Analyse des données décathlon

François

23/09/2020

## Importation des données

deca <- read.table("https://husson.github.io/img/decathlon04-18.csv", header=TRUE, sep=";", row.names=1, check.names=FALSE, stringsAsFactors = TRUE)  
summary(deca)

100m Longueur Poids Hauteur   
 Min. :10.30 Min. :6.620 Min. :12.84 Min. :1.870   
 1st Qu.:10.75 1st Qu.:7.190 1st Qu.:13.71 1st Qu.:1.960   
 Median :10.90 Median :7.490 Median :14.32 Median :2.025   
 Mean :10.92 Mean :7.434 Mean :14.46 Mean :2.023   
 3rd Qu.:11.13 3rd Qu.:7.603 3rd Qu.:15.24 3rd Qu.:2.083   
 Max. :11.40 Max. :7.960 Max. :16.36 Max. :2.170   
 400m 110m.haies Disque Perche   
 Min. :46.07 Min. :13.58 Min. :35.30 Min. :4.350   
 1st Qu.:48.25 1st Qu.:14.14 1st Qu.:43.20 1st Qu.:4.600   
 Median :48.96 Median :14.57 Median :44.99 Median :4.850   
 Mean :48.91 Mean :14.53 Mean :44.68 Mean :4.860   
 3rd Qu.:49.47 3rd Qu.:14.78 3rd Qu.:45.93 3rd Qu.:5.025   
 Max. :52.67 Max. :15.63 Max. :51.65 Max. :5.450   
 Javelot 1500m Rang Points Competition  
 Min. :49.45 Min. :254.6 Min. : 1.0 Min. :7276 CE18 :10   
 1st Qu.:57.19 1st Qu.:268.0 1st Qu.: 3.0 1st Qu.:8093 Decastar18:10   
 Median :61.12 Median :273.7 Median : 5.5 Median :8288 JO04 :10   
 Mean :61.52 Mean :274.2 Mean : 5.5 Mean :8265 JO16 :10   
 3rd Qu.:65.12 3rd Qu.:278.1 3rd Qu.: 8.0 3rd Qu.:8438   
 Max. :72.32 Max. :304.5 Max. :10.0 Max. :9126

### Question 1

Les individus sont les athlètes lors d’une compétition : il y a donc 40 individus statistiques, et les variables sont les perofrmances auxc 10 épreuves du décathlon. Les variables rang, nb de point et compétition seront supplémentaires. L’objectif est de comparer les performances des différents athlètes lors des 10 épreuves; on peut aussi comparer les résultats des différentes épreuves et savoir quelles épreuves font appel aux mêmes qualités physiques ou au contraire à des qualités physiques différentes. Enfin, ces 2 études sont liées et les différences entre athlètes pourront être interprétées par les variables de performance.

### Question 2

Les variables centrées-réduites indiquent tout d’abord si les athlètes ont des performances inférieurs ou supérieurs à la moyenne (selon le signe) : attention, pour les variables de vitesse, un athlète court vite si son temps est faible ! Elles montrent par exemple que certains ahlètes ont des performances exceptionnelles (valeurs centrée-réduiite supérieures à 1.96 en valeur absolue). Par exemple Sebrle lance le poids très loin (2.02), que MAYER n’a aucune performance exceptionnelle.

round(scale(deca[,1:10]),2)

100m Longueur Poids Hauteur 400m 110m.haies Disque Perche  
Sebrle -0.24 1.39 2.02 1.20 -0.47 -0.97 1.23 0.47  
Clay -1.72 1.80 0.82 0.46 0.23 -0.81 1.65 0.13  
Karpov -1.51 1.29 1.57 0.83 -1.77 -1.13 2.12 -0.87  
Macey -0.10 0.12 1.35 1.57 0.05 0.07 1.11 -1.55  
Warners -1.08 1.05 0.03 -0.66 -0.79 -1.05 -0.29 0.13  
Zsivoczky -0.03 -1.01 0.91 1.20 0.41 0.87 0.29 -0.54  
Hernu 0.19 -0.84 0.21 0.09 -0.16 -0.56 0.01 -0.20  
Nool -0.43 0.33 -0.21 -1.77 -0.09 0.56 -0.80 1.82  
Bernard -0.82 0.16 0.37 1.20 0.18 -0.73 0.02 -1.55  
Schwarzl 0.22 0.19 -0.47 -1.03 0.71 -0.56 -0.69 0.81  
Eaton -1.65 1.73 0.29 -0.16 -2.39 -1.48 0.25 1.14  
Mayer -0.39 0.57 1.39 0.21 -0.53 -1.03 0.64 1.82  
Warner -2.23 0.81 -0.85 0.21 -1.32 -1.93 0.08 -0.54  
Kazmirek -0.50 0.87 -0.27 0.95 -1.82 0.19 -0.44 0.47  
Bourrada -0.61 0.29 -0.72 0.95 -0.79 -0.77 -0.70 -0.87  
Suarez 1.06 -1.01 -0.20 0.58 -0.64 -0.09 0.73 0.13  
Ziemek -0.75 0.19 -1.08 0.95 0.77 0.50 1.44 1.14  
Van Der Plaetsen 1.16 0.77 -1.72 1.70 0.60 0.99 -0.34 1.82  
Felix 0.04 -0.05 0.33 0.58 0.19 0.54 0.13 -1.21  
Araujo -0.53 0.16 0.85 -1.28 -0.65 -0.73 0.13 0.13  
Abele -0.21 -0.05 1.26 -1.15 -0.76 -1.20 0.23 -0.87  
Samuelsson 0.44 0.29 -0.50 0.71 0.58 0.36 -0.23 -1.21  
Shkurenyov 0.73 0.40 -1.09 -0.04 0.03 -0.17 0.26 1.48  
Dolezal 0.51 -1.08 -0.45 -0.04 0.42 0.11 0.34 -0.20  
Roe -0.21 0.60 1.09 -0.78 0.42 1.60 -0.75 -0.20  
Kaul 1.60 -0.80 -0.64 0.71 0.31 0.52 0.49 -0.54  
Braun 1.49 0.29 -0.25 -0.04 -0.33 0.07 -0.28 -0.87  
Cairoli 0.08 0.19 -1.28 0.33 -0.12 0.28 -2.86 -0.87  
Duckworth -0.97 0.46 -0.90 1.82 0.80 0.05 -0.84 0.81  
Zhuk 0.73 -1.32 1.27 -0.41 -0.42 0.28 0.24 0.13  
MAYER -1.33 1.25 1.64 0.33 -0.42 -1.58 1.79 1.98  
ABELE -0.28 -0.84 0.79 -1.15 -0.37 -1.36 0.00 -0.37  
NOWAK 0.98 0.43 0.05 -0.78 0.32 0.40 0.19 -0.03  
BRAUN 0.91 -0.49 -0.09 -0.04 0.15 0.17 0.56 -1.38  
GADO -0.61 -0.02 -1.65 -1.52 -0.74 1.46 -2.20 0.97  
AUZEIL 1.71 -2.00 0.50 -0.78 0.94 0.52 0.11 -0.03  
WIESIOLEK 0.95 -1.25 -1.39 -0.78 1.48 1.28 0.02 -0.03  
NILSSON 1.74 -2.79 -0.77 -0.78 3.16 1.40 -0.49 0.30  
LELIEVRE 1.20 -0.94 -0.87 -1.89 0.99 2.26 -1.39 -1.71  
BARAN 0.44 -1.15 -1.35 -1.52 1.85 1.69 -1.77 -0.03  
 Javelot 1500m  
Sebrle 1.62 0.63  
Clay 1.47 0.85  
Karpov -1.07 0.42  
Macey -0.55 -0.95  
Warners -1.10 0.42  
Zsivoczky 0.35 -0.50  
Hernu -0.67 -1.06  
Nool -0.03 0.23  
Bernard -1.12 0.23  
Schwarzl -0.93 -0.07  
Eaton -0.31 -1.18  
Mayer 0.63 -0.94  
Warner 0.30 -1.01  
Kazmirek 0.55 -0.32  
Bourrada 0.89 -2.12  
Suarez 1.94 -0.64  
Ziemek -0.11 0.95  
Van Der Plaetsen 0.10 0.00  
Felix 1.51 -0.40  
Araujo -0.76 -0.29  
Abele 1.18 -0.36  
Samuelsson -0.83 0.42  
Shkurenyov -0.43 -0.30  
Dolezal 0.21 0.77  
Roe 0.54 0.78  
Kaul 1.11 -1.14  
Braun -0.36 -0.76  
Cairoli -0.34 -0.64  
Duckworth -1.21 2.61  
Zhuk 0.84 -0.37  
MAYER 1.86 0.21  
ABELE 0.69 -0.04  
NOWAK 0.45 -0.80  
BRAUN -0.90 0.24  
GADO -0.94 -0.81  
AUZEIL -0.33 1.03  
WIESIOLEK -0.70 1.22  
NILSSON 0.34 0.00  
LELIEVRE -1.73 0.40  
BARAN -2.17 3.28  
attr(,"scaled:center")  
 100m Longueur Poids Hauteur 400m 110m.haies Disque   
 10.91775 7.43450 14.45575 2.02300 48.91500 14.52525 44.68075   
 Perche Javelot 1500m   
 4.86000 61.51600 274.18550   
attr(,"scaled:scale")  
 100m Longueur Poids Hauteur 400m 110m.haies Disque   
0.27696605 0.29212704 0.94071137 0.08074017 1.18981791 0.48924320 3.28193219   
 Perche Javelot 1500m   
0.29746795 5.56825839 9.23776168

### Question 3

On remarque sur la matrice de corrélation de nombreuses corrélations assez élevées, mais jamais très élevées. Par exemple longueur et 100 m ont une corrélation de -0.72 (ceux qui courent vite sautent loin). On voit aussi par exemple une corrélation très faible entre 100 m et 1500m (0.016).

round(cor(deca[,1:10]),3)

100m Longueur Poids Hauteur 400m 110m.haies Disque Perche  
100m 1.000 -0.712 -0.294 -0.197 0.590 0.627 -0.228 -0.166  
Longueur -0.712 1.000 0.246 0.333 -0.636 -0.535 0.260 0.261  
Poids -0.294 0.246 1.000 0.110 -0.365 -0.494 0.618 -0.088  
Hauteur -0.197 0.333 0.110 1.000 -0.167 -0.246 0.409 -0.058  
400m 0.590 -0.636 -0.365 -0.167 1.000 0.659 -0.230 -0.059  
110m.haies 0.627 -0.535 -0.494 -0.246 0.659 1.000 -0.500 -0.107  
Disque -0.228 0.260 0.618 0.409 -0.230 -0.500 1.000 0.079  
Perche -0.166 0.261 -0.088 -0.058 -0.059 -0.107 0.079 1.000  
Javelot -0.053 0.124 0.422 0.260 -0.222 -0.309 0.442 0.186  
1500m 0.016 -0.120 -0.127 -0.114 0.508 0.323 -0.061 0.105  
 Javelot 1500m  
100m -0.053 0.016  
Longueur 0.124 -0.120  
Poids 0.422 -0.127  
Hauteur 0.260 -0.114  
400m -0.222 0.508  
110m.haies -0.309 0.323  
Disque 0.442 -0.061  
Perche 0.186 0.105  
Javelot 1.000 -0.381  
1500m -0.381 1.000

#### L’ACP avec Factoshiny

L’interface Factoshiny permet de facilement modifier les paramètres (choix de variables illustratives) et les graphes. Lorsqu’un graphe est bien travaillé, récupérer les lignes de code pour pouvoir relancer le graphe.

library(Factoshiny)  
res <- Factoshiny(deca)

### Question 4

res.PCA<-PCA(deca,quali.sup=c(13),quanti.sup=c(11,12),graph=FALSE)  
res.PCA$eig

eigenvalue percentage of variance cumulative percentage of variance  
comp 1 3.8177498 38.177498 38.17750  
comp 2 1.5341057 15.341057 53.51856  
comp 3 1.2362344 12.362344 65.88090  
comp 4 1.0567117 10.567117 76.44802  
comp 5 0.9447131 9.447131 85.89515  
comp 6 0.4258154 4.258154 90.15330  
comp 7 0.3527757 3.527757 93.68106  
comp 8 0.2347485 2.347485 96.02854  
comp 9 0.2037018 2.037018 98.06556  
comp 10 0.1934438 1.934438 100.00000

L’inertie du plan principal est de 53% (attention à mettre rang et nb de point en supplémentaire). Ce chiffre est à comparer à 38% qui correspond au quantile 95% de l’inertie obtenue en faisant des ACP avec des jeux de données de 40 individus et 10 variables au hasard : ainsi, ici il y a de l’information résumée sur le plan principal, ce n’est pas du hasard. Avec les dimensions 3 et 4, on arrive à 75%. Si le pouircentage d’inertie avec les 2 premiers axes étaient très élevé et proche de 100%, cela signifierait que les épreuves du décathlon ne font appel qu’à 2 dimensions physiques. Comme le décathlon couronne des athlètes complets (on parle des dieux du stade), il est préférable d’avoir besoin de plus de 2 dimensions pour résuler le tableau de données.

### Question 5

res.PCA$ind$contrib

Dim.1 Dim.2 Dim.3 Dim.4  
Sebrle 5.538217e+00 1.798768e+00 5.070802e+00 0.001072515  
Clay 4.845259e+00 1.521998e-04 7.162058e+00 0.133802583  
Karpov 6.820094e+00 4.092667e-01 9.171498e-01 12.789290409  
Macey 8.166876e-01 3.264255e+00 1.549833e-01 9.377110420  
Warners 6.424573e-01 6.350418e+00 2.602177e-02 0.753239766  
Zsivoczky 8.752585e-03 3.942407e+00 9.145571e-03 0.773890656  
Hernu 1.033772e-03 2.729674e-01 1.863400e+00 0.144066858  
Nool 2.048282e-01 3.741912e+00 9.039367e-02 8.078316355  
Bernard 1.866742e-01 1.277606e-01 3.859881e-03 12.676038183  
Schwarzl 4.426180e-01 1.976169e+00 1.236902e-02 0.929172062  
Eaton 7.130243e+00 6.492026e+00 2.044941e+00 1.031805395  
Mayer 3.342205e+00 8.134920e-02 2.706574e-01 4.878645184  
Warner 4.259313e+00 3.109890e+00 3.458671e+00 0.346323436  
Kazmirek 1.296972e+00 1.043833e+00 7.877977e-01 0.656854183  
Bourrada 1.166022e+00 6.583700e-02 1.035504e+01 0.036624206  
Suarez 1.508583e-01 7.017011e+00 3.017904e-01 3.198734006  
Ziemek 3.539518e-04 3.069121e-01 9.113948e+00 0.291792683  
Van Der Plaetsen 5.603953e-01 3.444962e-01 1.645270e+00 6.838938949  
Felix 5.892907e-02 3.628753e+00 3.216172e-01 0.275821652  
Araujo 3.705861e-01 7.587165e-01 4.741705e-01 0.080242456  
Abele 1.338367e+00 1.303056e+00 1.514641e+00 0.066670644  
Samuelsson 5.626552e-01 1.525277e-02 4.844825e-04 4.580351743  
Shkurenyov 4.255539e-02 6.598920e-01 1.265425e-01 4.316476237  
Dolezal 6.397310e-01 7.954752e-01 6.668918e-01 0.023375740  
Roe 2.990337e-01 8.088831e-03 5.015406e-01 0.001606131  
Kaul 2.885214e-01 7.383179e+00 1.257105e+00 0.807025450  
Braun 1.823334e-01 4.058832e-01 3.236096e+00 0.248091091  
Cairoli 1.186911e+00 2.763569e+00 8.718277e+00 0.063134252  
Duckworth 3.757880e-01 7.365062e+00 1.212791e+01 1.884780051  
Zhuk 1.108798e-04 4.544916e+00 1.785524e-01 0.704275312  
MAYER 8.711783e+00 2.146772e-01 7.570633e+00 4.905624095  
ABELE 3.093764e-01 5.198293e-01 5.666902e-01 0.001687142  
NOWAK 6.761713e-02 8.028149e-01 6.169924e-01 1.165927485  
BRAUN 5.501228e-01 7.627236e-01 1.094751e-01 4.812238808  
GADO 1.946398e+00 1.014206e+01 6.501184e+00 5.039871777  
AUZEIL 3.395769e+00 3.487574e+00 1.146276e+00 0.007140986  
WIESIOLEK 6.106391e+00 1.391561e-02 1.776571e+00 0.001550705  
NILSSON 1.094676e+01 7.142951e+00 6.679125e-01 3.102508428  
LELIEVRE 1.071513e+01 2.447300e-01 2.729404e+00 3.158480212  
BARAN 1.449215e+01 6.691446e+00 5.902736e+00 1.817401755  
 Dim.5  
Sebrle 0.003411015  
Clay 0.018772163  
Karpov 0.664702684  
Macey 1.348629198  
Warners 1.594193437  
Zsivoczky 0.959648183  
Hernu 0.049742081  
Nool 4.162164349  
Bernard 0.875134421  
Schwarzl 1.020787592  
Eaton 0.264806375  
Mayer 0.649886618  
Warner 0.364463268  
Kazmirek 3.033040193  
Bourrada 5.310972925  
Suarez 0.968023396  
Ziemek 3.721991511  
Van Der Plaetsen 19.069452361  
Felix 0.823846914  
Araujo 7.206559628  
Abele 7.745563563  
Samuelsson 2.169201970  
Shkurenyov 0.993512910  
Dolezal 0.004565760  
Roe 2.000508123  
Kaul 3.823479953  
Braun 0.355710141  
Cairoli 3.988724706  
Duckworth 6.540596009  
Zhuk 2.849766610  
MAYER 1.351000835  
ABELE 7.617989786  
NOWAK 0.228373633  
BRAUN 0.018883917  
GADO 0.034349449  
AUZEIL 3.668782583  
WIESIOLEK 0.007588222  
NILSSON 0.085243037  
LELIEVRE 2.068944462  
BARAN 2.336986017

Les individus qui contribuent le plus sont ceux qui ont la coordonnée la plus extrême sur un axe (BARAN 14.49 % et MAYER 8.71 % sur 1er axe; GADO 10.14% sur 2ème axe). Une forte contribution signifie que les dimensions de l’ACP sont influencées par ces individus. Mais ici les contributions sont plus fortes que pour les autres athlètes, mais pas exceptionnelles. Sans ces athlètes les axes resteraient identiques.

### Question 6

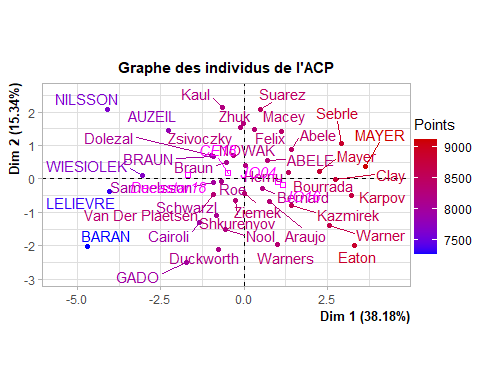
* 100 m haies - 400 m : corrélation positive et assez élevée mais comme variables pas parfaitement projetées, on ne peut connaître avec précision (vraie valeur 0.59)
* 100 m - Longueur : corrélation négative et forte en valeur absolue : ceux qui courent vite (temps faible) sautent loin (vraie valeur -0.712)
* Disque - Poids : corrélation positive (0.618)
* Disque - Longueur : corrélation faible (0.26)
* Perche - 1500 m : on ne peut rien dire, variables trop mal projetées

### Question 7

MAYER et LELIEVRE sont opposés sur la 1ère dimension, donc sur la principale dimension de variabilité. Ils ont donc des profils de performance très différents : ce n’est pas un bon signe pour LELIEVRE ! Contrairement à ce que son nom indique, il ne court pas vite (temps élevé) aux épreuves de vitesse (110m haies, 400 m 100 m) ! MAYER a un profil très complet car il court vite, mais aussi des valeurs élevées au javelot, disque, poids, hauteur, longueur. Les corrélations entre ces variables et la 1ère dimension ne sont pas très proches de 1, mais elles sont positives, ce qui indique qu’une coordonnée élevée sur une dimension implique une valeur plutôt élevée pour ces variables.

On peut colorier les athlètes en fonction de leur performance selon une variable ou selon le nombre de points.

plot.PCA(res.PCA,invisible=c('ind.sup'),habillage='Points',title="Graphe des individus de l'ACP",label =c('ind','quali'))



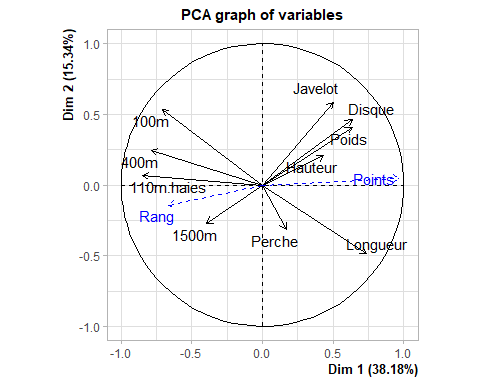
Kaul et Duckworth ont à peu près la même coordonnée sur la 1ère dimension, donc une performance globale assez proche. Par contre ils sont opposés sur la 2ème dimension. Donc par rapport à sa performance globale, Kaul lance plutôt loin disque et javelot mais ne saute pas loin et court pas vite le 100 m. Por Duckworth, c’est plutôt l’inverse. On imagine que KLaul est plutôt quelqu’un de costaud, un peu lourd pour courir et sauter, et pour Duckworth quelqu’un de léger qui court vite et saute loin, mais un peu frêle et qui ne lance pas loin disque et javelot (tout est relatif !).

Clay et Karpov sont proches sur le plan, et bien projetés sur le plan (cos² de 0.54 et 0.57, il faut additionner dim 1 et 2). On opeut donc dire que leurs profils de performance sont proches.

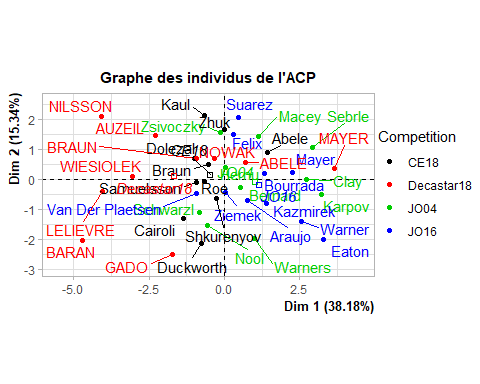
Pour Ziemek et Bernard les cos² sont très faibles, donc la proximité sur le plan ne traduit pas nécessairement une proximité dans l’espace. Il faudrait vois le plan 3-4 pour voir s’ils sont proches. Et on voit que ce n’est pas le cas.

### Question 8

plot(res.PCA, choix= "var")



plot.PCA(res.PCA,habillage=13,title="Graphe des individus de l'ACP")



La 1ère dimension oppose les éthlètes très bons au décathlon (MAYER, Sebrle) a des athlètes plus faibles (LELIEVRE, BARAN). Le nb de points est d’ailleurs très corrélé à cette dimension.

La 2ème dimension oppose des athlètes puissants (en haut, qui lancent loin disque, javelot et poids) a des athlètes véloce (qui court vite le 100 m et autent loin en longueur).

La 3ème dimension oppose à droite des athlètes qui ne court pas très vite 400 m et 1500 m mais qui lance plutôt loin le disque et saute plutôt haut à la perche (attention, la perche est surtout liée à la dim 4, mais un peu à la dim 3). A gauche, on a des athlètes qui courent vite 1500 m et plutot vite le 400 m, donc des athlètes endurants.

La 4ème dimension est liée à la perche et donc ceux qui sautent haut à la perche. Au passage, on peut voir qu’il n’y a pas de lien avec la hauteur.

### Question 9

La 1ère dimension est très liée au nombre de points donc les vainqueurs des décathlon excellent pour les variables bien projetées sur la 1ère dimension : les variables de vitesse, les lancers et les sauts. La perche est très particulière et peu liée au nombre de points.

### Question 10

On peut colorier les points selon la variable compétition, et construire le graphe avec seulement les modalités. On peut voir que les performances aux JO de 2004 et 2016 sont globalement proches et meilleures que lors des championnat d’Europe de 2018 ou que du decastar de 2018. Les performance du décastar de 2018 sont globalement faibles, les athlètes étant peut-être spectateurs du record mondial obtenu par MAYER.

plot.PCA(res.PCA,invisible='ind',habillage='Points',title="Graphe des individus de l'ACP")

