Ecole supérieure d'agro-développement international
istom

Faire une ACP avec R : le décathlon

Promotion 115

Table des matières

Aide-mémoire : l'ACP normée **2**

Exercice (guidé) : les performances de 41 décathloniens **4**

Indications **9**

Corrections **10**

Correction de l'exercice 1 10

Ce TD est une **ACP guidée**, à réaliser **avec R sous RStudio**, sur les performances de 41 décathloniens lors de deux compétitions de 2004 (le Décastar de Talence et les Jeux olympiques d'Athènes). L'ACP a deux objectifs :

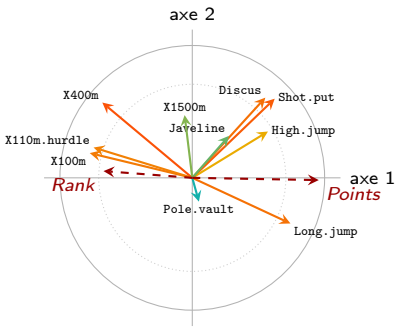
- **analyser les épreuves** : repérer celles qui sont liées entre elles, identifier des familles d'épreuves et celles qui pèsent le plus sur le total des points ;
- **analyser les athlètes** : repérer des profils semblables, ceux qui se distinguent dans certaines disciplines, et situer chacun selon ses points forts et ses points faibles.

Tout le code R est fourni ; à chaque étape, des encadrés **Lecture** expliquent les sorties et des lignes **✓ Contrôle** permettent de vérifier ses résultats. Le jeu de données est fourni dans le fichier texte `decathlon.txt`, à placer dans le dossier `data_raw` de votre projet RStudio. On utilise les packages `FactoMineR`, `factoextra` et `corrplot`.

Cet exercice entraîne à mener une ACP de bout en bout :

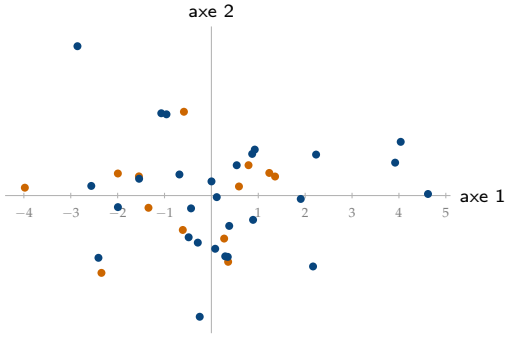
- **explorer les données** : valeurs manquantes, distributions, valeurs atypiques, matrice des corrélations ;
- **faire les bons choix** : ACP normée, variables actives et supplémentaires, nombre d'axes à retenir ;
- **lire les sorties de R** : valeurs propres, cercle des corrélations, carte des individus, $\cos^2$, contributions, valeurs-test, description automatique des axes (`dimdesc`) ;
- **interpréter** : nommer les axes (attention au sens des épreuves chronométrées !), revenir aux données brutes et rédiger une synthèse pour un entraîneur.

Un **aide-mémoire** (page 2) rassemble les formules, les commandes R correspondantes et les règles de lecture des graphiques. Les indications et la correction sont en fin de document.



Les épreuves

le cercle des corrélations : vitesse, force, endurance



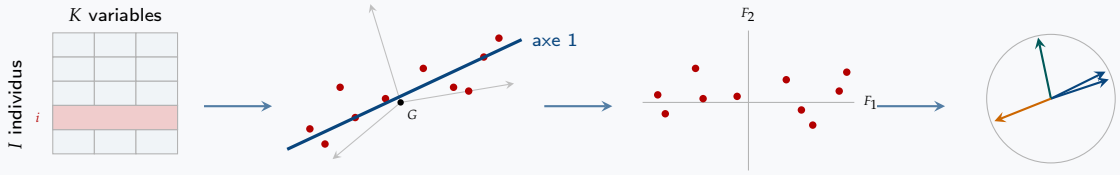
Les athlètes

la carte des 41 performances : Décastar, Jeux olympiques

Aide-mémoire : l'ACP normée

Ces deux pages rassemblent tout ce qui sert dans les quatre exercices : le principe de l'ACP, les formules « à la main » et leur équivalent sous R (`FactoMineR`), puis les règles de lecture des graphiques. Les notations sont celles du TD : I individus, K variables, z_{ik} la valeur centrée réduite de l'individu i pour la variable k.

Le principe : photographier un nuage de points sous son meilleur angle



1. le tableau
centré, réduit

2. le nuage des I points
axe 1 : direction d'étalement maximal

3. la carte (plan 1-2)
la « photographie » la plus fidèle

4. le cercle
les variables

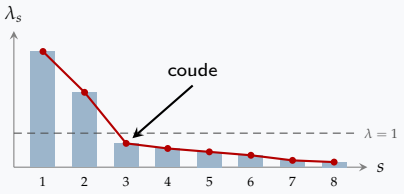
L'ACP cherche, dans l'espace à K dimensions, les directions orthogonales $U_1, U_2, \ldots$ le long desquelles le nuage est le plus étalé : la première porte la plus grande part d'inertie λ_1, la deuxième la plus grande part restante λ_2, etc. On lit ensuite les **individus** sur la carte et les **variables** sur le cercle, en faisant des allers-retours entre les deux.

Les étapes, à la main et sous R

Étape	À la main	Sous R (`res <- PCA(...)`)
Centrer, réduire	$z_{ik} = \dfrac{x_{ik} - \overline{X}_k}{\sigma_k}$ (σ_k : division par I)	`scale.unit = TRUE`
Distance au centre	$d^2(i,G) = \sum_k z_{ik}^2$ $\quad \mathcal{I}_{\text{tot}} = \frac{1}{I}\sum_i d^2(i,G) = K$	`res$ind$dist` (donne d)
Corrélations	$r_{kl} = \frac{1}{I}\sum_i z_{ik} z_{il}$ $\quad R = \frac{1}{I} Z^{\mathrm{T}} Z$	`cor(...)`
Valeurs propres	$\det(R - \lambda \mathrm{I}) = 0$ $\quad \sum_s \lambda_s = K$ $\quad \% = 100\,\lambda_s / K$	`res$eig`
Vecteurs propres	$R\,U_s = \lambda_s\,U_s$ $\quad \lVert U_s \rVert = 1$	`res$svd$V`
Coordonnées des individus	$F_s(i) = \sum_k u_{sk} z_{ik}$ $\quad \mathrm{Var}(F_s) = \lambda_s$	`res$ind$coord`
Qualité (individus)	$\cos^2_s(i) = \dfrac{F_s(i)^2}{d^2(i,G)}$ $\quad \sum_s \cos^2_s(i) = 1$	`res$ind$cos2`
Contribution (individus)	$\mathrm{contrib}_s(i) = 100\,\dfrac{F_s(i)^2}{I\,\lambda_s}$ $\quad$ moyenne $\dfrac{100}{I}$	`res$ind$contrib`
Coordonnées des variables	$r(X_k, F_s) = \sqrt{\lambda_s}\,u_{sk}$	`res$var$coord` (`$cor`)
Qualité (variables)	$\cos^2_s(X_k) = r^2(X_k, F_s)$ $\quad \sum_s r^2(X_k, F_s) = 1$	`res$var$cos2`
Contribution (variables)	$\mathrm{contrib}_s(X_k) = 100\,\dfrac{r^2(X_k, F_s)}{\lambda_s}$ $\quad$ moyenne $\dfrac{100}{K}$	`res$var$contrib`

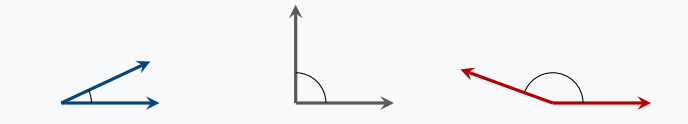
✓ **Contrôle** deux façons de retrouver la même valeur propre : $\lambda_s = \frac{1}{I}\sum_i F_s(i)^2$ (côté individus) $= \sum_k r^2(X_k, F_s)$ (côté variables).

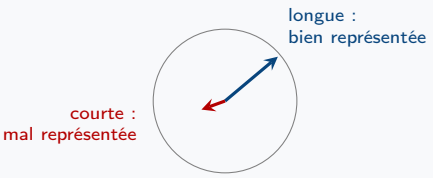
Combien d'axes ? Deux critères à croiser



- **Kaiser** : garder les axes tels que $\lambda_s \geq 1$ (ils résument au moins autant qu'une variable de départ).
- **Coude** : garder les axes situés avant la cassure de l'éboulis.
- Regarder aussi le **% d'inertie cumulé**, et ne garder que des axes que l'on sait **interpréter**.

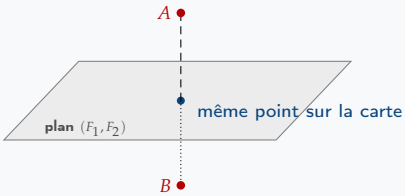
Lire le cercle des corrélations





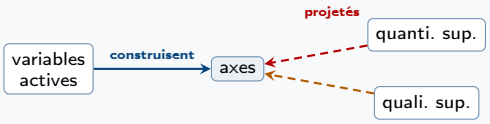
- La coordonnée d'une flèche sur un axe est la **corrélation** de la variable avec cet axe ; on nomme l'axe avec les variables qui y **contribuent** le plus (lues *axe par axe*).
- L'angle entre deux flèches approche leur corrélation ($r \approx \cos\theta$) **seulement si les deux flèches arrivent près du cercle**.

Lire la carte des individus



- Deux points proches sur la carte ne se ressemblent que s'ils sont **tous deux bien représentés** ($\cos^2$ sur le plan proche de 1) ; sinon, comme A et B, ils peuvent être éloignés.
- Un point situé dans la direction d'une flèche a une forte valeur de cette variable.
- Un individu de **contribution** très supérieure à $100/I$ « tire » l'axe : vérifier la donnée, refaire l'ACP en le mettant en supplémentaire.

Éléments supplémentaires



- Ils ne construisent pas les axes : on les place ensuite pour *interpréter* (ce que l'on cherche à expliquer, variables qualitatives).
- Une modalité est placée au barycentre de ses individus ; sa position sur un axe est significative si $|\text{v.test}| > 2$ (`res$quali.sup$v.test`, `dimdesc`).

Les pièges classiques. (1) Faire une ACP non normée sur des variables d'unités différentes : la variable de plus grande variance écrase les autres. (2) Interpréter la proximité de points ou l'angle de flèches mal représentés. (3) Croire qu'un axe classe du « meilleur » au « moins bon » : il oppose souvent deux profils. (4) S'inquiéter d'un axe inversé : son sens est arbitraire (un logiciel peut le retourner). (5) Nommer les axes avec les contributions *au plan* : elles mélangent les deux axes. (6) Conclure à une causalité : l'ACP décrit des associations.

Exercice 1 ACP guidée sous R : les performances de 41 décathloniens

Le décathlon réunit **dix épreuves** d'athlétisme disputées sur deux jours : courses (100 m, 400 m, 110 m haies, 1500 m), sauts (longueur, hauteur, perche) et lancers (poids, disque, javelot). Chaque performance est convertie en points par un barème, et le classement se fait sur le **total des points**. Un entraîneur se demande quelles épreuves « vont ensemble », quelles qualités font gagner, et quels profils d'athlètes on rencontre.

Le fichier `decathlon.txt` (données du package `FactoMineR`) contient les performances de **41 athlètes** lors de deux compétitions de 2004 : le *Décastar* de Talence et les *Jeux olympiques* d'Athènes. Chaque ligne correspond à un athlète ; le nom de l'athlète sert de nom de ligne. Les colonnes sont :

Colonne	Signification	Colonne	Signification
`100m`	100 m (s)	`Pole.vault`	saut à la perche (m)
`Long.jump`	saut en longueur (m)	`Javeline`	lancer du javelot (m)
`Shot.put`	lancer du poids (m)	`1500m`	1500 m (s)
`High.jump`	saut en hauteur (m)	`Rank`	classement dans la compétition
`400m`	400 m (s)	`Points`	total des points
`110m.hurdle`	110 m haies (s)	`Competition`	`Decastar` ou `OlympicG`
`Discus`	lancer du disque (m)		

☞ Récupérer les données

Scanner le QR code avec un téléphone, ou cliquer sur le lien `decathlon.txt`, puis télécharger le fichier avec le bouton Download raw file de la page GitHub. L'enregistrer dans le dossier `data_raw` de votre projet RStudio.

https://github.com/ant0in3g/ant0in3g.github.io/blob/main/sTa7/td-R-lab/R-lab/data_raw/decathlon.txt

Mode d'emploi

- Chaque question donne le code R à taper *dans un script* (*à enregistrer dans votre dossier OneDrive :* `STAT_Nom_Prenom`), pas dans la console : on garde ainsi une trace de tout ce que l'on fait. Pour exécuter une ligne, placer le curseur dessus et taper `Ctrl + Entrée`.
- Après chaque bloc de code : exécuter le code, comparer la sortie aux lignes ✓ **Contrôle**, lire les encadrés **Lecture**, puis répondre aux questions.
- Tout ce qui suit le signe # est un commentaire : R l'ignore.

1. Préparer la séance

a) Dans votre dossier de travail OneDrive, placer le fichier `decathlon.txt` (encadré « Récupérer les données » ci-dessus) dans `data_raw`. Lancer RStudio à partir du fichier `_stat_Istom.Rproj`.

b) Installer les packages `FactoMineR`, `factoextra` et `corrplot` avec RStudio, puis les charger (*à chaque séance*) :

```
library(FactoMineR)    # pour l'ACP
library(factoextra)    # pour les graphiques de l'ACP (fonctions fviz_...)
library(corrplot)      # pour la matrice des correlations en couleurs
```

c) Importer les données :

```
deca <- read.table("data_raw/decathlon.txt", header = TRUE, sep = "\t")
deca$Competition <- as.factor(deca$Competition)  # variable qualitative (facteur)
str(deca)                                         # structure du tableau
head(deca)                                        # les 6 premieres lignes
```

Lecture. `header = TRUE` : la première ligne du fichier contient les noms des colonnes ; `sep = "\t"` : les colonnes sont séparées par des tabulations. La ligne d'en-tête ayant une colonne de moins que les lignes de données, R comprend tout seul que la première colonne (le nom de l'athlète) donne les *noms de lignes*.

✓ **Contrôle** `str` affiche `41 obs. of 13 variables` ; les douze premières colonnes sont numériques (`num` ou `int`) et `Competition` est un `Factor w/ 2 levels`.

R a renommé `100m` en `X100m` : pourquoi, à votre avis ? Quels sont les numéros de colonne de `Rank`, `Points` et `Competition` ? Combien d'individus et de variables compte le tableau ?

2. Les valeurs manquantes

En pratique, il y a presque toujours des valeurs manquantes dans un jeu de données ; il faut les repérer *avant* toute analyse.

```
colSums(is.na(deca))      # nombre de valeurs manquantes par colonne
mean(is.na(deca))         # proportion de cases manquantes dans tout le tableau
```

Lecture. `is.na(deca)` remplace chaque case par `TRUE` si elle est manquante, `FALSE` sinon ; `colSums` compte les `TRUE` de chaque colonne. Pour une vue graphique, le package `VIM` propose `aggr(deca, numbers = TRUE, sortVars = TRUE)`.

Le tableau contient-il des valeurs manquantes ? Pourquoi est-ce important pour l'ACP ?

3. Décrire les données avant l'ACP

a) Afficher un résumé de chaque variable et les écarts-types des dix épreuves :

```
summary(deca)
round(apply(deca[, 1:10], 2, sd), 2)    # ecart-type de chacune des colonnes 1 a 10
```

✓ **Contrôle** la moyenne de `X100m` vaut 10,998 s et celle de `X1500m` 279,0 s.

Les épreuves ont-elles la même unité ? Pour quelles épreuves une *petite* valeur correspond-elle à une *bonne* performance ?

b) Tracer l'histogramme et la boîte à moustaches de chaque épreuve :

```
par(mfrow = c(2, 5))      # 10 graphiques sur 2 lignes et 5 colonnes
for (v in names(deca)[1:10]) hist(deca[[v]], main = v, xlab = "", col = "steelblue")
for (v in names(deca)[1:10]) boxplot(deca[[v]], main = v, col = "skyblue")
par(mfrow = c(1, 1))      # retour a un seul graphique par fenetre
```

Lecture. La boucle `for` répète la commande pour chacun des dix noms de colonnes ; `deca[[v]]` extrait la colonne dont le nom est contenu dans `v`. Sur une boîte à moustaches, les points isolés sont les valeurs situées à plus de 1,5 fois l'écart interquartile de la boîte.

Les distributions sont-elles à peu près symétriques ? Repérer les valeurs isolées, puis retrouver les athlètes correspondants, par exemple :

```
deca[deca$X1500m > 310, ]      # qui a couru le 1500 m en plus de 310 s ?
```

S'agit-il d'erreurs de saisie ? Faut-il retirer ces athlètes ?

c) Décrire la variable `Competition` :

```
table(deca$Competition)
round(prop.table(table(deca$Competition)), 3)
barplot(prop.table(table(deca$Competition)), col = c("orange", "steelblue"))
```

Les deux compétitions sont-elles également représentées ? Afficher `rownames(deca)` : que remarque-t-on sur l'écriture des noms ? Certains athlètes apparaissent-ils deux fois ?

4. Les corrélations entre épreuves

On représente la matrice des corrélations des dix épreuves (coefficients affichés en %) :

```
Matrice.Correlation <- cor(deca[, 1:10])
corrplot(Matrice.Correlation, method = "color", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, addCoef.col = "black",
         addCoefasPercent = TRUE, number.cex = 0.8)
```

Lecture. `order = "hclust"` réordonne les variables pour rapprocher celles qui sont corrélées entre elles : des blocs de couleur apparaissent alors le long de la diagonale.

✓ **Contrôle** $r(\texttt{X100m}, \texttt{X110m.hurdle}) \approx 0{,}58$ et $r(\texttt{Shot.put}, \texttt{Discus}) \approx 0{,}62$.

a) Quels groupes d'épreuves sont corrélés entre eux ? Ces liaisons vous semblent-elles logiques pour un athlète ?

b) La corrélation entre `X100m` et `Long.jump` vaut $-0{,}60$. Les bons sprinteurs sautent-ils donc moins loin ? Expliquer le signe.

c) Quelles épreuves sont peu corrélées à toutes les autres ? Qu'en attendre pour l'ACP ?

5. Centrer et réduire les données

```
data.CR <- scale(deca[, 1:10], center = TRUE, scale = TRUE)
round(head(data.CR, 3), 2)
```

a) Vérifier à la main le coefficient de `SEBRLE` pour le 100 m, à l'aide de la moyenne et de l'écart-type obtenus à la question 3. Que signifie une valeur positive pour une épreuve chronométrée ?

b) Pourquoi faut-il faire une ACP *normée* ? Calculer la part de la variance totale due au 1500 m si l'on ne réduisait pas les données (`v <- apply(deca[, 1:10], 2, var)` puis `round(100 * v / sum(v), 1)`).

6. Lancer l'ACP

a) On construit les axes avec les *dix épreuves* seulement, et l'on garde le classement, les points et la compétition pour *interpréter*. Taper :

```
res <- PCA(deca, scale.unit = TRUE, quanti.sup = 11:12, quali.sup = 13,
           graph = FALSE)
names(res)
```

Expliquer le rôle de chacun des arguments `scale.unit = TRUE`, `quanti.sup = 11:12`, `quali.sup = 13` et `graph = FALSE`.

b) Pourquoi `Points` et `Rank` doivent-ils être des variables *supplémentaires* et non actives ?

c) On n'a pas donné `data.CR` à la fonction `PCA`. Est-ce un problème ?

7. Combien d'axes faut-il regarder ?

Rappels. En ACP normée, chacune des K variables actives a une variance égale à 1 : l'inertie totale vaut K. Le **critère de Kaiser** retient les axes tels que $\lambda_s \geq 1$, c'est-à-dire ceux qui résument au moins autant qu'une variable de départ (ici, avec $K = 10$, au moins $10\,\%$ de l'inertie). Le **critère du coude** retient les axes situés *avant* la cassure de l'éboulis.

```
round(res$eig, 2)
fviz_eig(res, addlabels = TRUE)    # eboulis des valeurs propres (en %)
```

✓ **Contrôle** $\lambda_1 \approx 3{,}27$ et $\lambda_2 \approx 1{,}74$.

a) Que vaut la somme des valeurs propres (`sum(res$eig[, 1])`) ? Pourquoi ?

b) Quel pourcentage d'inertie le plan $(1,2)$ restitue-t-il ? Ce pourcentage vous semble-t-il élevé ? Le relier aux corrélations observées à la question 4.

c) Combien d'axes retenez-vous selon le critère de Kaiser ? selon le critère du coude ?

8. Le cercle des corrélations : interpréter les axes

> **Rappels.** Une flèche **longue** (proche du cercle) est bien représentée sur le plan ; une flèche **courte** est mal représentée, et l'on ne l'interprète pas. Pour deux flèches bien représentées, un angle proche de $0°$ traduit une forte corrélation positive, un angle proche de $90°$ une absence de corrélation, un angle proche de $180°$ une forte corrélation négative.

a) Tracer le cercle colorié selon la qualité de représentation, et afficher les résultats des variables :

```
fviz_pca_var(res, col.var = "cos2", repel = TRUE,
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"))
round(res$var$coord[, 1:3], 2)     # coordonnees = correlations avec les axes
round(res$var$cos2[, 1:3], 2)      # qualite de representation
round(res$var$contrib[, 1:3], 1)   # contributions (en %)
```

> **Lecture.** Le $\cos^2$ d'une variable sur le plan vaut $\cos_1^2 + \cos_2^2$: c'est le carré de la longueur de sa flèche. La contribution d'une variable à l'axe s vaut $100 \times r^2(X_k, F_s)/\lambda_s$: elle indique quelle part de l'axe est « fabriquée » par cette variable. La contribution moyenne vaut $100/10 = 10\,\%$.

b) Afficher les contributions des variables aux deux premiers axes :

```
fviz_contrib(res, choice = "var", axes = 1)   # contributions a l'axe 1
fviz_contrib(res, choice = "var", axes = 2)   # contributions a l'axe 2
fviz_cos2(res, choice = "var", axes = 1:2)    # qualite sur le plan (1,2)
```

> **Lecture.** Les barres sont triées par ordre décroissant ; la ligne rouge en pointillés indique la contribution moyenne ($10\,\%$).

c) **Axe 1.** Quelles épreuves lui sont fortement corrélées ? Dans quel sens ? Attention au sens des épreuves chronométrées ! Proposer un nom pour cet axe.

d) **Axe 2.** Mêmes questions.

e) Quelles épreuves sont mal représentées sur le plan $(1,2)$? Sur quel axe se trouve leur information ?

f) **Faire le lien avec le cours.** Pour `X100m`, vérifier avec les valeurs affichées que $\cos_1^2 = r^2(\texttt{X100m}, F_1)$ et que $\text{contrib}_1 = 100 \times r^2(\texttt{X100m}, F_1)/\lambda_1$.

g) Le cercle confirme-t-il les groupes d'épreuves repérés à la question 4 a) ? Quels conseils d'entraînement en tirer pour un décathlonien ?

9. La carte des athlètes

a) Tracer la carte des athlètes coloriée selon la qualité de représentation, et afficher leurs résultats :

```
fviz_pca_ind(res, col.ind = "cos2", repel = TRUE) +
  scale_color_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0.5)
round(res$ind$coord[, 1:2], 2)                    # coordonnees F1, F2
round(rowSums(res$ind$cos2[, 1:2]), 2)            # qualite sur le plan (1,2)
round(sort(res$ind$contrib[, 1], decreasing = TRUE)[1:5], 1)  # 5 plus fortes contrib.
    axe 1
round(sort(res$ind$contrib[, 2], decreasing = TRUE)[1:5], 1)  # idem pour l'axe 2
```

✓ **Contrôle** Karpov a la plus forte coordonnée sur l'axe 1 : $F_1 \approx 4{,}62$.

b) Quels athlètes ont les meilleurs profils d'ensemble ? les moins bons ? Quels athlètes contribuent le plus aux axes 1 et 2 (comparer à la contribution moyenne $100/41 \approx 2{,}4\,\%$) ?

c) Quel athlète est le plus mal représenté sur le plan ? Peut-on dire qu'il est « moyen dans toutes les épreuves » ?

d) **Retour aux données brutes.** Afficher les performances de trois athlètes situés à trois coins différents de la carte :

```
deca[c("Karpov", "Casarsa", "Drews"), 1:10]
```

Leurs performances sont-elles cohérentes avec les noms donnés aux axes ?

e) Tracer le *biplot* des athlètes et des variables bien représentés :

```
fviz_pca_biplot(res, select.ind = list(cos2 = 0.5), select.var = list(cos2 = 0.5),
                repel = TRUE, col.ind = "blue", col.var = "red")
```

Quel lien fait-on entre la position d'un athlète et la direction des flèches ?

10. Les variables supplémentaires

Rappel. Une variable **supplémentaire** n'intervient pas dans le calcul des axes : elle est projetée *après* coup pour aider à les interpréter. Quantitative, elle se lit sur le cercle ; qualitative, ses modalités se lisent sur la carte des individus (barycentres).

a) Afficher la corrélation des points et du classement avec les axes :

```
round(res$quanti.sup$coord[, 1:2], 2)
round(res$quanti.sup$cos2[, 1:2], 2)
```

Que confirme la position de `Points` sur le cercle ? Pourquoi `Rank` est-il nettement moins corrélé à l'axe 1 que `Points` ?

b) Tracer la carte des athlètes coloriée selon la compétition, et afficher les barycentres et leurs valeurs-test :

```
fviz_pca_ind(res, habillage = 13, addEllipses = TRUE, ellipse.level = 0.68,
             repel = TRUE)
round(res$quali.sup$coord[, 1:2], 2)
round(res$quali.sup$v.test[, 1:2], 2)
```

Lecture. La valeur-test (`v.test`) mesure si les athlètes d'une compétition ont, sur un axe, une position moyenne significativement différente de 0 : le lien est significatif (au seuil de $5\,\%$) si $|\texttt{v.test}| > 1{,}96$, c'est-à-dire environ 2.

Les athlètes des Jeux olympiques ont-ils significativement de meilleurs profils que ceux du Décastar ?

c) Neuf athlètes ont participé aux deux compétitions (par exemple `SEBRLE` au Décastar et `Sebrle` aux Jeux olympiques) :

```
deca[c("SEBRLE", "Sebrle", "NOOL", "Nool"), ]
res$ind$coord[c("SEBRLE", "Sebrle", "NOOL", "Nool"), 1:2]
```

Comment leur position évolue-t-elle sur la carte entre les deux compétitions ? Cela nuance-t-il la conclusion de la question précédente ?

d) Obtenir la description automatique des axes :

```
dimdesc(res, axes = 1:2, proba = 0.05)
```

Lecture. Pour chaque axe, `$quanti` liste les variables (actives ou supplémentaires) significativement corrélées à l'axe, de la plus positive à la plus négative ; `$quali` et `$category` n'apparaissent que si une variable qualitative est significativement liée à l'axe.

Quelles variables décrivent le mieux chaque axe ? Pourquoi la compétition n'apparaît-elle pas ?

11. Synthèse

Rédiger, pour l'entraîneur, une synthèse de l'ACP en une dizaine de lignes : les groupes d'épreuves, les deux grandes dimensions du décathlon, ce qui fait gagner, les profils d'athlètes remarquables, et des pistes d'entraînement. Préciser les limites de l'analyse.

Indication ▼ Correction ▼ [stat-0122]

Indications

Indication pour l'exercice 1 ▲

1. Si `read.table` ne trouve pas le fichier, vérifier qu'il est bien dans le dossier `data_raw`, que RStudio a été lancé depuis le fichier `.Rproj` (`getwd()` affiche le dossier de travail) et que le chemin s'écrit avec une barre oblique `/`. En R, un nom de variable ne peut pas commencer par un chiffre.
2. Pour l'ACP, chaque individu doit avoir une valeur pour chaque variable active.
3. Un temps court est une bonne performance ; une distance ou une hauteur grande est une bonne performance. Les noms en majuscules correspondent à l'une des deux compétitions.
4. Comparer les corrélations entre épreuves de course, entre lancers, et entre courses et sauts. Penser au sens des performances avant d'interpréter un signe.
5. Une valeur centrée réduite vaut $(x - \overline{x})/s$. Le 1500 m se mesure en secondes et varie d'une trentaine de secondes d'un athlète à l'autre.
6. Le total des points est calculé *à partir* des dix performances. `PCA` centre et réduit elle-même les variables actives quand `scale.unit = TRUE`.
7. L'inertie totale d'une ACP normée vaut le nombre de variables actives.
8. Pour une épreuve chronométrée, une flèche dirigée vers la gauche signifie que les athlètes situés à droite ont des temps *courts*. La qualité de représentation d'une variable sur le plan est la somme de ses $\cos^2$ sur les axes 1 et 2.
9. La contribution moyenne d'un athlète vaut $100/41$. Un athlète mal représenté peut être éloigné du centre dans d'autres dimensions.
10. Le classement est calculé *séparément* dans chaque compétition. Pour comparer les deux compétitions, regarder la valeur absolue de la valeur-test.

Corrections

Correction de l'exercice 1 ▲

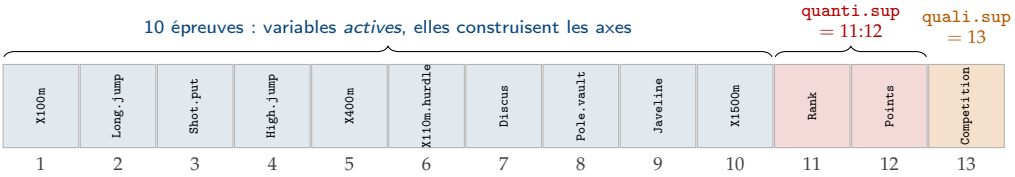
Comment lire cette correction

- Chaque réponse suit le même fil : **ce que R affiche** (tableau ou figure), **ce qu'on y lit**, puis **la conclusion pour l'entraîneur**.
- Les valeurs sont celles de `FactoMineR`. Selon l'ordinateur, le signe d'un axe peut être inversé (toute la carte est alors retournée en miroir), sans conséquence sur l'interprétation.
- Les figures dessinées ici reprennent les sorties de R, avec les mêmes valeurs, pour les rendre plus lisibles.

1. Préparer la séance

c) `str(deca)` affiche `'data.frame': 41 obs. of 13 variables` : dix colonnes `num` (les épreuves), deux colonnes `int` (`Rank` et `Points`) et `Competition : Factor w/ 2 levels "Decastar","OlympicG"`.

- En R, un nom de variable doit commencer par une lettre (ou un point) : `read.table` corrige automatiquement les noms invalides (`check.names = TRUE`) en ajoutant un `X` devant `100m`, `400m`, `110m.hurdle` et `1500m`.
- Le tableau compte **41 individus** (les athlètes, en noms de lignes) et **13 variables**, numérotées ainsi :



`Rank` et `Points` sont les colonnes 11 et 12, `Competition` la colonne 13 : ce sont ces numéros qu'on utilisera dans `PCA`.

2. Les valeurs manquantes

`colSums(is.na(deca))` renvoie 0 pour les treize colonnes et `mean(is.na(deca))` vaut 0 : **aucune valeur manquante**. C'est important car l'ACP calcule des distances entre individus à partir de *toutes* les variables actives : un individu incomplet ne peut pas être placé. Face à des cases vides, `PCA` les remplace sans prévenir par la moyenne de la colonne, ce qui rapproche artificiellement l'individu du centre ; il faut donc toujours savoir s'il y en a, et combien.

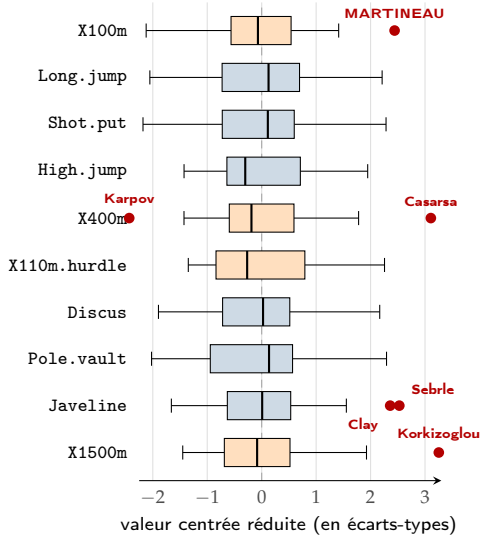
3. Décrire les données avant l'ACP

a) **Unités et sens des performances.**

	X100m	Long.jump	Shot.put	High.jump	X400m	X110m.hurdle	Discus	Pole.vault	Javeline	X1500m
Unité	s	m	m	m	s	s	m	m	m	s
Moyenne	11,00	7,26	14,48	1,98	49,62	14,61	44,33	4,76	58,32	279,0
Écart-type	0,26	0,32	0,82	0,09	1,15	0,47	3,38	0,28	4,83	11,67
Minimum	10,44	6,61	12,68	1,85	46,81	13,97	37,92	4,20	50,31	262,1
Maximum	11,64	7,96	16,36	2,15	53,20	15,67	51,65	5,40	70,52	317,0

Les unités diffèrent (secondes, mètres) et les écarts-types vont de 0,09 m (hauteur) à 11,7 s (1500 m). Pour les **quatre courses** (100 m, 400 m, 110 m haies, 1500 m), une *petite* valeur est une *bonne* performance ; pour les sauts et les lancers, c'est l'inverse. Il faudra s'en souvenir à chaque lecture d'un signe.

b) **Distributions.** Pour comparer les dix épreuves sur un même graphique, on a dessiné les boîtes à moustaches des valeurs *centrées réduites* (en orange, les épreuves chronométrées : à droite = plus lent) :

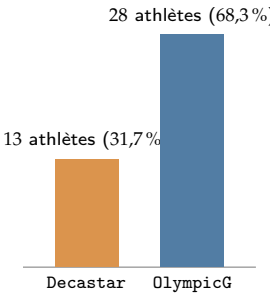


Les distributions sont à peu près symétriques (médiane au milieu de la boîte), avec une légère asymétrie à droite pour le 1500 m, le 400 m et le 110 m haies : quelques athlètes nettement plus lents que les autres. Six valeurs sont isolées :

- 100 m : MARTINEAU (11,64 s, le plus lent) ;
- 400 m : Karpov (46,81 s, le plus rapide) et Casarsa (53,20 s, le plus lent) ;
- javelot : Sebrle (70,52 m) et Clay (69,71 m), les deux meilleurs lanceurs ;
- 1500 m : Korkizoglou (317,0 s, soit plus de 15 s derrière tous les autres).

Ce ne sont **pas des erreurs de saisie** : ces performances sont tout à fait plausibles pour des décathloniens de haut niveau (un athlète blessé ou épuisé en fin de décathlon court facilement un 1500 m très lent). On **les garde** : ce sont de vraies informations sur les athlètes. Il faudra seulement vérifier qu'ils ne dominent pas un axe à eux seuls (contributions, question 9).

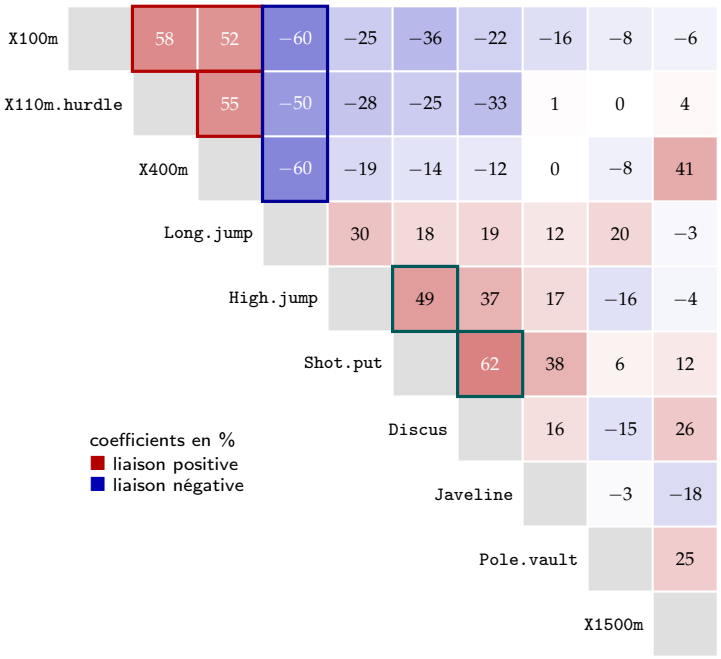
c) **Compétition.**



Les Jeux olympiques sont deux fois plus représentés que le Décastar. Les noms des athlètes du **Décastar** sont écrits en **majuscules** (`SEBRLE`), ceux des **Jeux olympiques** en minuscules (`Sebrle`). Neuf athlètes apparaissent deux fois : SEBRLE, CLAY, KARPOV, BERNARD, WARNERS, ZSIVOCZKY, HERNU, BARRAS et NOOL. Les 41 lignes ne sont donc pas 41 athlètes différents mais 41 *performances* : on en tirera parti à la question 10.

4. Les corrélations entre épreuves

La matrice est reproduite ci-dessous, les épreuves étant rangées par famille (courses, sauts, lancers, puis perche et 1500 m) ; l'ordre de `corrplot` avec `"hclust"` peut être légèrement différent, mais les blocs sont les mêmes.



a) **Trois familles d'épreuves.**

- **Les courses de vitesse** (encadrés rouges) : 100 m, 110 m haies et 400 m sont corrélés positivement entre eux (r de 0,52 à 0,58). Un athlète rapide l'est sur les trois distances.
- **Le saut en longueur** (encadré bleu) est corrélé *négativement* aux trois courses ($r \approx -0{,}60$, $-0{,}51$ et $-0{,}60$) : c'est en réalité une liaison *positive* entre vitesse et longueur (voir b).
- **Les lancers** (encadrés verts) : poids et disque sont fortement liés ($r = 0{,}62$), et le poids est aussi lié à la hauteur ($r = 0{,}49$). Ces épreuves demandent de la force et de l'explosivité.

Ces liaisons sont logiques : le sprint et la longueur reposent sur la vitesse de course, les lancers sur la force et la masse musculaire.

b) **Le signe piège.** Un *bon* sprinteur a un temps *petit* au 100 m et saute *loin* : quand le temps diminue, la longueur augmente, d'où une corrélation négative. Ici, $r = -0{,}60$ signifie donc que **les plus rapides sautent le plus loin**. Avec des variables de sens opposés, il faut toujours traduire le signe en « bonne » ou « mauvaise » performance avant de conclure.

c) **Les épreuves isolées.** Le saut à la perche (au plus $|r| = 0{,}25$, avec le 1500 m), le javelot (au plus 0,37, avec le poids) et le 1500 m (au plus 0,41, avec le 400 m) sont peu liés aux autres épreuves. Une ACP résume les liaisons *communes* à plusieurs variables : ces trois épreuves seront mal représentées sur les premiers axes, et l'on retrouvera leur information sur des axes suivants.

5. Centrer et réduire les données

a) R affiche pour les trois premiers athlètes (trois premières colonnes) :

	X100m	Long.jump	Shot.put
SEBRLE	0,16	1,01	0,43
CLAY	−0,91	0,44	−0,26
KARPOV	0,08	0,13	0,36

À la main, pour SEBRLE au 100 m :

$$z = \frac{x - \bar{x}}{s} = \frac{11{,}04 - 10{,}998}{0{,}263} = \frac{0{,}042}{0{,}263} \approx \mathbf{0{,}16}.$$

SEBRLE a couru 0,16 écart-type *au-dessus* de la moyenne : pour une épreuve chronométrée, une valeur positive signifie un temps *plus long*, donc une performance *moins bonne* que la moyenne (de peu). À l'inverse, CLAY ($-0{,}91$) a été nettement plus rapide que la moyenne.

b) **Pourquoi une ACP normée ?** Sans réduction, l'inertie totale serait la somme des variances, 173,5, et chaque épreuve y pèserait selon sa variance :

	100m	Long.	Poids	Haut.	400m	110m h.	Disque	Perche	Jav.	1500m
Variance	0,07	0,10	0,68	0,01	1,33	0,22	11,41	0,08	23,30	**136,27**
Part (%)	0,0	0,1	0,4	0,0	0,8	0,1	6,6	0,0	13,4	**78,6**

Le 1500 m ferait à lui seul 78,6 % de l'inertie, simplement parce qu'il se mesure en secondes sur une course longue : l'ACP deviendrait une « analyse du 1500 m », et le saut en hauteur (variance 0,008 m^2) ne compterait pour rien. L'ACP **normée** donne à chaque épreuve une variance égale à 1, donc le *même poids* (10 % de l'inertie) : c'est le seul choix sensé quand les unités diffèrent.

6. Lancer l'ACP

a) **Rôle des arguments.**

- `scale.unit = TRUE` : les variables actives sont centrées *et réduites* : c'est une ACP normée (c'est d'ailleurs le réglage par défaut) ;
- `quanti.sup = 11:12` : les colonnes 11 et 12 (`Rank`, `Points`) sont quantitatives *supplémentaires* : elles ne participent pas au calcul des axes, elles sont projetées sur le cercle ;
- `quali.sup = 13` : la colonne 13 (`Competition`) est qualitative supplémentaire : ses modalités sont placées au barycentre de leurs athlètes sur la carte ;
- `graph = FALSE` : pas de graphique automatique ; on les trace ensuite nous-mêmes. Toutes les autres colonnes (1 à 10) sont *actives*.

`names(res)` affiche sept éléments, `eig`, `var`, `ind`, `svd`, `quanti.sup`, `quali.sup` et `call` : valeurs propres, résultats des variables, des individus, décomposition en valeurs singulières, éléments supplémentaires, et paramètres de l'appel.

b) **Pourquoi `Points` et `Rank` en supplémentaires ?**

- `Points` est *calculé* à partir des dix performances (somme des points du barème) : ce n'est pas une information nouvelle. En variable active, il compterait une deuxième fois les mêmes épreuves et tirerait l'axe 1 vers lui. Surtout, la question est « *qu'est-ce qui fait gagner ?* » : il faut construire les axes avec les épreuves seules, puis regarder si le score s'aligne avec l'un d'eux. En active, la réponse serait biaisée d'avance.
- `Rank` est un rang (variable ordinale) calculé *dans chaque compétition* : il ne mesure pas une performance et ne se compare pas d'une compétition à l'autre.

c) **Et `data.CR` ?** Ce n'est pas un problème : avec `scale.unit = TRUE`, `PCA` centre et réduit elle-même les variables actives. Lui donner `data.CR` donnerait exactement les mêmes axes (la fonction `scale` divise par $n-1$ et `PCA` par n, mais réduire deux fois ne change rien aux corrélations), mais on perdrait `Rank`, `Points` et `Competition`, absents de `data.CR`. Il vaut donc mieux donner le tableau complet.

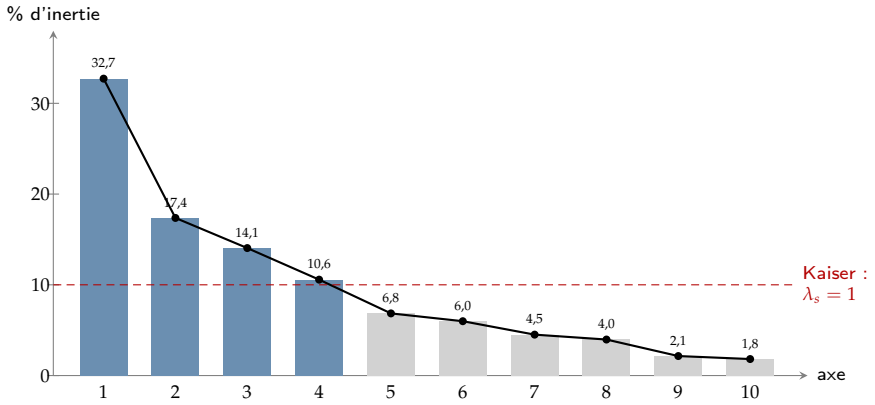
7. Combien d'axes ?

Axe	1	2	3	4	5	6	7	8	9	10
λ_s	**3,27**	**1,74**	**1,40**	**1,06**	0,68	0,60	0,45	0,40	0,21	0,18
% d'inertie	32,72	17,37	14,05	10,57	6,85	5,99	4,51	3,97	2,15	1,82
% cumulé	32,72	50,09	64,14	74,71	81,56	87,55	92,06	96,03	98,18	100

a) $\sum_s \lambda_s = 3{,}27 + 1{,}74 + \cdots + 0{,}18 = \mathbf{10}$: c'est l'inertie totale d'une ACP normée, égale au nombre de variables actives (10 épreuves de variance 1). Les valeurs propres « se partagent » ces 10 unités d'inertie.

b) Le plan $(1,2)$ restitue $32{,}72 + 17{,}37 = \mathbf{50{,}1}$ % de l'inertie : la carte ne montre que la moitié de l'information. C'est modeste, mais attendu : les corrélations entre épreuves sont moyennes (au plus 0,62) et trois épreuves sont presque indépendantes des autres (question 4). Quand les variables partagent peu d'information, aucun axe ne peut en résumer beaucoup à lui seul. Il faudra donc contrôler la qualité de représentation avant d'interpréter un point.

c) **Nombre d'axes.**



- *Kaiser* : λ_1 à λ_4 dépassent 1 (barres bleues), soit **4 axes** et 74,7 % de l'inertie.
- *Coude* : la plus forte chute a lieu entre les axes 1 et 2 ; ensuite les barres décroissent régulièrement jusqu'à l'axe 4, puis se stabilisent (de 6,9 à 6,0 %) : on peut placer le coude après l'axe 4.

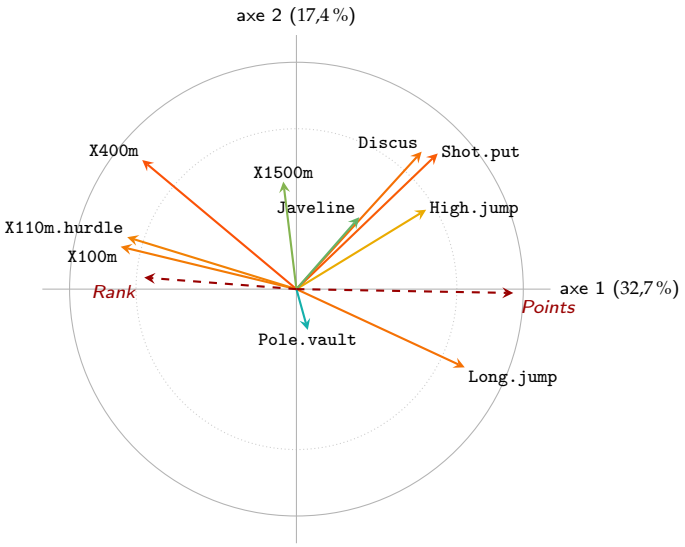
En pratique, on interprète en détail le **plan** $(1,2)$, puis on jette un œil aux axes 3 et 4 pour les épreuves mal représentées (perche, 1500 m, javelot).

8. Le cercle des corrélations

a) **Résultats des variables** (en gras : contributions supérieures à la moyenne de 10 %) :

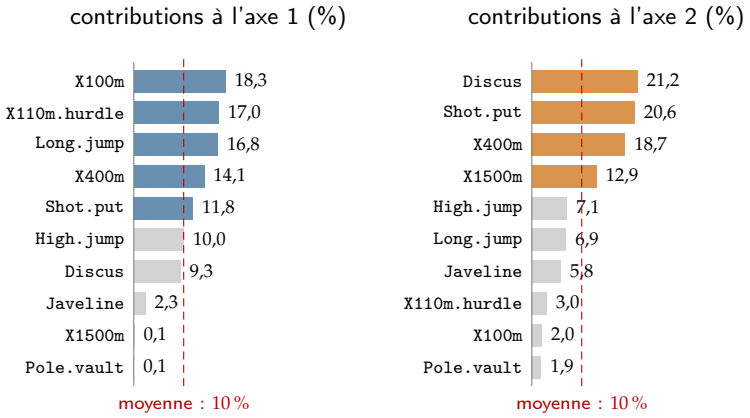
	coordonnée $r(X_k, F_s)$			$\cos^2$			contribution (%)		
	axe 1	axe 2	axe 3	axe 1	axe 2	axe 3	axe 1	axe 2	axe 3
`X100m`	−0,77	0,19	−0,18	0,60	0,04	0,03	18,3	2,0	2,4
`Long.jump`	0,74	−0,35	0,18	0,55	0,12	0,03	16,8	6,9	2,4
`Shot.put`	0,62	0,60	−0,02	0,39	0,36	0,00	11,8	20,6	0,0
`High.jump`	0,57	0,35	−0,26	0,33	0,12	0,07	10,0	7,1	4,8
`X400m`	−0,68	0,57	0,13	0,46	0,32	0,02	14,1	18,7	1,2
`X110m.hurdle`	−0,75	0,23	−0,09	0,56	0,05	0,01	17,0	3,0	0,6
`Discus`	0,55	0,61	0,04	0,31	0,37	0,00	9,3	21,2	0,1
`Pole.vault`	0,05	−0,18	0,69	0,00	0,03	0,48	0,1	1,9	34,1
`Javeline`	0,28	0,32	−0,39	0,08	0,10	0,15	2,3	5,8	10,8
`X1500m`	−0,06	0,47	0,78	0,00	0,22	0,61	0,1	12,9	43,5

Le cercle, avec la couleur de chaque flèche selon sa qualité de représentation sur le plan (les variables supplémentaires sont en pointillés) :





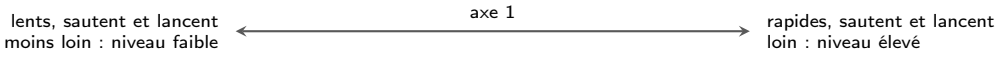
b) **Contributions et qualité de représentation.**



c) **Axe 1 (32,7 %) : la vitesse et la performance d'ensemble.**

- À gauche : les trois courses rapides, `X100m` ($r = -0{,}77$), `X110m.hurdle` ($-0{,}75$) et `X400m` ($-0{,}68$). Ce sont des temps : les athlètes situés à *droite* ont des temps *courts*, donc sont *rapides*.
- À droite : `Long.jump` (0,74), puis `Shot.put` (0,62), `High.jump` (0,57) et `Discus` (0,55) : les athlètes de droite sautent haut et loin et lancent loin.
- Les six premières épreuves fournissent 88 % de l'axe : 100 m (18,3 %), 110 m haies (17,0 %), longueur (16,8 %), 400 m (14,1 %), poids (11,8 %), hauteur (10,0 %).

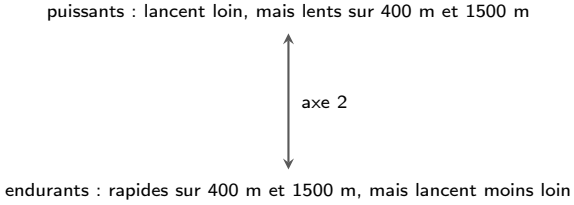
Une fois traduit le sens des temps, toutes les flèches bien représentées pointent vers la droite pour une *bonne* performance. L'axe 1 est un axe de « **niveau général** », dominé par la **vitesse** : à droite, les athlètes rapides et performants partout ; à gauche, ceux qui sont en dessous de la moyenne dans la plupart des épreuves. La variable supplémentaire `Points`, presque confondue avec l'axe ($r = 0{,}96$), le confirme (question 10).



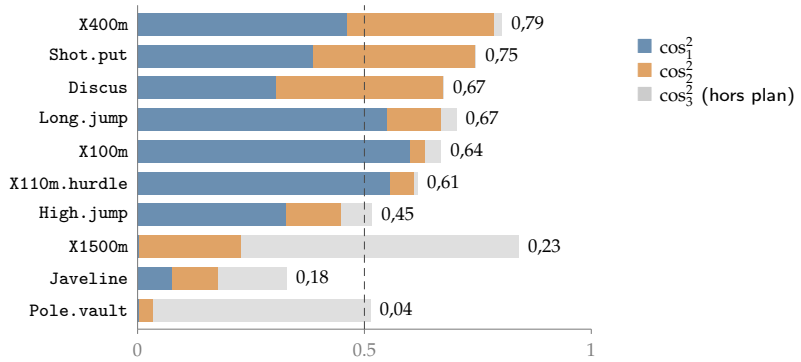
d) **Axe 2 (17,4 %) : la force contre l'endurance.**

- En haut : `Discus` ($r = 0{,}61$), `Shot.put` (0,60), mais aussi `X400m` (0,57) et `X1500m` (0,47).
- Ces quatre épreuves fournissent 73 % de l'axe : disque (21,2 %), poids (20,6 %), 400 m (18,7 %), 1500 m (13,0 %).

Pour le 400 m et le 1500 m, une forte valeur est un temps *long*. En haut se trouvent donc les athlètes **puissants** (bons lanceurs) mais **moins endurants** (lents sur 400 m et 1500 m) ; en bas, les athlètes plus **légers et endurants**, à l'aise sur les courses longues mais qui lancent moins loin. L'axe 2 oppose deux qualités difficiles à réunir : **la force et l'endurance**.



e) **Épreuves mal représentées.**



`Pole.vault` ($\cos^2 = 0{,}03$), `Javeline` (0,18) et `X1500m` (0,23) sont mal représentés sur le plan : leurs flèches sont courtes, on ne les interprète pas sur ce plan (le 1500 m n'est utilisé ci-dessus que comme appoint de l'axe 2). Leur information est ailleurs :

- sur l'**axe 3** pour la perche ($r = 0{,}69$, $\cos^2_3 = 0{,}48$) et le 1500 m ($r = 0{,}78$, $\cos^2_3 = 0{,}61$), qui fournissent à eux deux 78 % de l'axe 3 ;
- sur l'**axe 4** pour le javelot ($r = 0{,}71$, $\cos^2_4 = 0{,}51$).

Ce sont les épreuves « isolées » repérées à la question 4 : des épreuves techniques ou d'endurance, qui forment chacune une dimension à part.

f) **Lien avec le cours (`X100m`).** R donne $r(\texttt{X100m}, F_1) = -0{,}7747$ et $\lambda_1 = 3{,}272$.

$$\cos^2_1 = r^2 = (-0{,}7747)^2 = \mathbf{0{,}600} \qquad \text{contrib}_1 = 100 \times \frac{r^2}{\lambda_1} = 100 \times \frac{0{,}600}{3{,}272} = \mathbf{18{,}3}\,\%,$$

ce sont bien les valeurs affichées par `res$var$cos2` et `res$var$contrib`. On vérifie aussi que

$$\textstyle\sum_k r^2(X_k, F_1) = 0{,}600 + 0{,}550 + \cdots + 0{,}003 = 3{,}272 = \lambda_1,$$

c'est pourquoi les contributions d'un axe font 100 %.

g) **Retour à la matrice des corrélations et conseils.** Le cercle retrouve les trois familles de la question 4 : les trois courses de vitesse groupées à gauche, la longueur à l'opposé (angle proche de 180°), les lancers groupés en haut à droite. Pour deux flèches bien représentées, le produit scalaire approche la corrélation : pour le 100 m et la longueur,

$$(-0{,}775)(0{,}742) + (0{,}187)(-0{,}345) = -0{,}639 \quad \text{contre} \quad r = -0{,}60$$

(l'écart vient des 50 % d'information hors du plan). Pour l'entraîneur :

- travailler la **vitesse** est le meilleur investissement : elle améliore d'un coup le 100 m, le 110 m haies, le 400 m et la longueur, les épreuves qui « fabriquent » l'axe 1, donc le score ;
- la **force** (poids, disque) est un second bloc, qui se paie souvent par une moins bonne endurance : il faut doser musculation et travail foncier ;
- la **perche** et le **javelot**, peu liés aux autres, demandent un travail technique spécifique : ce sont des épreuves où un athlète peut gagner des points sans que le reste de son profil n'en soit affecté.

9. La carte des athlètes

a) **Carte des athlètes** (orange : Décastar ; bleu : Jeux olympiques ; seuls quelques noms sont écrits) :



b) **Meilleurs et moins bons profils.**

- À droite : **Karpov** ($F_1 = 4{,}62$), **Sebrle** (4,04) et **Clay** (3,92), qui sont justement les trois médaillés des Jeux olympiques (rangs 3, 1 et 2). Puis Macey (2,23) et Warners (2,17).
- À gauche : BOURGUIGNON ($-3{,}98$, dernier du Décastar), Casarsa ($-2{,}86$, dernier des Jeux) et Uldal ($-2{,}56$, avant-dernier).

Contributions les plus fortes (moyenne : $100/41 \approx 2{,}4\,\%$) :

axe 1		axe 2	
Karpov	15,9 %	Casarsa	**20,3** %
Sebrle	12,2 %	Drews	13,3 %
BOURGUIGNON	11,8 %	YURKOV	6,4 %
Clay	11,4 %	Parkhomenko	6,2 %
Casarsa	6,1 %	Korkizoglou	6,0 %

Quatre athlètes font la moitié de l'axe 1, et ils sont aux deux extrémités : l'axe 1 oppose les meilleurs aux moins bons. Sur l'axe 2, **Casarsa** fournit à lui seul un cinquième de l'axe (huit fois la moyenne) : sa position atypique (en haut à gauche) oriente fortement l'axe 2. Ce n'est pas une erreur, mais il faut le garder en tête ; on pourrait vérifier la stabilité de l'axe 2 en refaisant l'ACP avec Casarsa en individu supplémentaire (argument `ind.sup`).

Vérification pour Karpov ($F_1 = 4{,}620$, $d(i,G) = 5{,}006$, $I = 41$) :

$$\text{contrib}_1 = 100 \times \frac{F_1^2}{I\,\lambda_1} = 100 \times \frac{4{,}620^2}{41 \times 3{,}272} = 100 \times \frac{21{,}34}{134{,}2} = \mathbf{15{,}9}\,\%$$

$$\cos_1^2 = \frac{F_1^2}{d^2(i,G)} = \frac{21{,}34}{25{,}06} = \mathbf{0{,}85}.$$

c) **L'athlète le plus mal représenté** est **Schoenbeck** : $\cos_1^2 + \cos_2^2 = 0{,}005$ (puis Barras, 0,03). Il est tout près du centre de la carte ($F_1 = 0{,}11$, $F_2 = -0{,}04$), mais on **ne peut pas** en conclure qu'il est « moyen partout » : sa distance au centre vaut 1,80, et presque toute cette distance est sur les axes 3 et 4 ($F_3 = 0{,}74$, $F_4 = 0{,}93$), ceux de la perche, du 1500 m et du javelot. Il saute d'ailleurs 5,00 m à la perche et lance le javelot à 60,9 m, au-dessus des moyennes (4,76 m et 58,3 m). Un point proche du centre *de la carte* peut être loin du centre *du nuage*.

d) **Retour aux données brutes** (moyennes en dernière ligne ; en gras, les performances meilleures que la moyenne) :

	100m	Long.	Poids	Haut.	400m	110m h.	Disque	Perche	Jav.	1500m
Karpov	**10,50**	**7,81**	**15,93**	**2,09**	**46,81**	**13,97**	**51,65**	4,60	55,54	**278,1**
Casarsa	11,36	6,68	**14,92**	1,94	53,20	15,39	**48,66**	4,40	**58,62**	296,1
Drews	**10,87**	**7,38**	13,07	1,88	**48,51**	**14,01**	40,11	**5,00**	51,53	**274,2**
Moyenne	11,00	7,26	14,48	1,98	49,62	14,61	44,33	4,76	58,32	279,0

- **Karpov** (tout à droite) est meilleur que la moyenne dans huit épreuves sur dix, avec le meilleur 400 m, le meilleur 110 m haies et le meilleur disque : le profil type de la droite de l'axe 1.
- **Casarsa** (en haut à gauche) est lent (le plus mauvais 400 m, 110 m haies et longueur faibles) mais bon lanceur (poids, disque) : lent donc à gauche, puissant mais peu endurant donc en haut.
- **Drews** (en bas) est rapide et endurant (400 m, 1500 m), mais lance peu loin (poids, disque, javelot tous faibles) : c'est le profil « endurant » du bas de l'axe 2.

Les données brutes confirment les noms donnés aux axes.

e) **Le biplot** superpose les athlètes bien représentés et les flèches bien représentées. Un athlète situé dans la direction d'une flèche a une valeur *supérieure* à la moyenne pour cette variable ; dans la direction opposée, une valeur inférieure. Ainsi Karpov, dans la direction opposée aux flèches des courses, a des temps courts ; Casarsa, du côté des flèches `X400m` et `Discus`, est lent au 400 m et lance loin le disque. Attention : cette lecture ne vaut que pour des points *et* des flèches bien représentés, ce que garantit la sélection `cos2 = 0.5`.

10. Les variables supplémentaires

a) **Points et classement.**

	$r(\cdot, F_1)$	$r(\cdot, F_2)$	$\cos^2_1$	$\cos^2_2$
`Points`	0,96	−0,02	0,91	0,00
`Rank`	−0,67	0,05	0,45	0,00

La flèche de `Points` est presque confondue avec l'axe 1 et presque de longueur 1 : l'axe 1, construit *sans* le score, explique à lui seul 91 % de la variance des points. L'axe 1 est donc bien l'axe de la **performance globale**, et la vitesse (qui le construit) est ce qui fait gagner. Le score ne dépend presque pas de l'axe 2 : être plutôt « puissant » ou plutôt « endurant » ne change pas le total, ce sont deux façons d'arriver au même score.

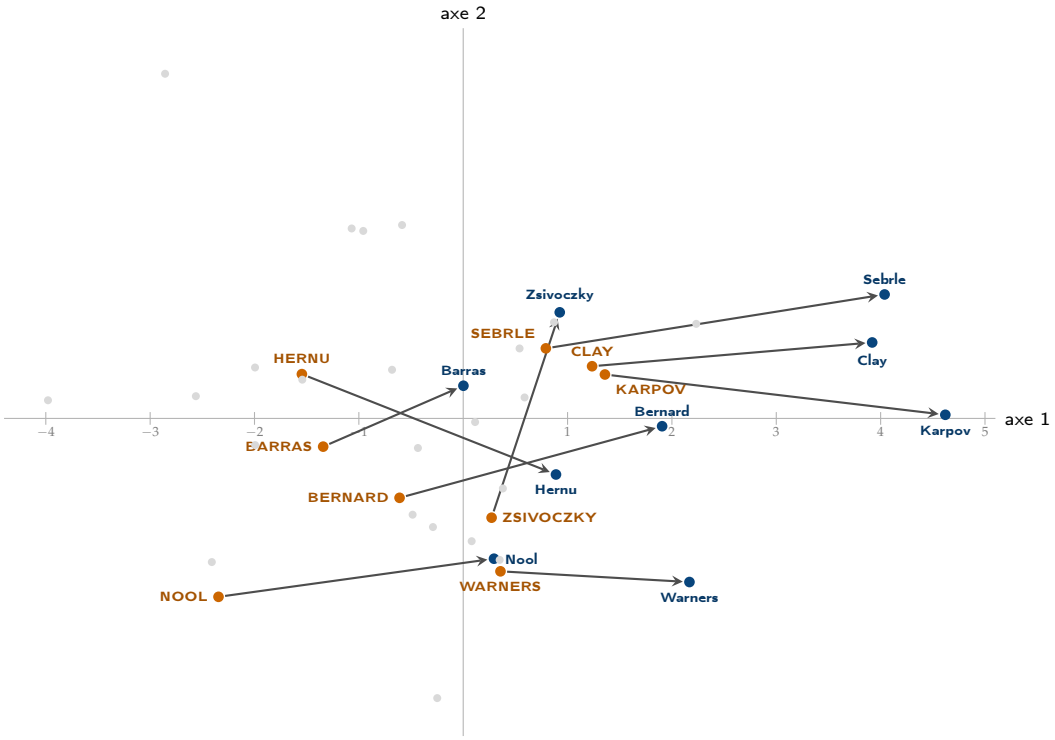
`Rank` est corrélé négativement à l'axe 1 (le rang 1 est le meilleur), mais plus faiblement ($-0{,}67$), pour deux raisons : c'est un rang, qui écrase les écarts de points ; et surtout il est calculé *dans chaque compétition*. SEBRLE, 1er du Décastar avec 8217 points, aurait été 10e aux Jeux olympiques ; Casarsa, 28e et dernier des Jeux avec 7404 points, aurait été 13e au Décastar.

b) **Les deux compétitions.**

	coordonnées		valeurs-test	
	axe 1	axe 2	axe 1	axe 2
Decastar	−0,60	−0,04	−1,43	−0,12
OlympicG	0,28	0,02	1,43	0,12

Le barycentre des Jeux olympiques est un peu à droite, celui du Décastar un peu à gauche (les petits carrés blancs de la carte de la question 9) : les performances olympiques sont en moyenne meilleures (8052 points de moyenne contre 7906). Mais $|\texttt{v.test}| = 1{,}43 < 1{,}96$: l'écart **n'est pas significatif**. Les ellipses des deux groupes se recouvrent largement : la différence entre compétitions est petite devant les différences entre athlètes. Sur l'axe 2, aucune différence.

c) **Les neuf athlètes présents deux fois.** On relie chaque performance au Décastar (majuscules) à celle du même athlète aux Jeux olympiques :



	Sebrle	Clay	Karpov	Bernard	Warners	Zsivoczky	Hernu	Barras	Nool
Décastar	8217	8122	8099	8067	8030	8004	7733	7708	7651
Jeux	8893	8820	8725	8225	8343	8287	8237	8067	8235
Gain	+676	+698	+626	+158	+313	+283	+504	+359	+584

Les **neuf** flèches vont vers la droite : chaque athlète a fait un meilleur décathlon aux Jeux qu'au Décastar (de 158 à 698 points de plus). Si la compétition n'avait aucun effet, chaque athlète aurait une chance sur deux de progresser, et la probabilité que les neuf progressent serait $(1/2)^9 \approx 0{,}002$: l'effet de la compétition est en réalité très net.

Cela **nuance** la conclusion de b) : la valeur-test compare deux groupes d'athlètes *en partie différents*, dont le niveau varie beaucoup ; ces écarts de niveau « noient » l'effet de la compétition. En comparant chaque athlète à lui-même, on élimine ces écarts et l'effet apparaît (enjeu olympique, préparation ciblée sur les Jeux, conditions de compétition). Une absence de significativité ne prouve donc pas une absence d'effet.

d) **Description automatique des axes.** `dimdesc` donne, au seuil de 5 % :

`$Dim.1$quanti`		`$Dim.2$quanti`	
Points	0,96	`Discus`	0,61
`Long.jump`	0,74	`Shot.put`	0,60
`Shot.put`	0,62	`X400m`	0,57
`High.jump`	0,57	`X1500m`	0,47
`Discus`	0,55	`High.jump`	0,35
Rank	−0,67	`Javeline`	0,32
`X400m`	−0,68	`Long.jump`	−0,35
`X110m.hurdle`	−0,75		
`X100m`	−0,77		

On retrouve exactement les interprétations de la question 8 : l'axe 1 est décrit d'abord par les *Points*, puis par la longueur (à droite) et les courses rapides (à gauche) ; l'axe 2 par les lancers et les courses longues. La compétition n'apparaît pas (pas de rubrique `$quali` ni `$category`) : son lien avec les axes n'est significatif ni sur l'axe 1 ($R^2 = 0{,}05$, $p = 0{,}16$), ni sur l'axe 2, ce que disaient déjà les valeurs-test.

11. Synthèse

Exemple de synthèse. L'ACP normée des dix épreuves de 41 décathlons (2004) fait apparaître trois familles d'épreuves : les courses de vitesse et le saut en longueur, les lancers, et des épreuves plus isolées (perche, javelot, 1500 m). Le premier axe (33 % de l'inertie) est un axe de **niveau général dominé par la vitesse** : il est presque confondu avec le total des points ($r = 0{,}96$). Les médaillés olympiques Karpov, Sebrle et Clay y sont en tête, BOURGUIGNON et Casarsa en queue. Le deuxième axe (17 %) oppose les athlètes **puissants** (lancers) mais peu endurants aux athlètes **endurants** (400 m, 1500 m) qui lancent moins loin ; il est sans lien avec le score : ce sont deux manières d'atteindre un même total. Les neuf athlètes présents aux deux compétitions ont tous progressé aux Jeux olympiques, même si l'écart moyen entre les deux compétitions n'est pas significatif. *Pistes* : la vitesse est la qualité la plus rentable ; la perche et le javelot, indépendants du reste, sont des réserves de points par le travail technique. *Limites* : le plan ne restitue que 50 % de l'information, l'axe 2 dépend beaucoup d'un athlète (Casarsa), les 41 lignes ne sont pas 41 athlètes indépendants, et l'échantillon (deux compétitions de 2004) est petit : ces conclusions décrivent ces données, elles ne se généralisent pas sans précaution.

Ecole supérieure d'agro-développement international

istom



Étudiant(e) (Nom, Prénom) : .....

Promotion, groupe : .....

Email : .....