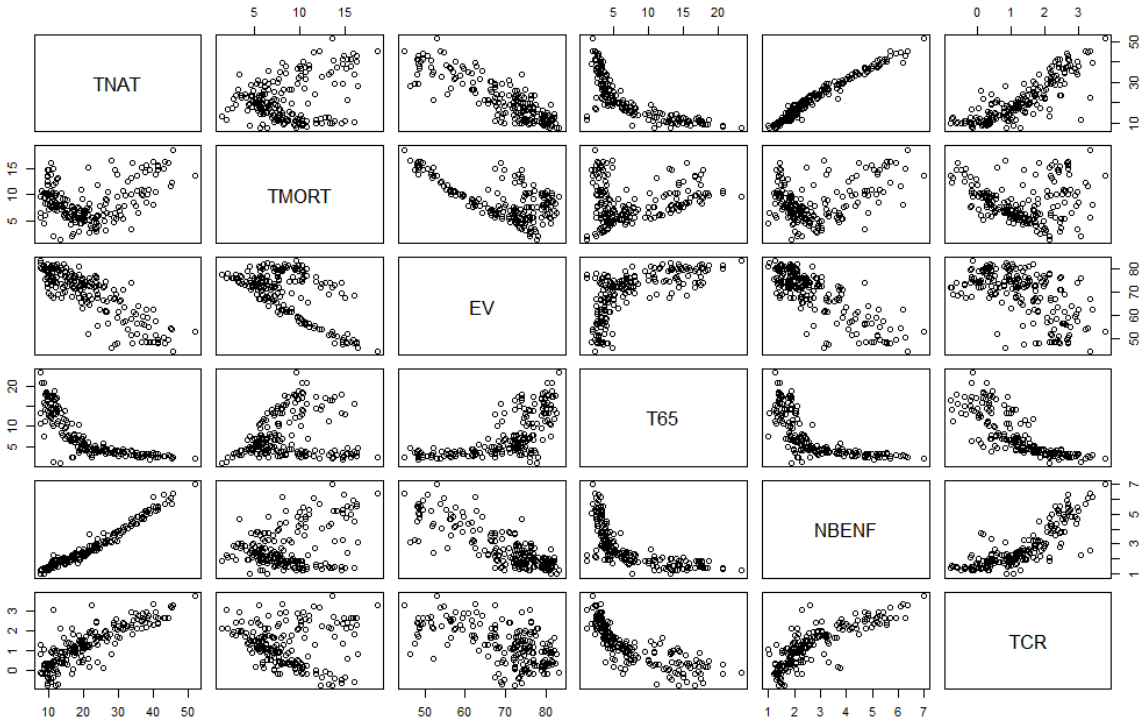
TP : Analyse en Composantes Principales

Exercice 1 — *Données : EspVieACPData.txt*

L'objectif de cet exercice est d'apprendre à utiliser les packages R permettant de faire une ACP : `FactoMineR, ade4, explor`.

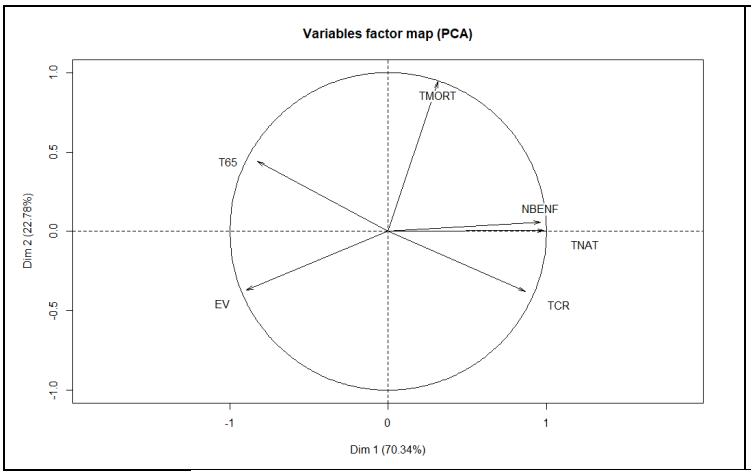
1) Installer et charger ces trois packages dans votre session de travail.
2) Lire le jeu de données utilisé en illustration du cours : EspVieACPData.txt
3) Préparation des données
   - Représenter les nuages de points des données. Y-a-t'il des individus atypiques ? Quelles sont les variables corrélées ?
   - Centrer et réduire les variables.

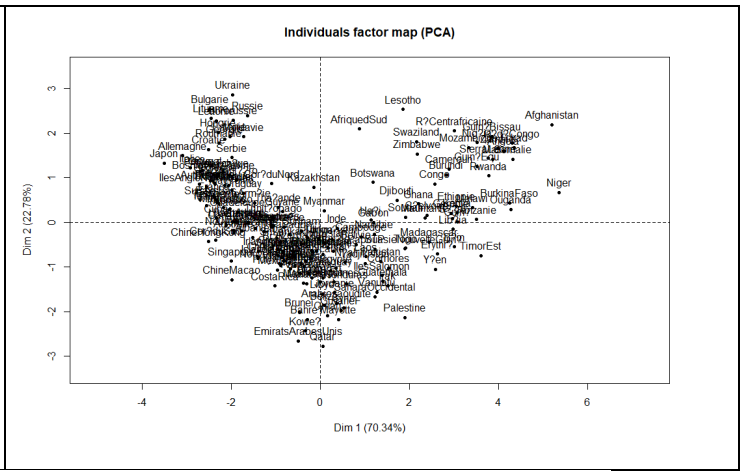
```
tab=read.table("EspVieACPData.txt",header=T)
pairs(tab[,1:6])
tab[,1:6]=scale(tab[,1:6])
```



4) Faire une ACP avec FactoMineR
   - Afficher l'aide R concernant la fonction PCA
   - 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.

```
res.PCA=PCA(tab[,1:6],scale.unit=T,ncp=6,graph=T,axes=c(1,2))
```



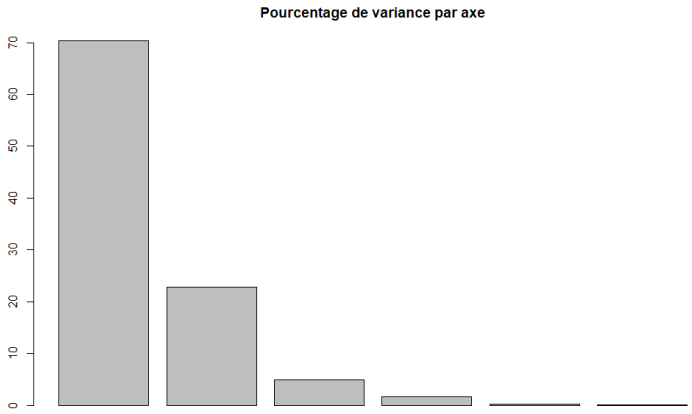


- Afficher le diagramme des valeurs propres

L'attribut $eig donne un tableau avec les valeurs propres, le pourcentage de variance correspondant et le pourcentage cumulé

```
res.PCA$eig
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 4.220696272              70.3449379                           70.34494
comp 2 1.366814683              22.7802447                           93.12518
comp 3 0.292394892               4.8732482                           97.99843
comp 4 0.097342316               1.6223719                           99.62080
comp 5 0.013669905               0.2278318                           99.84863
comp 6 0.009081932               0.1513655                          100.00000
```

```
barplot(res.PCA$eig[,2],main="Pourcentage de variance par axe")
```



- 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 ?

```
res.PCA$var$cos2
          Dim.1        Dim.2        Dim.3        Dim.4        Dim.5        Dim.6
TNAT  0.9712846 0.0000335445 0.0084609538 1.376190e-02 3.021247e-03 3.437726e-03
TMORT 0.0996249 0.8911353397 0.0001426413 5.988691e-03 2.160944e-03 9.474837e-04
EV    0.7973222 0.1360834253 0.0582323428 2.860889e-03 3.321286e-03 2.179845e-03
T65   0.6789212 0.1939980998 0.1235362609 2.078493e-05 3.070905e-03 4.527491e-04
NBENF 0.9236236 0.0034352851 0.0449950663 2.378890e-02 2.093170e-03 2.063993e-03
TCR   0.7499197 0.1421289886 0.0570276265 5.092115e-02 2.353008e-06 1.361461e-07

sum(res.PCA$var$cos2[1,])
[1] 1
```

Cela signifie que le cos² peut s'interpréter comme un pourcentage. Par exemple, TMORT est représentée (projetée) à 89,11% sur l'axe 2.

```
res.PCA$var$contrib
          Dim.1       Dim.2       Dim.3       Dim.4       Dim.5       Dim.6
TNAT  23.01243  0.00245421  2.89367360 14.13763663 22.10144786 37.852361234
TMORT  2.36039 65.19796361  0.04878378  6.15219725 15.80804312 10.432622074
EV    18.89077  9.95624550 19.91564986  2.93899826 24.29633645 24.001995514
T65   16.08553 14.19344570 42.24980136  0.02135241 22.46471156  4.985162874
NBENF 21.88320  0.25133511 15.38845842 24.43839504 15.31224796 22.726359215
TCR   17.76768 10.39855588 19.50363299 52.31142041  0.01721306  0.001499087
```

- Afficher les résultats sur les individus

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

Comme pour les variables, le cos² donne le pourcentage d'un individu par axe. On voit donc que le Bangladesh est projeté à 46,1% sur l'axe 2 et 49.8% sur l'axe 3.

```
res.PCA$ind$cos2
…
Bangladesh  4.181514e-03 4.611016e-01 4.977998e-01 1.962746e-02 1.618735e-02
```

Quelle est la contribution moyenne d'un pays à la construction des axes ? Y-at'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.

Il y a 196 pays donc la contribution moyenne est 1/196.

```
sort(res.PCA$ind$contrib[,1],decreasing=T)
      Niger           Afghanistan                 Tchad               Somalie
3.478377e+00          3.275963e+00          2.292943e+00          2.259864e+00
   R?d?Congo              Ouganda            BurkinaFaso                Angola
2.250734e+00          2.220869e+00          2.179999e+00          2.051130e+00
     Zambie                 Mali           Guin?Bissau                Rwanda
…
```

Certains pays ont une contribution bien plus importante que la contribution moyenne. Ils pourraient être atypiques et entrainer des perturbations dans les résultats. On supprime donc celui qui a la plus grande contribution et on regarde si cela change la construction des axes. NB. Le Niger est la 127ème ligne.

```
res=PCA(tab[,1:6],ind.sup=c(127),scale.unit=T,ncp=6,graph=T,axes=c(1,2))
```

Le fait de mettre le Niger en individu supplémentaire, le retire des calculs lors de la construction des axes. Il est ensuite ajouté mais uniquement pour la représentation graphique.
Cela ne change rien donc pas d'individu atypique.

- Ajouter la variable Continent sur le graphique des individus

```
PCA(tab,quali.sup=7,scale.unit=T,ncp=6,graph=T,axes=c(1,2))
```

Attention à remettre la colonne 7 dans le tableau !

- Visualisation avec explor

```
explor(res.PCA)
```

Le package explor est une application Shiny qui permet d'avoir une représentation dynamique et interactive des résultats. Très pratique.
Il peut y avoir un problème de Proxy avec le package explor sous Windows. Il faut aller dans les paramètres de votre ordinateur, rubrique Réseaux et Internet, puis Proxy et désactiver le Proxy.

- Faire une ACP avec le package ade4

```
res=dudi.pca(tab[,1:6])
```

Exercice 1 — Données : DecathlonData.txt

- Les colonnes 1 à 12 sont des variables continues:
  - 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 (Jeux Olympiques de 2004 ou Décastar 2004).
- Les lignes désignent les athlètes.

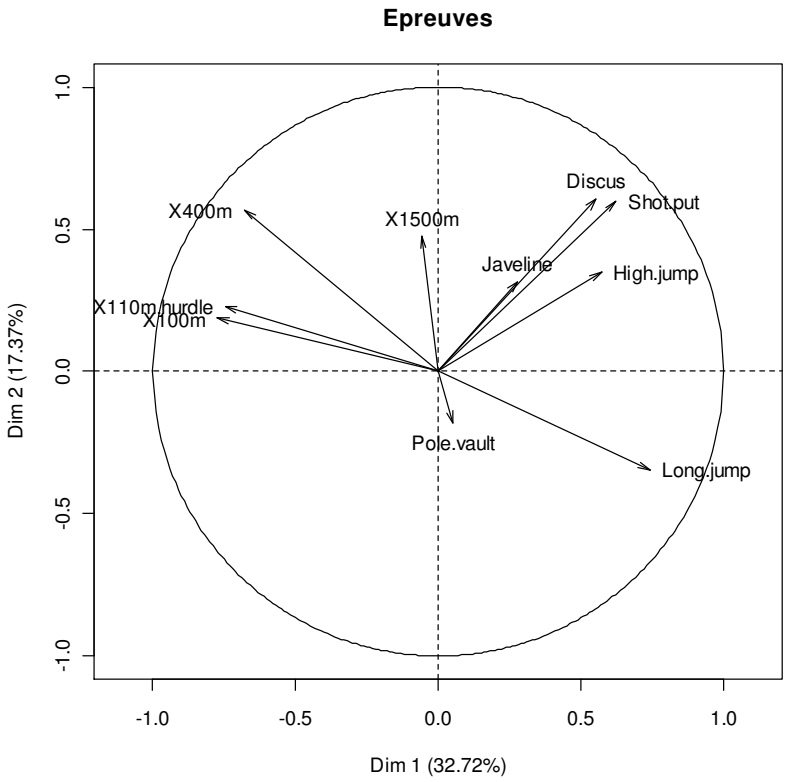
Nous allons faire une ACP sur les colonnes de 1 à 10.

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 ?

Réponse : Sur les graphiques, on voit que la 1ère composante contient 32,7% et la deuxième 17,4% donc en tout 50,1% de l'inertie totale. Si on regarde les résultats des valeurs propres, on voit qu'il faut 4 composantes principales pour atteindre plus de 70% de l'inertie totale.

```
eigenvalue percentage of variance cumulative percentage of variance
comp 1   3.2719055              32.719055                          32.71906
comp 2   1.7371310              17.371310                          50.09037
comp 3   1.4049167              14.049167                          64.13953
comp 4   1.0568504              10.568504                          74.70804
comp 5   0.6847735               6.847735                          81.55577
comp 6   0.5992687               5.992687                          87.54846
comp 7   0.4512353               4.512353                          92.06081
comp 8   0.3968766               3.968766                          96.02958
comp 9   0.2148149               2.148149                          98.17773
comp 10  0.1822275               1.822275                         100.00000
```

2) Etude des variables.



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

Réponse : Il faut faire attention car pour ces variables, une petite valeur correspond à un score élevé.

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

Réponse : Les deux variables très corrélées (cos ~ -1). Cela signifie qu'un athlète qui réalise un temps faible au 100m peut aussi sauter loin.

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

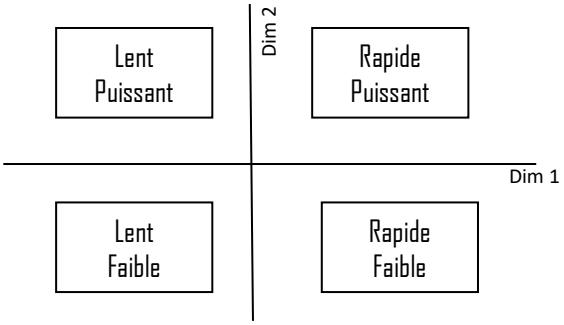
Réponse : On distingue 2 groupes. Le 1er constitué par les variables « X100m », « X400m », « X110m.hurdle », « long.jump » représente des performances de vitesse. Le 2ème constitué des variables « Discus », « Shot.put », « High.jump » représente des performances de force. Ces deux groupes sont non corrélés (cos~0), ce qui signifie que force et vitesse ne sont pas liées.

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 ?

Réponse : Les variables qui contribuent significativement à la première composante principale sont :

Dim 1 (96%)		Dim 2 (94%)	
X100	18%	Discus	21%
X110.hurdle	17%	Shot.put	21%
Long.jump	17%	X400m	19%
X400m	14%	X1500m	3%
Shot.put	11%	High.jump	7%
High.jump	10%	Long.jump	7%
Discus	9%	Javeline	6%

Rq. Les variables Javeline, Pole.vault, X1500m ou High.jump contribuent très peu aux deux premières composantes mais contribuent aux composantes suivantes.
Le premier axe semble plutôt représenter des épreuves de vitesse alors que le deuxième est identifié (moins clairement) à des épreuves de force. On peut en déduire la cartographie suivante

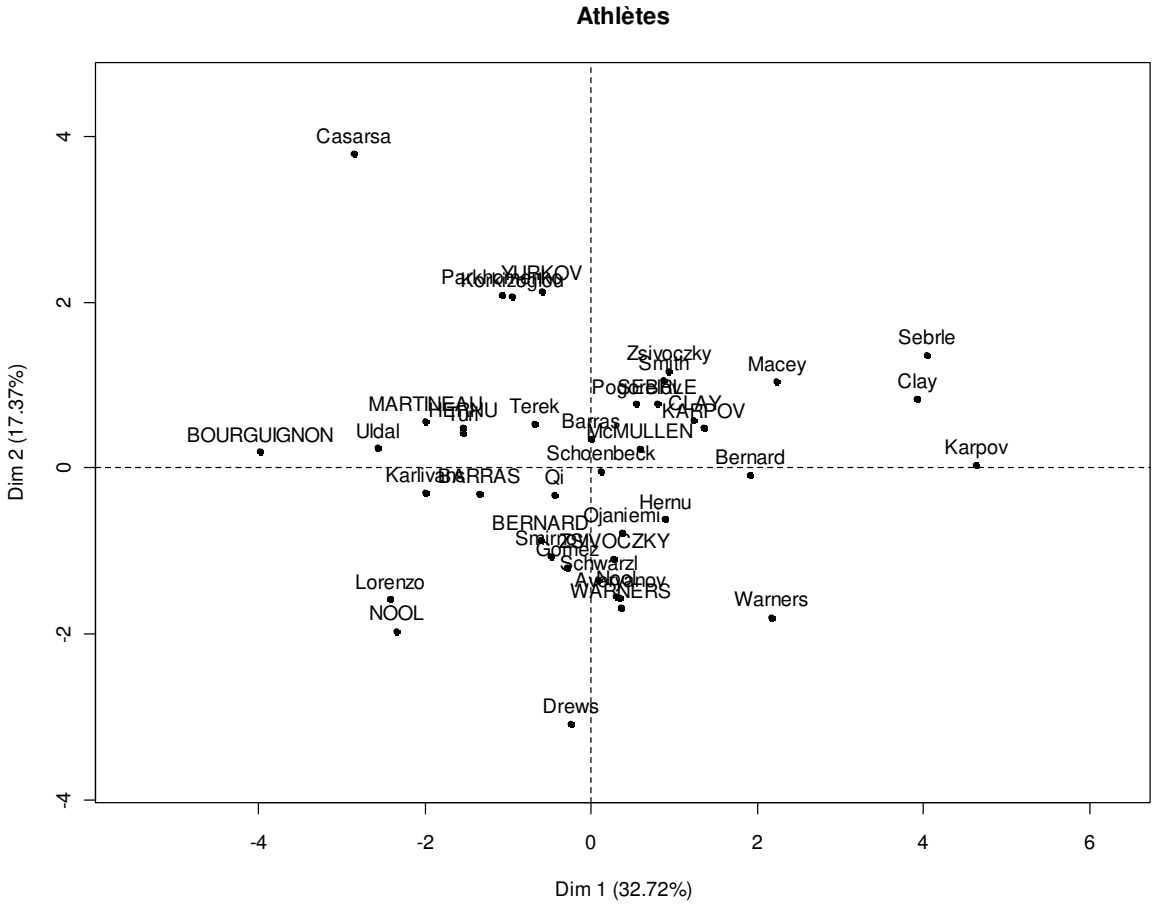



```
$contrib
                 Dim.1     Dim.2       Dim.3       Dim.4       Dim.5
X100m        18.34376957  2.016090  2.42049891  0.13532858 13.3361842
Long.jump    16.82246707  6.868559  2.36319121  0.98030118  0.1964560
Shot.put     11.84353954 20.606785  0.03890276  3.43711486  1.8041739
High.jump     9.99788710  7.063694  4.79362526  1.73967752 45.0533061
X400m        14.11622887 18.666374  1.23027094  0.08124195  1.1229714
X110m.hurdle 17.02011495  3.013382  0.61083225  8.00327927  3.9431102
Discus        9.32848615 21.162245  0.13131711  6.38020830  1.6047240
Pole.vault    0.07745541  1.872547 34.06090024 28.78266727 15.8991470
Javeline      2.34696326  5.784369 10.80714169 48.00480246 13.5962699
X1500m        0.10308808 12.945954 43.54331962  2.45537861  3.4436573

$Dim.1
$Dim.1$quanti
             correlation      p.value
Long.jump      0.7418997 2.849886e-08
Shot.put       0.6225026 1.388321e-05
High.jump      0.5719453 9.362285e-05
Discus         0.5524665 1.802220e-04
X400m         -0.6796099 1.028175e-06
X110m.hurdle  -0.7462453 2.136962e-08
X100m         -0.7747198 2.778467e-09
```

```
$Dim.2
$Dim.2$quanti
           correlation      p.value
Discus       0.6063134 2.650745e-05
Shot.put     0.5983033 3.603567e-05
X400m        0.5694378 1.020941e-04
X1500m       0.4742238 1.734405e-03
High.jump    0.3502936 2.475025e-02
Javeline     0.3169891 4.344974e-02
Long.jump   -0.3454213 2.696969e-02
```

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 ?

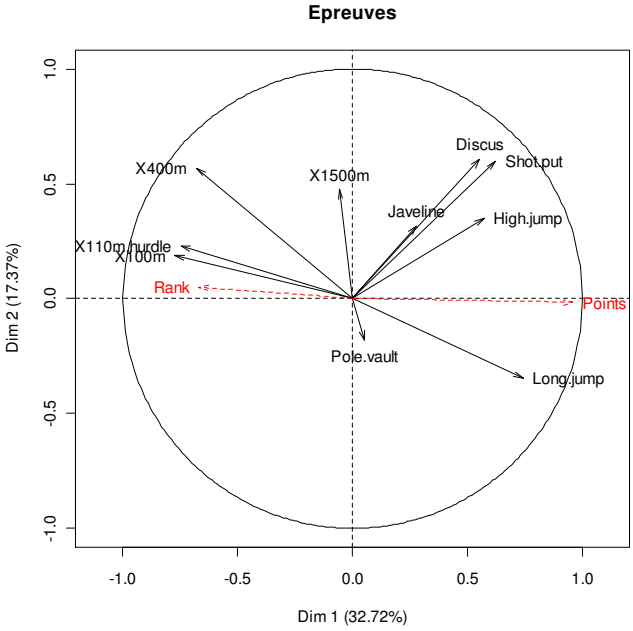
Réponse : Lorenzo semble être un athlète lent et faible. Il gagne donc ces points sur les épreuves qui ne sont pas représentées par les deux composantes principales. En effet, il est excellent au 1500m et au saut en hauteur, épreuves dans lesquels il termine premier.

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

Réponse : Casarsa est un athlète puissant mais lent, Karpov est rapide et Sebrle est rapide et puissant. Karpov et Serble terminent premiers de leur compétition et Casarsa termine dernier.

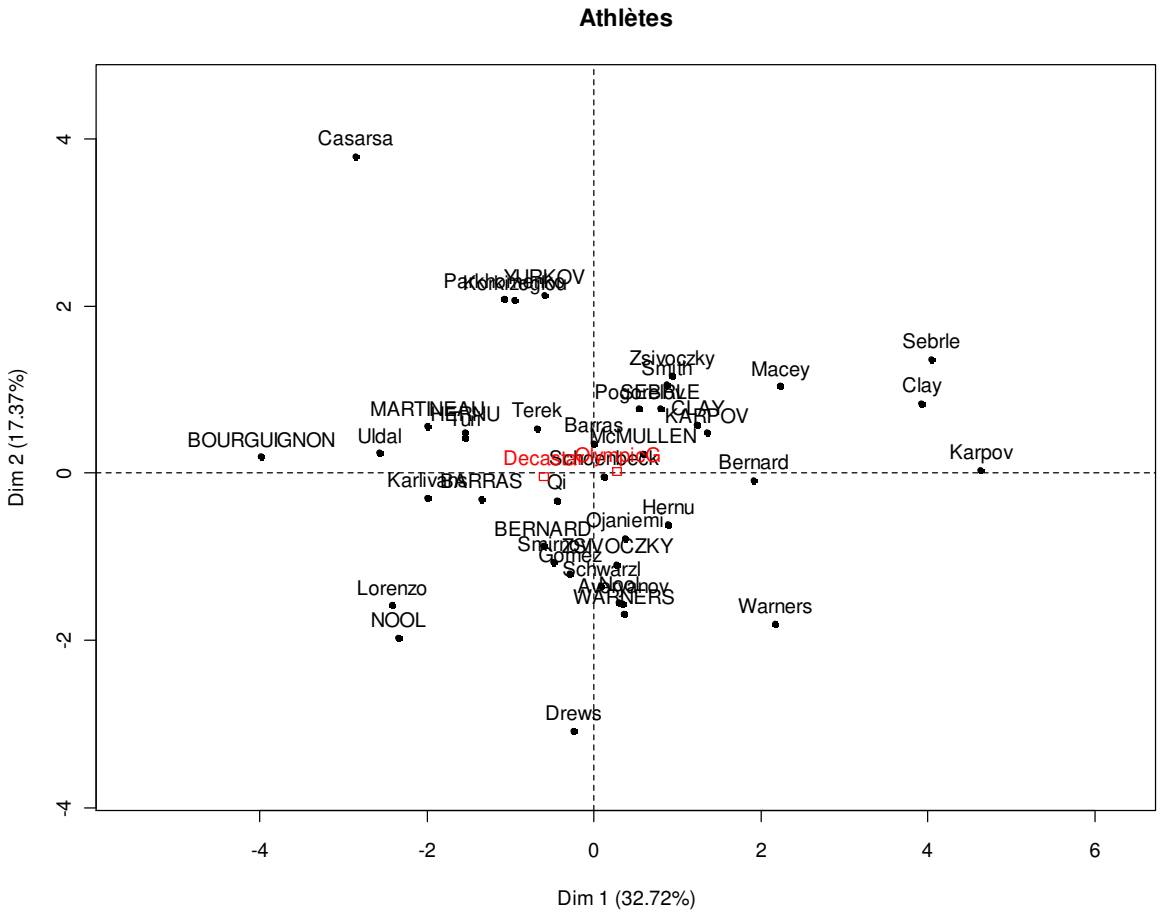
c) Peut-on en conclure qu'il faut être rapide pour gagner le décathlon et que la puissance ne suffit pas ? Pour répondre à cette question, on ajoute les variables supplémentaires « Rank » et « Points ». Ces variables n'entrent pas en compte dans le calcul des composantes principales mais aident à une meilleure compréhension des axes. Que pouvez-vous en conclure ?

Réponse : Rank et Points sont bien évidemment deux variables fortement corrélées (cos~-1). Elles sont négativement corrélées car plus on a de points et plus le rang est petit. Les variables les plus liées au nombre de points sont donc celles de la première composante. Les athlètes qui sont bons au 1500m et au pole.vault ne sont pas favorisés.



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 représentants un individu moyen pour chaque compétition sont ajoutés au graphique. Que pouvez-vous en conclure ?

Réponse : Le premier axe étant quasiment confondu avec la variable « Points », on peut conclure que les scores sont moins bons pour le décastar. Est-ce parce que les athlètes participants aux jeux olympiques sont meilleurs ou bien est-ce parce qu'ils sont plus motivés ? En revanche, les deux compétitions se situent au même niveau de l'axe 2, ce qui signifie que les individus ne change pas de profil.



Exercice 2 — Données :

A faire

Exercice 3 — Données : PoissonsData.txt

24 mulets ont été répartis dans trois aquariums, A1, A2 et A3, contaminés par radioactivité de façon identique mais, par contre, avec un temps d'exposition différent. Le poisson 17 est mort pendant l'expérience et n'apparaît pas dans le tableau.

On cherche à savoir s'il y a une influence du temps d'exposition à la radioactivité sur la taille des poissons. Nous disposons de mesures sur les 16 variables suivantes :

- 9 concernant la contamination à radioactivité, des yeux (ŒIL), des branchies (BRAN), des opercules (OPER), des nageoires (NAGE), du foie (FOIE), du tube digestif (TUDI), des reins (REINS), des écailles (ECAI), des muscles (MUSC).

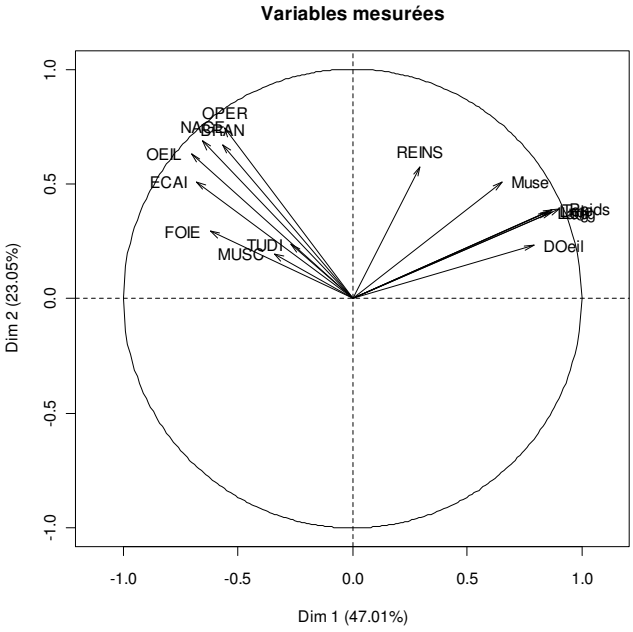
- 7 concernant la morphologie du poisson : le poids (Poid), la longueur (Long), la longueur standard (Lng), la largeur de la tête (Tete), la largeur (Larg), la largeur du museau (Muse), le diamètre des yeux (Doeil)

Effectuer une analyse en composantes principales pour répondre à la question.

Réponse :

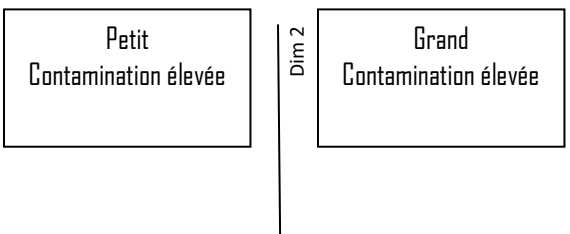
- Inertie : Les deux premières composantes principales expliquent déjà 70% de l'inertie.
- Lien entre variables : Exceptée la variable REINS, les autres variables se regroupent naturellement suivant leur catégorie. Le groupe des variables décrivant la morphologie de poisson et non corrélé au groupe des variables mesurant la contamination.
- Contribution aux axes :

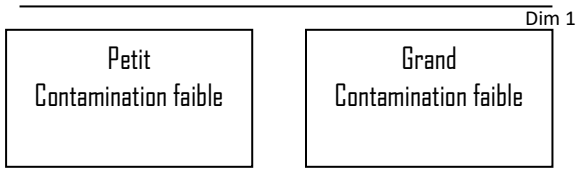
Les contributions sont très dispersées sur les variables. On note cependant que les variables liées à la morphologie contribuent majoritairement à la première composante principale (64% contre 36%) et celles liées à la contamination contribuent à la deuxième composante principale (72% contre 28%)




```
$contrib
           Dim.1      Dim.2        Dim.3      Dim.4      Dim.5
OEIL    6.5845064 10.752292 2.486881e-01  0.5559070  1.7388163
BRAN    4.2694318 12.255132 1.701309e+00  4.4499647  0.2103769
OPER    4.1529125 15.022110 4.257690e+00  2.4687557  0.9837650
NAGE    5.7014079 12.801906 1.963809e+00  1.3666213  2.5591510
FOIE    5.1298493  2.327672 1.565391e+01  5.1064535  5.8700527
TUDI    0.9556331  1.514304 7.440412e+00 65.3129242  4.0037845
REINS   1.1609291  8.996653 1.185718e+00  6.5181954 54.2036810
ECAI    6.1643181  7.032766 8.710107e-04  0.3202767  8.5881472
MUSC    1.5481717  1.027760 5.204269e+01  1.5017937  5.1690526
Poids  10.8344612  4.175412 7.883131e-02  0.2369399  0.4689956
Long   10.0635332  3.829962 1.425501e+00  0.1485902  0.1319147
Lng     9.8054557  3.932740 5.136667e-01  1.9602537  0.4458415
Tete   10.0857499  4.075143 8.709293e-01  1.2213263  0.2093709
Larg    9.5884679  3.822296 1.109204e+00  0.4078502  5.7586852
Muse    5.6428423  6.979339 9.135785e+00  2.8650169  1.2693719
DOeil   8.3123299  1.454514 2.370987e+00  5.5591308  8.3889931
```

On peut en déduire la cartographie suivante





Les poissons de l'aquarium 3 sont clairement localisés dans la zone petits poissons avec une contamination élevée. Ce qui semblerait indiquer que plus le poisson est exposé longtemps et plus sa taille est petite. Cependant la conclusion n'est pas confirmée par les autres aquariums, notamment le 1 qui a sa population divisée en deux sous-populations opposées avec d'un coté des grands poissons fortement contaminés et de l'autre des petits poissons peu contaminées.