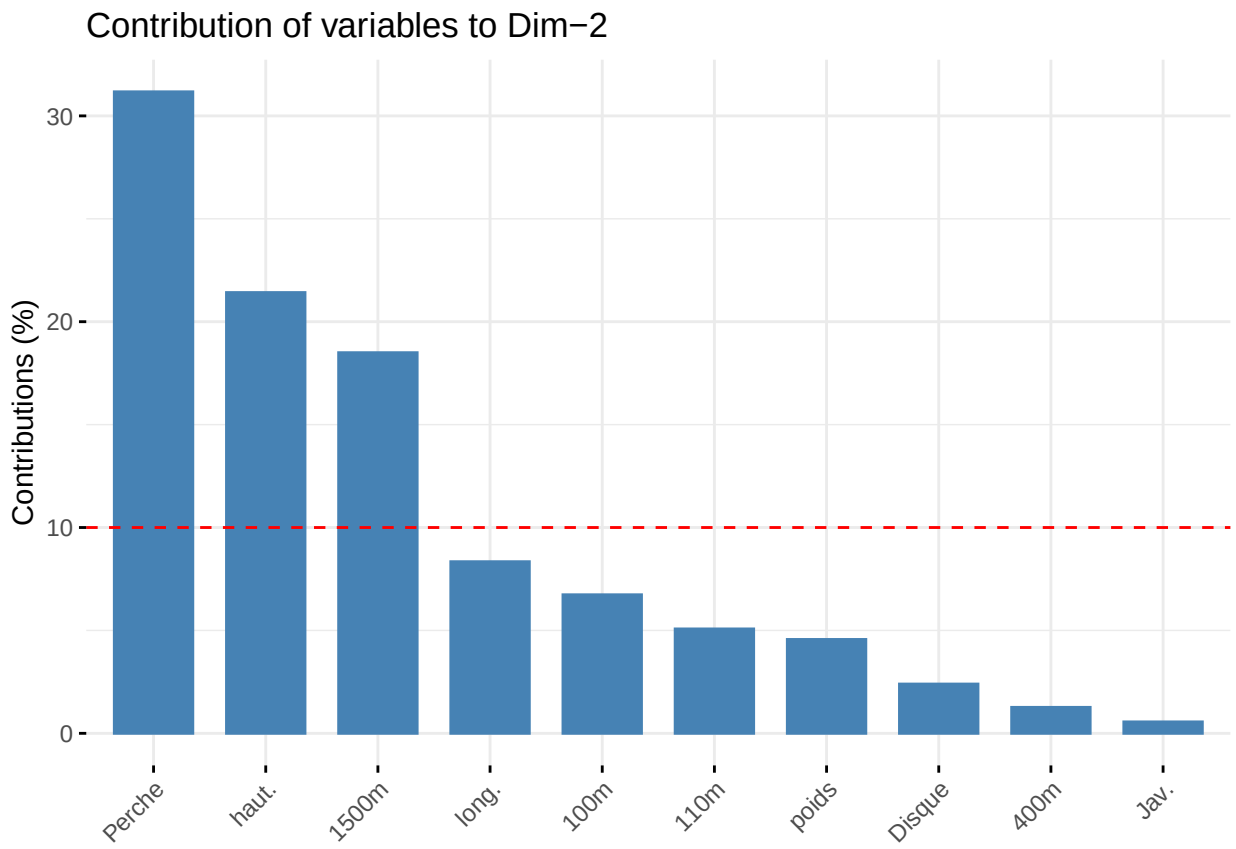


Comme on pouvait s'y attendre les épreuves de sprint sont bien corrélées entre elles ainsi que le poids et le disque.

Cependant il faut se méfier et bien avoir à l'esprit que nous ne voyons que les projections des variables sur un plan, attention aux interprétations rapides ! Deux flèches proches l'une de l'autre ne veut pas dire qu'elles sont fortement corrélées...



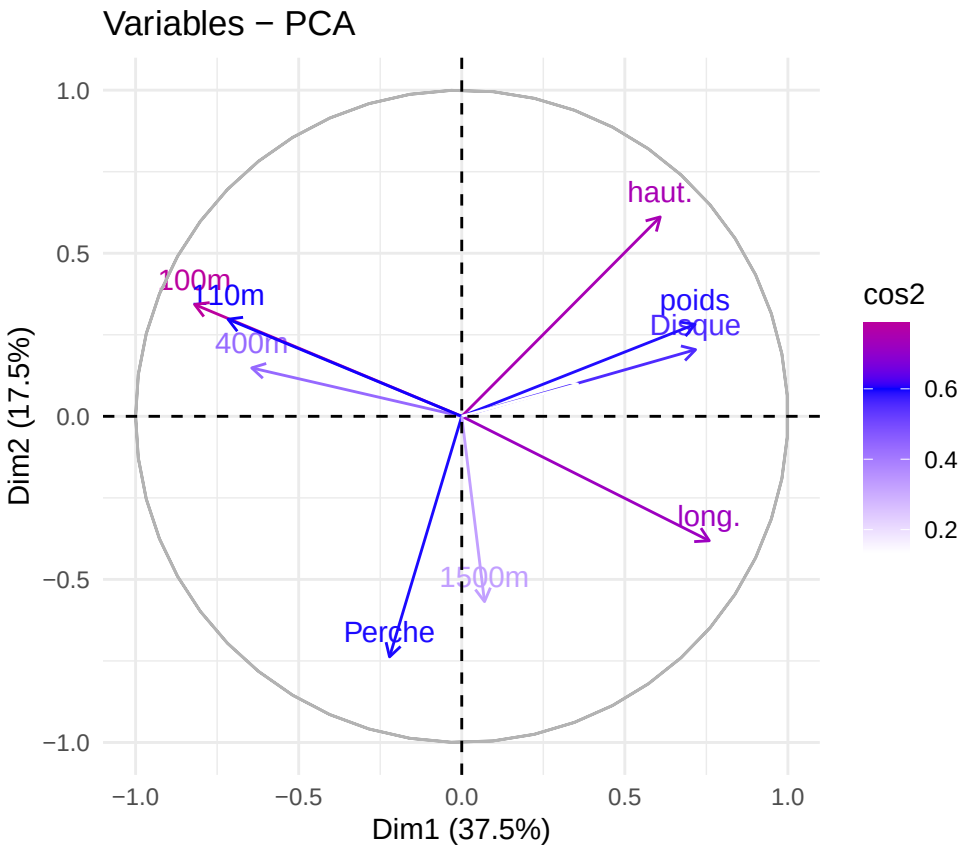
Il est important de connaître les contributions des variables sur les axes factoriels :



Ainsi que leurs corrélations avec les axes factoriels :

```
##              Dim.1      Dim.2
## 100m    -0.81895206  0.3427787
## long.    0.75889854 -0.3814931
## poids    0.71507829  0.2821167
## haut.    0.60849326  0.6113542
## 400m    -0.64384815  0.1484225
## 110m    -0.71642027  0.2975519
## Disque   0.71688812  0.2043979
## Perche -0.22141731 -0.7375479
## Jav.     0.35517566  0.0985309
## 1500m    0.06971223 -0.5681197
```

Ou de façon plus visuelle :



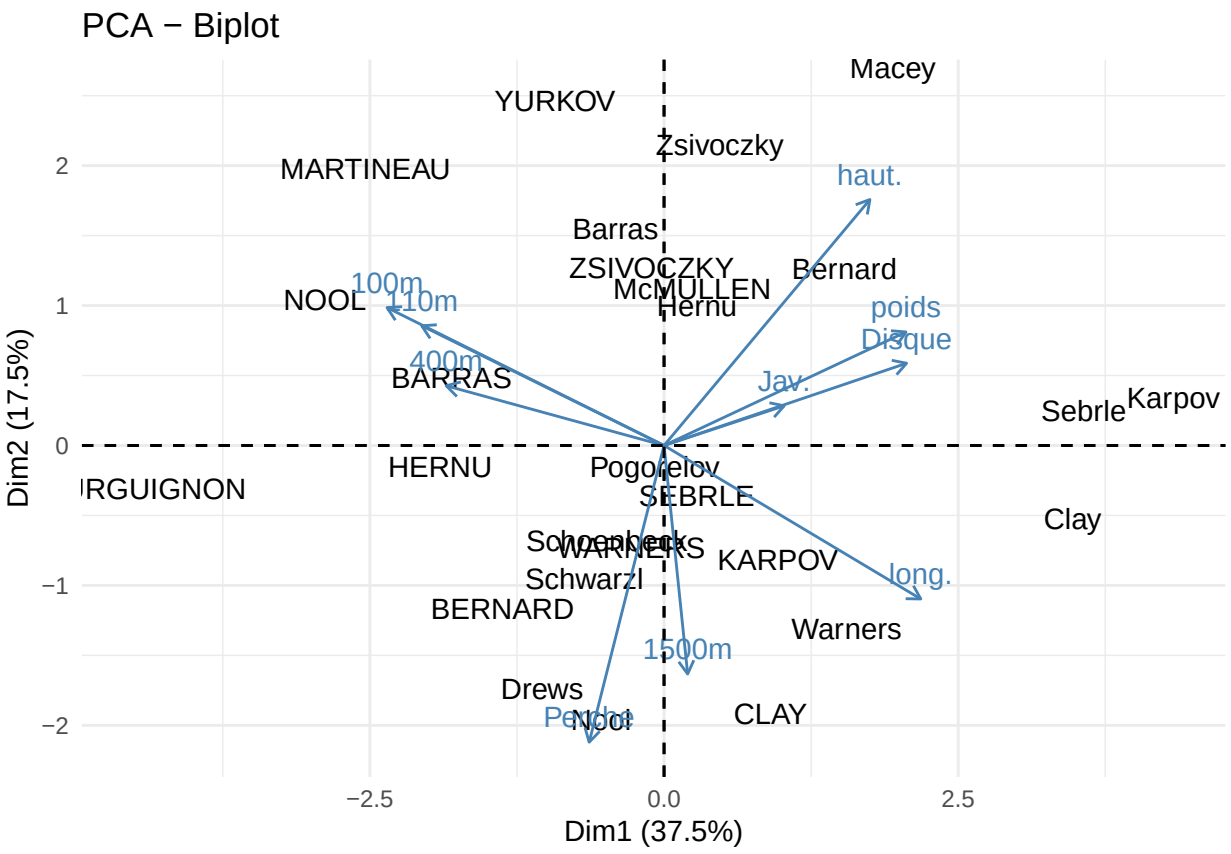
On voit tout de suite que les variables 100m, haut, long sont très bien représentées dans le plan factoriel.

Détails des cos² (carré des coordonnées):

```
##              Dim.1       Dim.2
## 100m   0.670682478 0.117497253
## long.  0.575926998 0.145537006
## poids  0.511336968 0.079589834
## haut.  0.370264042 0.373753996
## 400m   0.414540443 0.022029236
## 110m   0.513258007 0.088537114
## Disque 0.513928570 0.041778514
## Perche 0.049025625 0.543976848
## Jav.   0.126149749 0.009708339
## 1500m  0.004859795 0.322759992
```

Nous pouvons superposer sur le même graphique les deux graphes précédents afin de pouvoir interpréter simultanément les informations mais **attention les coordonnées des individus et des variables ne sont pas comparables** !

Il faut garder à l'esprit que **le graphe des variables est un cercle de corrélation**, les variables dont les vecteurs unitaires sont proches les uns des autres sont dites positivement corrélées, ce qui signifie que leur influence sur le positionnement des individus est similaire (là encore, ces proximités se reflètent dans les projections des variables sur le graphique des individus). Cependant, les variables éloignées les unes des autres seront définies comme étant corrélées négativement (ou anti-corrélées). Les variables qui ont un vecteur unité perpendiculaire ne sont pas corrélées



- **Un athlète qui est du même côté d'une variable aura une valeur élevée pour cette variable** ; (par exemple Nool à la perche a fait 5.40m est opposé à Zsivoczky qui a fait 4.70m, ces deux athlètes sont de plus assez bien représentés dans le plan factoriel)
- **Un athlète qui est à l'opposé d'une variable aura une faible valeur pour cette variable** (ne pas être étonné de trouver Zsivoczky à l'opposé de la variable 1500m ou Karpov et Clay qui sont à l'opposés de la variable 100m, ils sont très performants sur ces épreuves donc opposés à cette variable car leurs temps sont inférieurs aux autres athlètes !)

Pour conclure, que ce soit dans l'espace des individus ou l'espace des variables notre objectif a été le même, déterminer les axes factoriels de ces deux espaces, or on démontre que ce sont les mêmes !!

En effet les résultats concernant les variables (les colonnes) se déduisent de ceux obtenus par les individus (les lignes) il suffit de remplacer les lignes et les colonnes pour s'en convaincre (en d'autres termes, on remplace la matrice initiale par sa transposée). On a ce qu'on appelle **une relation de dualité entre les deux nuages**, celui des individus et celui des variables.

Application avec R

J'utiliserai principalement deux packages de R pour l'ACP :

- FactomineR développé par F. HUSSON de l'université de Rennes
- Factoextra développé par Alboukadel Kassambara

```
library(FactoMineR)
library(factoextra)
```

Nous allons étudier un jeu de données de 406 observations décrites par 9 variables.

Chargeons les données (récupérables ici : https://sjaubert.github.io/ACP/cars.csv):

```
cars<-read.csv2(file ="cars.csv",sep = ";",dec = ".")
```

Avant toute chose il faut examiner les données :

```
str(cars)
```

```
## 'data.frame':    406 obs. of  9 variables:
##  $ Car         : Factor w/ 308 levels "AMC Ambassador Brougham",..: 51 38 236 15 165 144 56 227 247 1
##  $ MPG         : num  23 11 23 14 18 11 8 8 8 11 ...
##  $ Cylinders   : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ Displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ Horsepower  : num  17 35 29 29 24 42 47 46 48 40 ...
##  $ Weight      : num  250 268 244 243 247 324 325 321 333 281 ...
##  $ Acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ Model       : Factor w/ 13 levels "70,00","71,00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Origin      : Factor w/ 3 levels "Europe","Japan",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
head(cars)
```

```
##                        Car MPG Cylinders Displacement Horsepower Weight
## 1 Chevrolet Chevelle Malibu  23         5          307         17    250
## 2         Buick Skylark 320  11         5          350         35    268
## 3        Plymouth Satellite  23         5          318         29    244
## 4             AMC Rebel SST  14         5          304         29    243
## 5               Ford Torino  18         5          302         24    247
## 6          Ford Galaxie 500  11         5          429         42    324
##   Acceleration Model Origin
## 1         12.0 70,00     US
## 2         11.5 70,00     US
## 3         11.0 70,00     US
## 4         12.0 70,00     US
## 5         10.5 70,00     US
## 6         10.0 70,00     US
```

ça mérite quelques explications :

- mpg: Consommation en Miles/(US) gallon
- cylinders : Nombre de cylindres
- displacement : La cylindrée
- Horsepower : la puissance
- Weight : l'unité est 1000 lbs
- Acceleration : Temps pour effectuer un 1/4 de mile
- Model : Année de sortie du véhicule
- Origin : Pays d'origine

Pour simplifier renommons quelques variables :

```
library(dplyr)
cars <-cars %>%
  rename(
    Nb_Cyl = Cylinders,
    Cylindrée = Displacement,
    Puissance = Horsepower)
str(cars)
```

```
## 'data.frame':    406 obs. of  9 variables:
##  $ Car         : Factor w/ 308 levels "AMC Ambassador Brougham",..: 51 38 236 15 165 144 56 227 247 1
##  $ MPG         : num  23 11 23 14 18 11 8 8 8 11 ...
##  $ Nb_Cyl      : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ Cylindrée   : num  307 350 318 304 302 429 454 440 455 390 ...
##  $ Puissance   : num  17 35 29 29 24 42 47 46 48 40 ...
##  $ Weight      : num  250 268 244 243 247 324 325 321 333 281 ...
##  $ Acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ Model       : Factor w/ 13 levels "70,00","71,00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Origin      : Factor w/ 3 levels "Europe","Japan",..: 3 3 3 3 3 3 3 3 3 3 ...
```

Lançons l'ACP sur les variables actives (les colonnes c(1,8,9) étant qualitatives)

```
res_ACP<-PCA(cars,scale.unit = T,quali.sup = c(1,8,9),graph = F)
```

Les valeurs propres sont obtenues avec :

```
res_ACP$eig
```

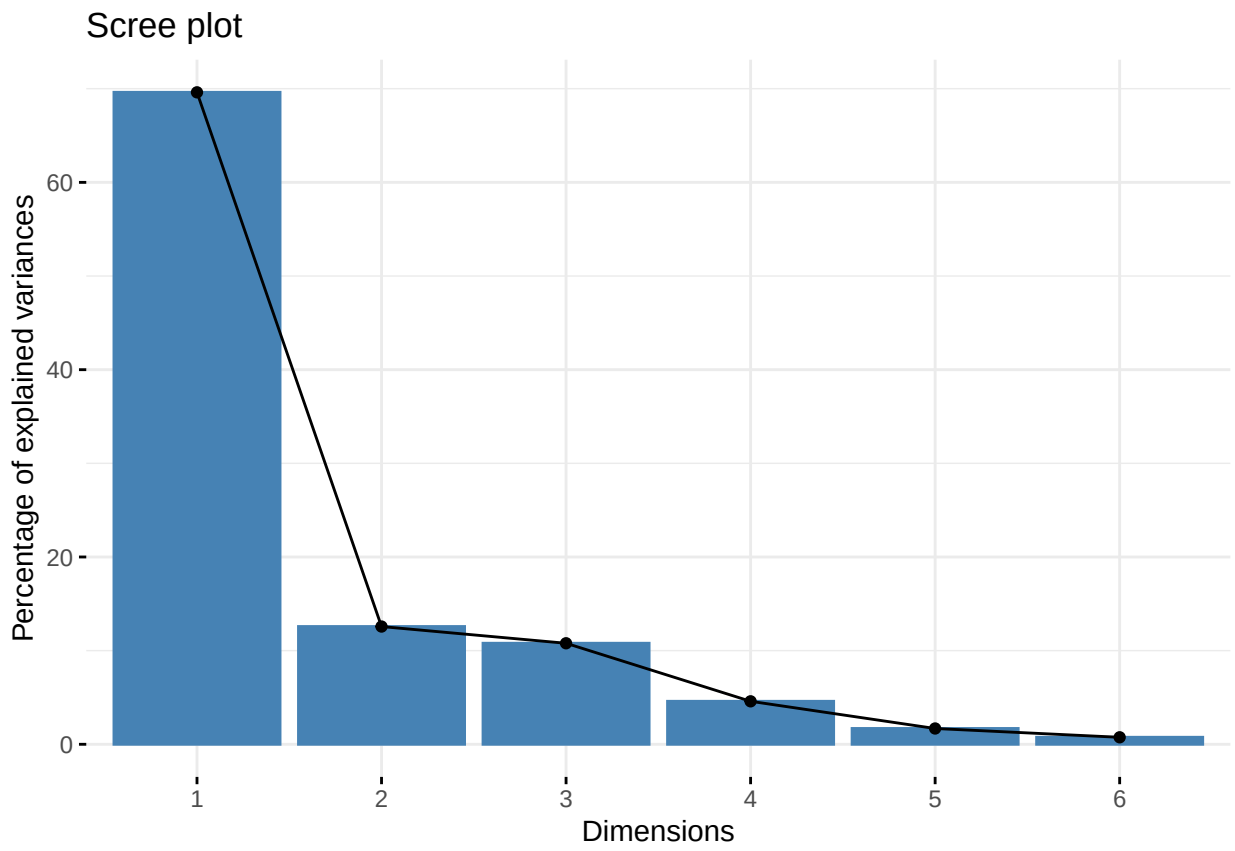
```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.17671948              69.611991                           69.61199
## comp 2 0.75440177              12.573363                           82.18535
## comp 3 0.64728993              10.788165                           92.97352
## comp 4 0.27545370               4.590895                           97.56441
## comp 5 0.10125709               1.687618                           99.25203
## comp 6 0.04487802               0.747967                          100.00000
```

La proportion de variation expliquée par chaque valeur propre est donnée par la 2ème colonne, le pourcentage cumulé est donné par la 3ème colonne.

Ici les 3 premières composantes expliquent 93% des variations

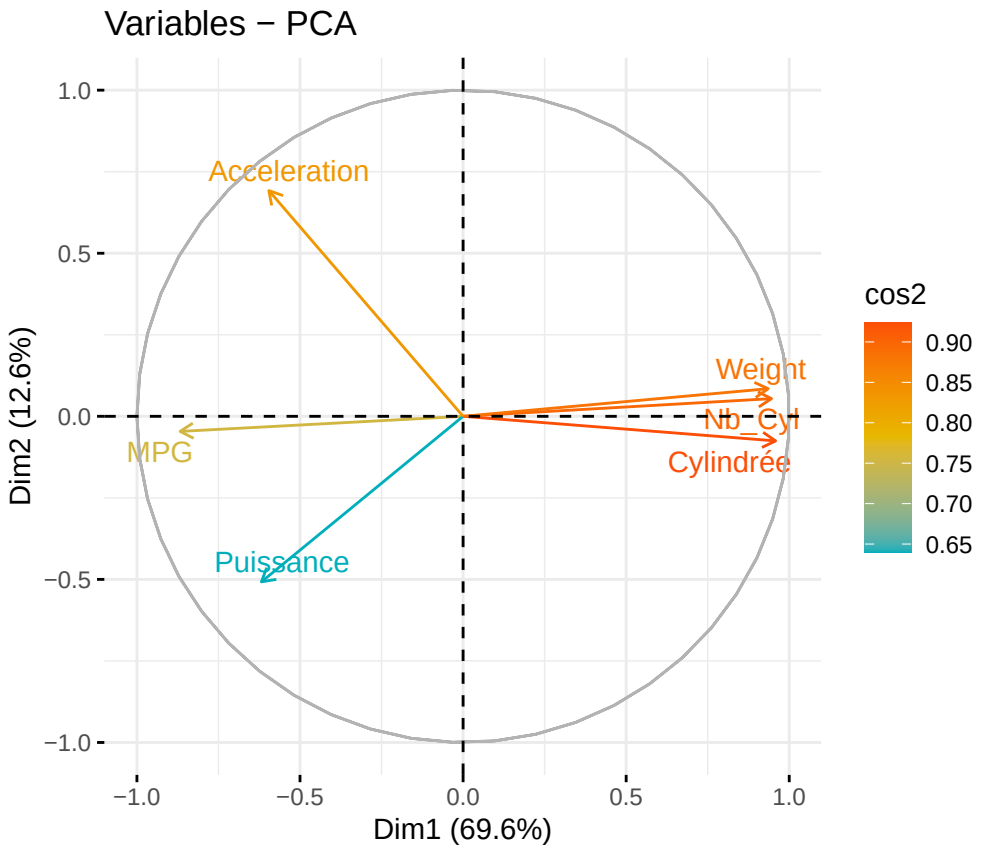
Nous pouvons le visualiser par l'éboulis des valeurs propres :

```
fviz_eig(res_ACP)
```



visualisation des variables :

```
fviz_pca_var(res_ACP,repel = TRUE,col.var = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
```



Ce n'est pas une surprise, *Weight*, *Nb_cyl* et la *Cylindrée* sont très bien representées et corrélées sur l'axe 1 et la consommation, *MPG* qui est aussi bien représentée sur cet axe et anti-corrélée à ces variables.

Leurs coordonnées dans le plan factoriel sont :

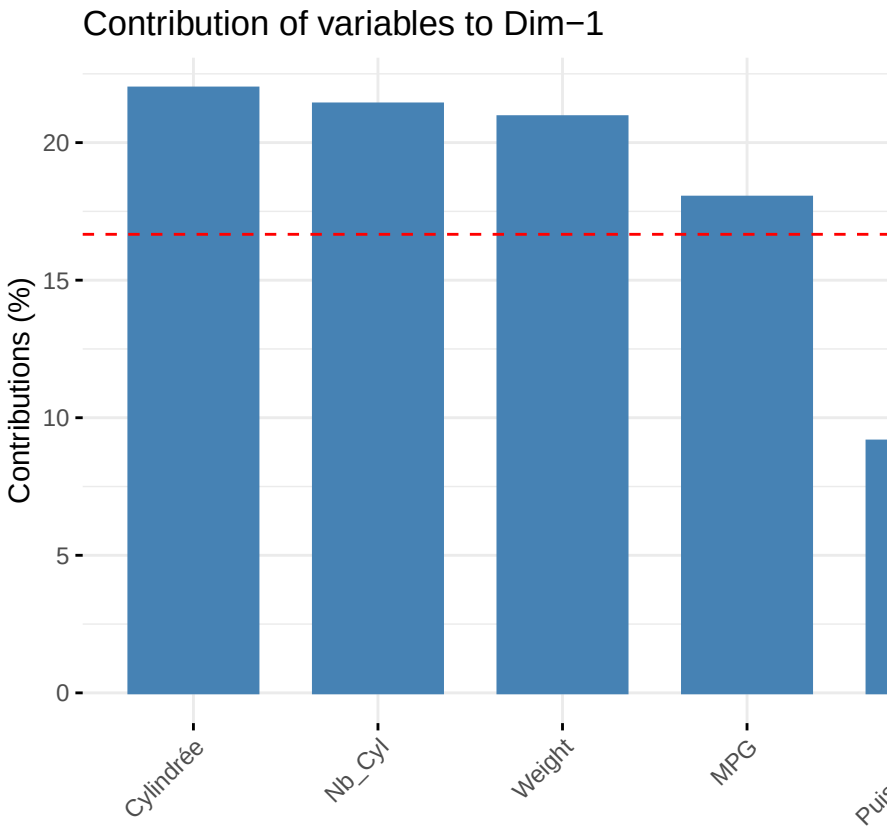
```
res_ACP$var$coord[,1:2]
```

```
##                    Dim.1       Dim.2
## MPG          -0.8675245 -0.04629692
## Nb_Cyl        0.9455754  0.05421210
## Cylindrée     0.9582673 -0.07538234
## Puissance    -0.6183437 -0.50750888
## Weight        0.9352739  0.08459446
## Acceleration -0.5955212  0.69203716
```

et les $\cos^2$:

```
res_ACP$var$cos2[,1:2]
```

```
##                   Dim.1       Dim.2
## MPG          0.7525988 0.002143405
## Nb_Cyl       0.8941129 0.002938952
## Cylindrée    0.9182762 0.005682498
## Puissance    0.3823489 0.257565268
## Weight       0.8747372 0.007156222
## Acceleration 0.3546455 0.478915430
```

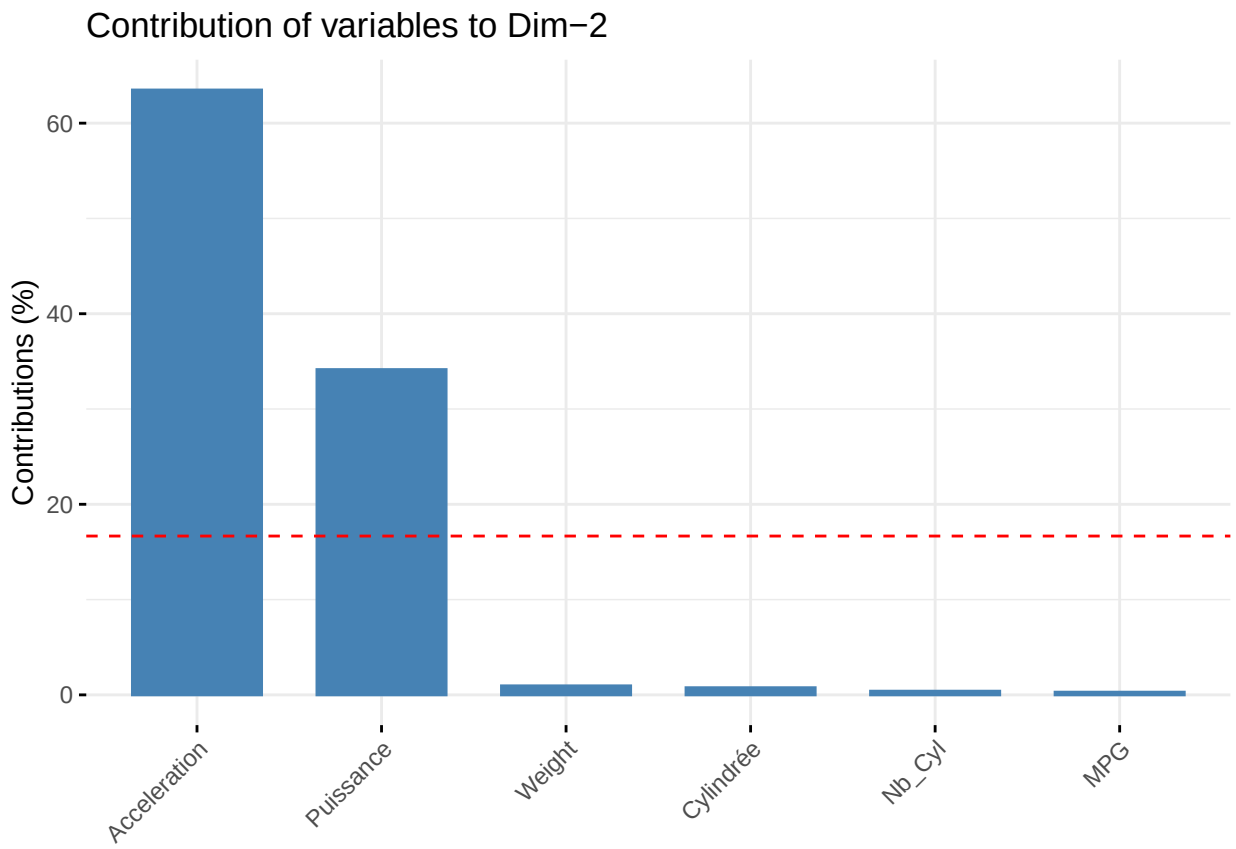


Graphiquement, leurs contributions sur les axes

Et sur l'axe 2

```
fviz_pca_contrib(res_ACP,choice = "var",axes = 2)
```

```
## Warning in fviz_pca_contrib(res_ACP, choice = "var", axes = 2): The function
## fviz_pca_contrib() is deprecated. Please use the function fviz_contrib() which
## can handle outputs of PCA, CA and MCA functions.
```



Il est possible d'extraire plus simplement les résultats des variables avec l'instruction :

```
var<-get_pca_var(res_ACP)
```

puis on appelle simplement les variables associées à l'objet *var* :

```
var$coord
```

```
##                  Dim.1       Dim.2       Dim.3      Dim.4         Dim.5
## MPG          -0.8675245 -0.04629692 -0.22750853 0.43380526  0.0721135408
## Nb_Cyl        0.9455754  0.05421210  0.08091566 0.21920421 -0.1966726541
## Cylindrée     0.9582673 -0.07538234  0.12734447 0.17717138 -0.0005445788
## Puissance    -0.6183437 -0.50750888  0.59826070 0.03943591 -0.0205695650
## Weight        0.9352739  0.08459446  0.22887193 0.05896650  0.2374450478
## Acceleration -0.5955212  0.69203716  0.40307271 0.05286097 -0.0239382397
```

```
var$cor
```

```
##                  Dim.1       Dim.2       Dim.3      Dim.4         Dim.5
## MPG          -0.8675245 -0.04629692 -0.22750853 0.43380526  0.0721135408
## Nb_Cyl        0.9455754  0.05421210  0.08091566 0.21920421 -0.1966726541
## Cylindrée     0.9582673 -0.07538234  0.12734447 0.17717138 -0.0005445788
## Puissance    -0.6183437 -0.50750888  0.59826070 0.03943591 -0.0205695650
## Weight        0.9352739  0.08459446  0.22887193 0.05896650  0.2374450478
## Acceleration -0.5955212  0.69203716  0.40307271 0.05286097 -0.0239382397
```

```
var$cos2
```

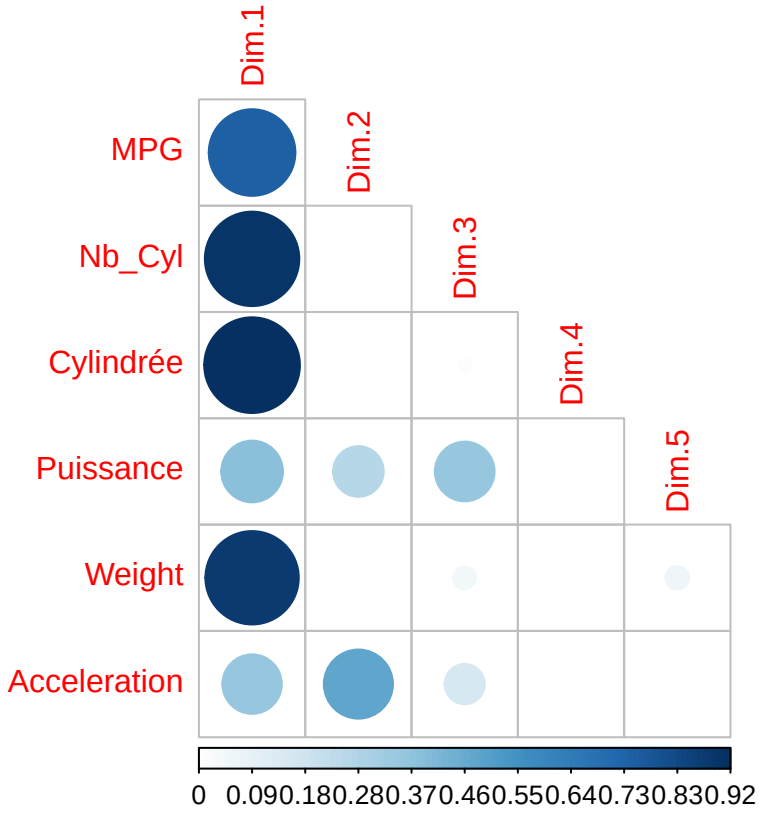
```
##                  Dim.1       Dim.2       Dim.3       Dim.4        Dim.5
## MPG          0.7525988 0.002143405 0.051760132 0.188187000 5.200363e-03
## Nb_Cyl       0.8941129 0.002938952 0.006547344 0.048050485 3.868013e-02
## Cylindrée    0.9182762 0.005682498 0.016216614 0.031389698 2.965661e-07
## Puissance    0.3823489 0.257565268 0.357915864 0.001555191 4.231070e-04
## Weight       0.8747372 0.007156222 0.052382360 0.003477048 5.638015e-02
## Acceleration 0.3546455 0.478915430 0.162467613 0.002794282 5.730393e-04
```

```
var$contrib
```

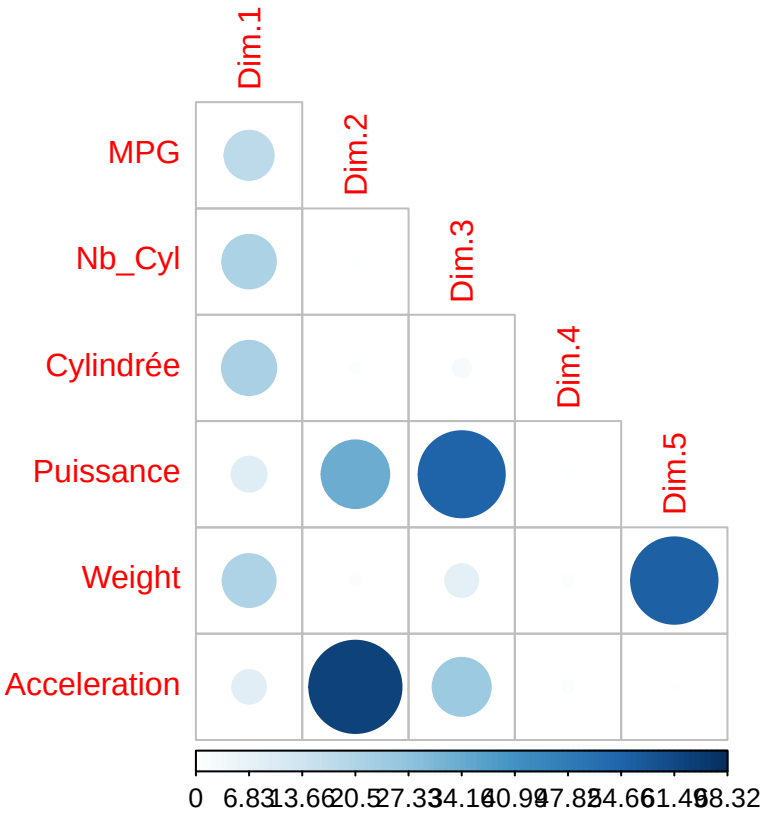
```
##                  Dim.1      Dim.2     Dim.3      Dim.4        Dim.5
## MPG          18.018897  0.2841198  7.996437 68.3189217 5.135801e+00
## Nb_Cyl       21.407060  0.3895738  1.011501 17.4441237 3.819993e+01
## Cylindrée    21.985585  0.7532455  2.505309 11.3956347 2.928842e-04
## Puissance     9.154287 34.1416573 55.294521  0.5645926 4.178542e-01
## Weight       20.943165  0.9485956  8.092565  1.2622983 5.568020e+01
## Acceleration  8.491006 63.4828080 25.099667  1.0144290 5.659251e-01
```

Avec la librairie *corrplot* il est aussi possible de représenter graphiquement les $\cos^2$ et les contributions des variables

```
library("corrplot")
corrplot(var$cos2, is.corr=FALSE,type = "lower")
```

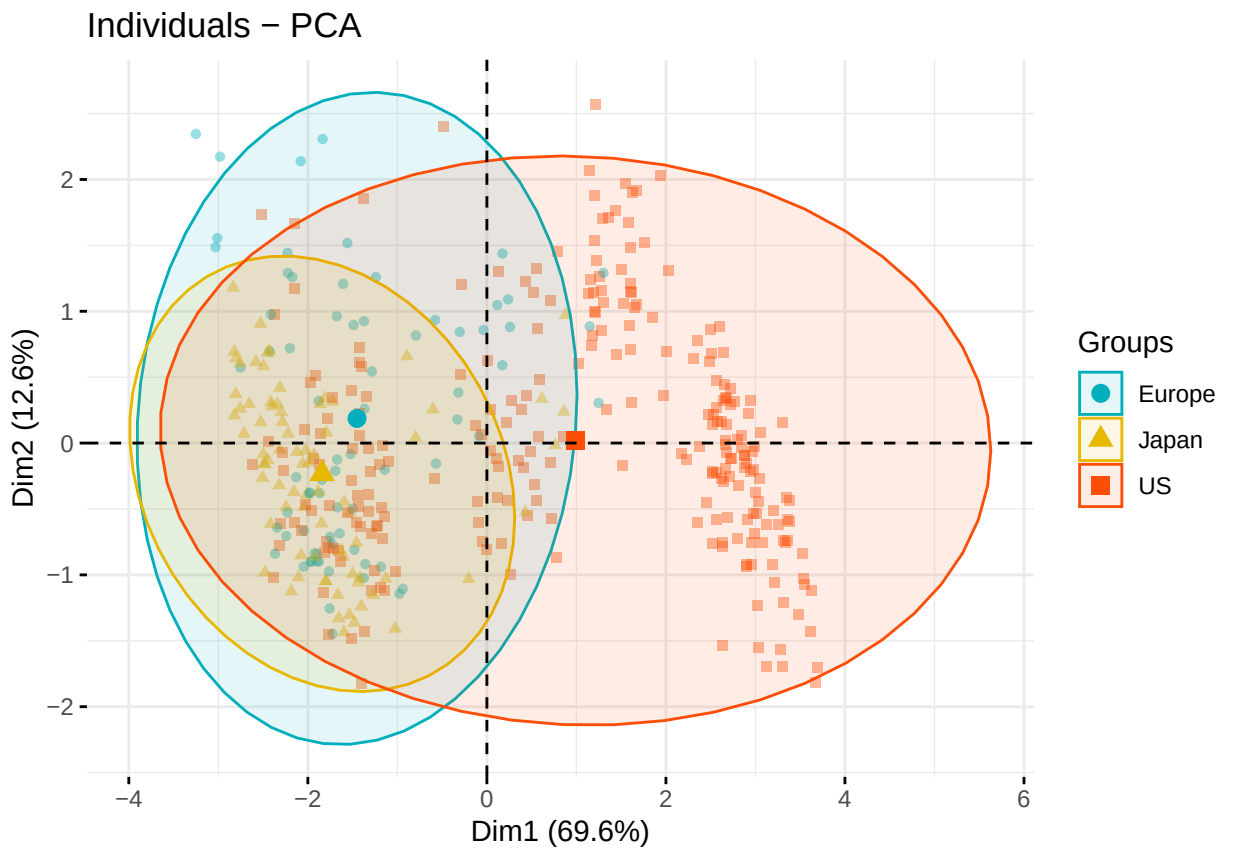


```
corrplot(var$contrib,is.corr = FALSE,type = "lower")
```



Représentons les individus :

```
fviz_pca_ind(res_ACP,label="none",habillage = 9,addEllipses = T,ellipse.level=0.95,alpha.ind = 0.4,pale
```



Les individus “similaires” sont regroupés dans les ellipses sur le graphique. Comme pour les variables, on peut obtenir tous les résultats concernant les individus par :

```
indiv<-get_pca_ind(res_ACP)
```

Annexe

Soit X la matrice de nos données :

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & & \vdots & x_{2,p} \\ \vdots & \vdots & x_{i,j} & \vdots \\ & & \vdots & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

, Cette matrice donne en $x_{i,j}$ la valeur de la jème variable pour le ième individu

Déterminer les écarts entre les individus soulève un problème, celui des unités choisies. Il ne faut pas que la distance entre deux points dépende des unités. La solution sera de normaliser les données, on remplacera les $x_{i,j}$ par :

$$x'_{i,j} = \frac{x_{i,j} - \bar{v}_j}{s_j \sqrt{n}}$$

où $\bar{v}_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}$ moyenne de la variable j et s_j est l'écart-type de la variable j

Dans la suite on considérera que les données sont normalisées et on conservera la notation $x_{i,j}$ (dans ce cas la variance de chaque variable égale 1)

Deux individus $e_i = (x_{i,1}, \ldots, x_{i,p})$ et $e_j = (x_{j,1}, \ldots, x_{j,p})$ sont très proches (ou homogènes) si les p coordonnées qui les décrivent sont très voisines.

Le coefficient de corrélation

Références

- Analyse de données avec R (Data Analysis with R) F. Husson, S. Lê & J. (Presses Universitaire de Rennes)
- Probabilités, analyse des données et statistique G. SAPORTA (Technip)
- Statistiques avec R PA Cornillon, A. Guyader, F. Husson, N. Jegou, J. Josse, M. Kloareg, E. Matzner-Lober & L. Rouvière 2012 (3rd edition Presses Universitaires de Rennes)