TD1 - Analyse en composante Principale

MAF 2025-2026

Avant-propos

Les TD et TP du cours MAF servent trois objectifs:

- Revenir sur certaines notions de cours pour mieux les appréhender
- donner des outils pratiques pour mettre en oeuvre les méthodes vues en cours,
- réfléchir à leur mise en oeuvre et leur application dans des cas concrets.

Il est donc fréquent que l'on fasse certaine partie avec peu d'outils (à la main) pouur bien faire le lien avec les concets de cours, tandis que d'autres servent d'exemple pour illustrer comment dérouler une analyse sur un jeu de données en particulier.

Faire le lien avec le cours

On s'intéresse à un jeu de données de 100 individus, pour lesquels on mesure 4 variables

- le poids en Kg (variable `P`)
- la taille en m,
- la longueur de pas en cm,
- la vitesse de marche en $m.s^{-1}$.

```
# A tibble: 4 x 3
  variable     moyenne variance
  <chr>          <dbl>    <dbl>
1 weight         68.8    153.
2 height          1.69     0.01
3 steplength     70.5     52.1
4 walkingspeed    1.35     0.02
```

1. Que peut-on dire de la dispersion le long de chacun des axes initiaux.

2. Combien vaut l'inertie projeté sur l'axe de vecteur directeur $(1, 0, 0, 0)^\top$?

3. Combien vaut l'inertie projeté sur l'axe de vecteur directeur $(0, 1, 0, 0)^\top$?

On note X de dimension 100×4 la matrice des données centrées. Chaque individu a le même poids.

4. On souhaite utiliser la distance euclidienne puis réaliser une ACP. le cours indique qu'on diagonalise la matrice MVM. Indiquer ce que sont les matrices V et M impliquées.

5. On diagonalise la matrice MVM et on obtient les valeurs propres suivantes. Quelle est l'inertie portée par le premier plan principal, quel pourcentage de l'inertie totale ceci représente-t-il ? Pensez vous qu'on ait ainsi une bonne manière de représenter les données inditiales en seulement deux dimensions ?

```
[1] 160.754  44.752   0.004   0.002
```

6. Quelle est l'inertie portée par l'axe de vecteur directeur $(1, 0, 0, 0)^\top$ dans ce cas ?

7. Que deviennent les matrices V et M dans ce cas ?

8. Les valeurs propres de MVM sont données ci-dessous. Quelle est maintenant la part d'inertie représentée sur le premier axe principal, sur le premier plan principal.

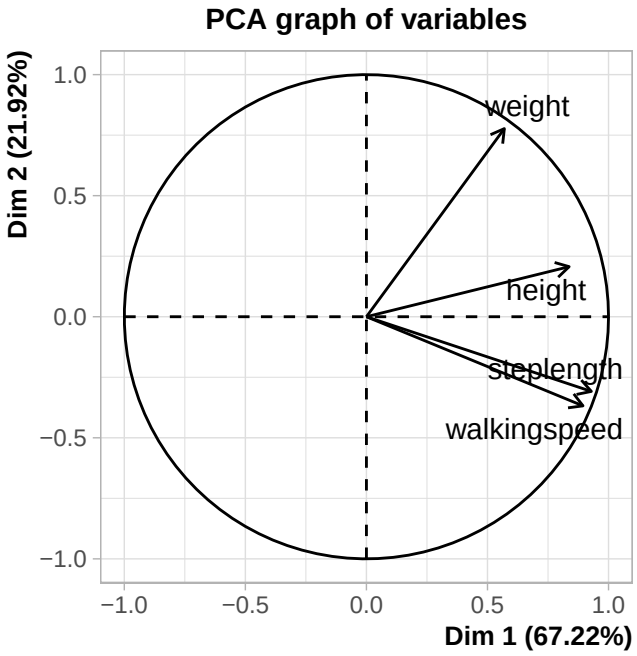
```
[1] 2.689 0.877 0.377 0.058
```

9. On utilise le package `FactoMineR` pour réaliser l'ACP. Peut on deviner quelle est la métrique utilisée par défaut ? Quelles sont les variables bien représentées ? Discuter du lien entre les variables.

```
walking_dta_pca <- PCA(walking_dta, graph = FALSE)
walking_dta_pca$eig
```

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.68869333              67.217333                           67.21733
comp 2 0.87687700              21.921925                           89.13926
comp 3 0.37659068               9.414767                           98.55403
comp 4 0.05783899               1.445975                          100.00000
```

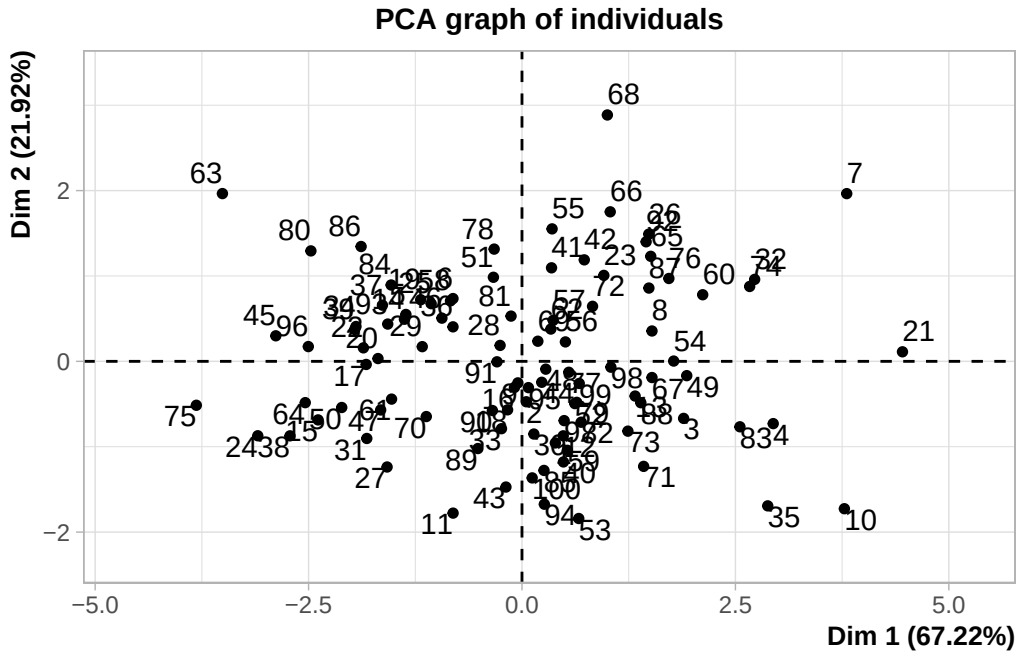
```
plot(walking_dta_pca, choix = "var")
```



10. En utilisant le graphique ci-dessous, que pouvez-vous dire

- sur l'individu 10 ?
- sur l'individu 94 ?

```
plot(walking_dta_pca, choix = "ind")
```



11. Quels sont les individus qui contribuent le plus à la formation du premier axe ?

```
walking_dta_pca$ind$contrib |> as_tibble() |> rowid_to_column("Ind") |> arrange(-Dim.1)
```

```
# A tibble: 100 x 5
     Ind Dim.1  Dim.2  Dim.3  Dim.4
   <int> <dbl>  <dbl>  <dbl>  <dbl>
 1    21  7.39 0.0142 1.09   0.0424
 2    75  5.41 0.302  2.09   1.94
 3     7  5.38 4.40   1.15   2.90
 4    10  5.30 3.40   2.80   2.01
 5    63  4.58 4.40   0.233  0.233
 6    24  3.56 0.866  0.333  3.48
 7     4  3.22 0.607  0.645  0.594
 8    45  3.09 0.102  0.0596 0.242
 9    35  3.08 3.27   0.0590 0.715
10    32  2.76 1.05   1.54   0.663
# i 90 more rows
```

12. Quels sont les individus les mieux représentés sur le premier axe ?

```
walking_dta_pca$ind$cos2 |> as_tibble() |> rowid_to_column("Ind") |> arrange(-Dim.1)
```

```
# A tibble: 100 x 5
     Ind Dim.1      Dim.2     Dim.3    Dim.4
   <int> <dbl>      <dbl>     <dbl>    <dbl>
 1    17 0.994 0.000398   0.00178   0.00398
 2    54 0.990 0.00000331 0.00686   0.00347
 3    45 0.985 0.0106     0.00266   0.00166
 4    21 0.979 0.000615   0.0202    0.000121
 5    96 0.972 0.00465    0.0184    0.00483
 6    29 0.962 0.0209     0.0000642 0.0169
 7    64 0.957 0.0348     0.00753   0.000453
 8    34 0.936 0.0412     0.0111    0.0120
 9    75 0.926 0.0169     0.0502    0.00713
10     8 0.925 0.0505     0.0144    0.0100
# i 90 more rows
```

13. Comment sont définies ces deux quantités ?
14. Comment mesurer la qualité de la représentation sur le premier plan ?

Une première ACP. Etude des performance au Décathlon

```
data("decathlon")
```

1. Discuter de la nature des variables

On a 11 variables quantitatives, les performances aux épreuves et le nombre total de points. Le rang est une variable numérique mais qui a un rôle particulier, on ne peut pas vraiment donner de sens au rang moyen par exemple. enfin on a une variable qualitative la competition dont il s'agit.

Le code ci-dessous calcule des résumés quantitatifs pour les variables pertinentes.

```
decathlon |> summarise(across(is.numeric, list(moyenne=mean, variance=var ))) |>
  pivot_longer(
    everything(),
    names_to = c("variable", ".value"),
    names_sep = "_"
  ) |>
   mutate(variance = variance *(n-1)/n) |> # pour corriger la variance par défaut de R
  mutate(across(is.numeric, \(x) round(x, 2)))
```

```
# A tibble: 12 x 3
   variable    moyenne  variance
   <chr>         <dbl>     <dbl>
 1 100m          11         0.07
 2 Long.jump      7.26      0.1
 3 Shot.put      14.5       0.67
 4 High.jump      1.98      0.01
 5 400m          49.6       1.32
 6 110m.hurdle   14.6       0.22
 7 Discus        44.3      11.3
 8 Pole.vault     4.76      0.08
 9 Javeline      58.3      23.1
10 1500m        279.      135.
11 Rank          12.1      62.1
12 Points      8005.   116055.
```

2. Pour comprendre les relations entre les variables, nous pouvons réaliser une ACP.

- Faut il prendre en compte toutes les variables quantitatives ?
- Quelle choix de distance feriez vous ?

3. l'ACP est réalisée grâce aux commandes suivantes. A quoi servent les options `quanti.sup` et `quali.sup` ? pourquoi ce choix ? Combien y a t il de variables quantitatives au total ?

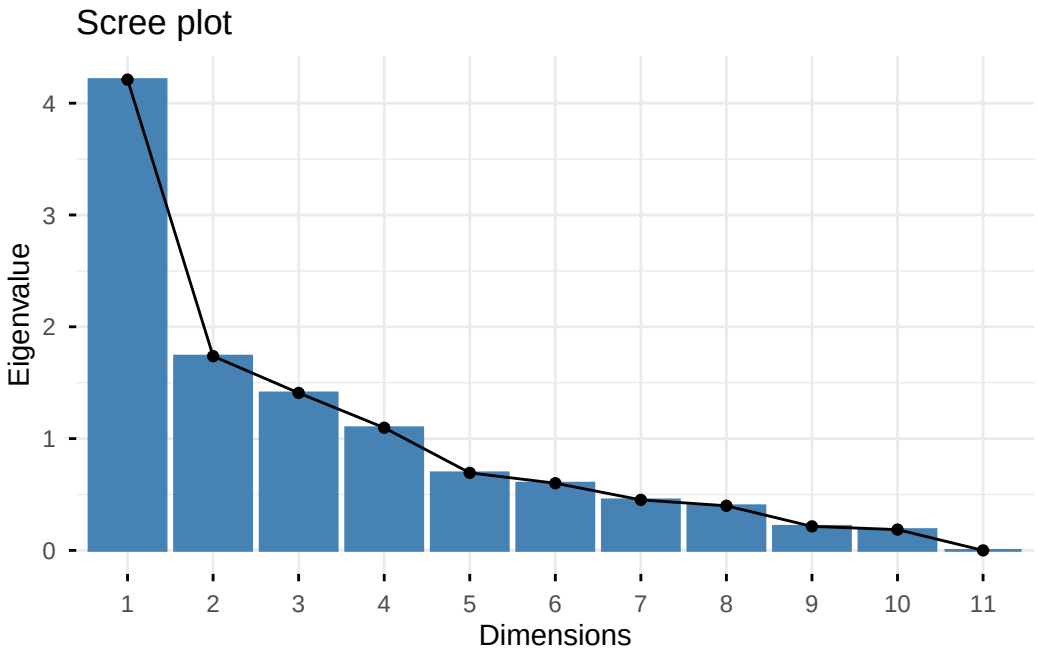
```
decathlon_pca <- PCA(decathlon, scale.unit = TRUE, quanti.sup = c("Rank"), quali.sup = "Compe
```

4. Combien suggérez d'axes vous semble-t-il pertinent de regarder ?

```
library(factoextra) # for nice vizualisation
decathlon_pca$eig
```

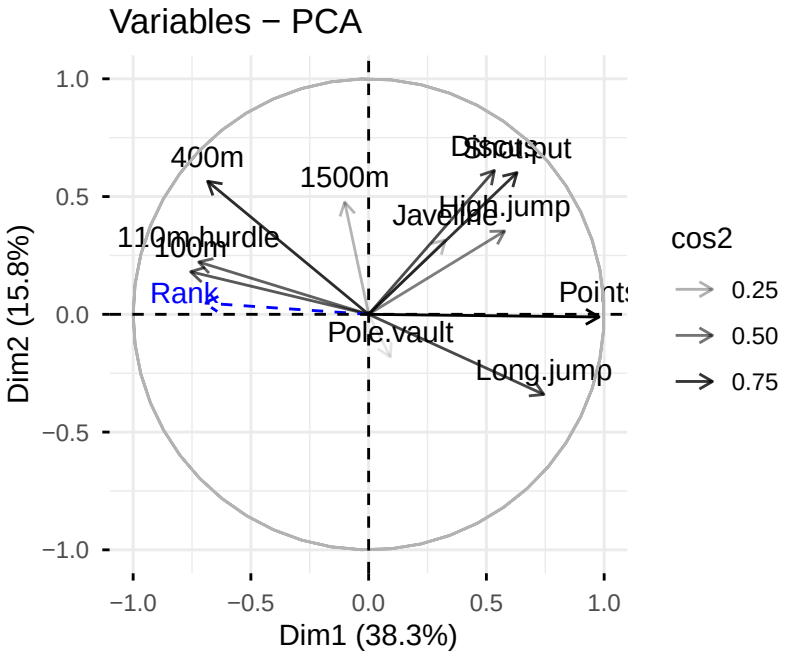
```
          eigenvalue percentage of variance cumulative percentage of variance
comp 1  4.210772e+00           3.827975e+01                             38.27975
comp 2  1.737315e+00           1.579378e+01                             54.07352
comp 3  1.408296e+00           1.280269e+01                             66.87621
comp 4  1.097035e+00           9.973046e+00                             76.84926
comp 5  6.936226e-01           6.305660e+00                             83.15492
comp 6  6.010811e-01           5.464374e+00                             88.61929
comp 7  4.516987e-01           4.106352e+00                             92.72564
comp 8  3.991488e-01           3.628626e+00                             96.35427
comp 9  2.148244e-01           1.952949e+00                             98.30722
comp 10 1.861517e-01           1.692288e+00                             99.99951
comp 11 5.439085e-05           4.944623e-04                            100.00000
```

```r
fviz_eig(decathlon_pca, choice = "eigenvalue", ncp = 11)
```



5. Quelles sont les variables bien représentées sur le premier plan principal ?

```r
library(factoextra) # for nice vizualisation
fviz_pca_var(decathlon_pca, axes = c(1,2), alpha.var = "cos2")
```



6. Discuter du lien entre les variables. Cela semble-t-il surprenant ?
7. Le saut à la perche `Pole.vault` est il bien représenté ? Sur quel plan faut-il projeté pour pouvoir visualiser les performances au saut à la perche ?

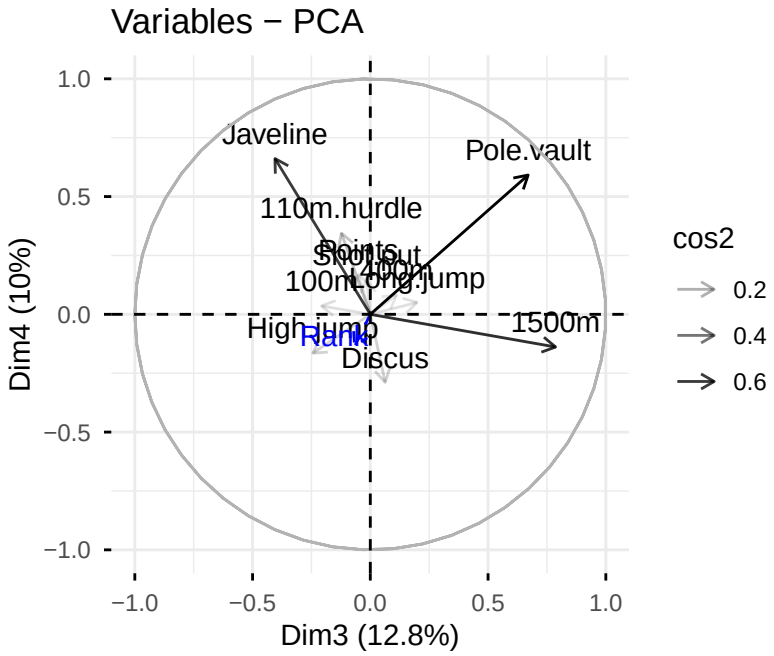
```
decathlon_pca$var$cos2
```

```
                Dim.1        Dim.2        Dim.3       Dim.4        Dim.5
100m        0.57139704 0.0327242707 0.0430338384 0.001274172 1.148933e-01
Long.jump   0.55357323 0.1157488070 0.0395671585 0.002557195 1.048259e-05
Shot.put    0.39706705 0.3621620493 0.0002382196 0.021897220 7.677701e-03
High.jump   0.33342735 0.1250871798 0.0595892073 0.027148070 2.787889e-01
400m        0.46743733 0.3191990038 0.0125288513 0.006698959 5.756890e-03
110m.hurdle 0.52071355 0.0493917625 0.0150700801 0.118904929 2.642918e-02
Discus      0.28508308 0.3733202080 0.0040004975 0.083457249 7.389082e-03
Pole.vault  0.00854153 0.0328341249 0.4490187957 0.349771085 9.983515e-02
Javeline    0.10647521 0.0996192520 0.1639116219 0.437392911 1.247085e-01
1500m       0.01034718 0.2271041183 0.6187491252 0.019227369 2.210338e-02
Points      0.95670958 0.0001246526 0.0025882124 0.028705925 6.029924e-03
                Dim.6        Dim.7        Dim.8        Dim.9       Dim.10
100m        0.051885058 0.0684509434 0.0819624457 2.464186e-03 0.031913545
Long.jump   0.060860974 0.1740178871 0.0009616512 5.047724e-02 0.002223441
Shot.put    0.046794392 0.0484975072 0.0444866878 3.987637e-02 0.031301962
High.jump   0.153216454 0.0042214724 0.0048514263 1.265994e-02 0.001007847
```

```
400m         0.059019037 0.0053962448 0.0229309061 6.457339e-02 0.036458409
110m.hurdle 0.006952002 0.0543807117 0.2006915068 4.967260e-03 0.002497964
Discus       0.124992676 0.0825979330 0.0002560070 5.261503e-03 0.033640071
Pole.vault   0.032468459 0.0049944097 0.0101802056 1.478939e-03 0.010875042
Javeline     0.016107763 0.0051044820 0.0310317324 1.290937e-02 0.002737377
1500m        0.047647885 0.0037097648 0.0002353866 2.014860e-02 0.030725498
Points       0.001136448 0.0003273564 0.0015608778 7.628454e-06 0.002770533
                   Dim.11
100m         1.196104e-06
Long.jump    1.938446e-06
Shot.put     8.482581e-07
High.jump    2.125406e-06
400m         9.753387e-07
110m.hurdle 1.056204e-06
Discus       1.696489e-06
Pole.vault   2.261812e-06
Javeline     1.732012e-06
1500m        1.694150e-06
Points       3.886663e-05
```

8. Le saut à la perche est la huitième épreuve au décathlon, le javelot la neuvième, le 1500 m la dernière. Que laisse espérer une bonne performance au saut à la perche, quant à la performance au javelot à venir ?

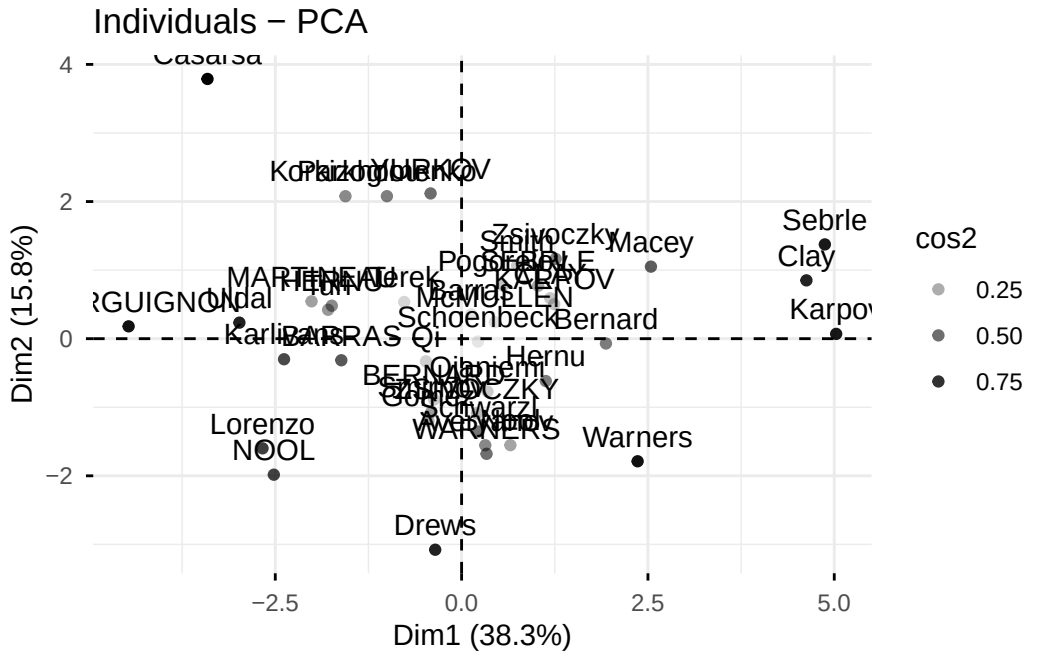
```r
fviz_pca_var(decathlon_pca, axes = c(3,4), alpha.var = "cos2")
```



9. La variable nombre de points est très bien représentée par l'axe 1. que pensez de l'intérêt des trois dernières épreuves sur le résultat final ?

10. Ci-dessous le code pour obtenir la représentation des individus dans le premier plan principal. Citer un excellent décathlonien. Discuter de la performance de Casarsa (en haut à gauche) au 400m. Peut-on affirmer que Qi (près de l'origine) est un décathlonien moyen ?

```
fviz_pca_ind(decathlon_pca, axes = c(1,2), alpha.ind = "cos2")
```



11. Le graphique suivant ajoute l'information sur la Competition. Que pouvez vous déduire des résultats au Decastar et aux Jeux Olympiques ?