TD4 : Performance des athlètes au décathlon

(Correction)

```
> require(ade4)
> options(width=100)
> deca=read.table("decathlon2.dat")
> deca[1:10]=signif(deca[1:10],3)
> source("fonctions.R")
> deca1=deca[,1:10]
> pca1=dudi.pca(deca1, scannf=F, nf=3)
> pca1 = dudi.fixsigns(pca1, sign.co=c(1,-1,1))
> inert1=inertia.dudi(pca1,ro=T,co=T)
> keeprowabs = inert1$row.abs[1,1]
> #inert1$row.abs[1,1]=NA
> inert1$row.rel = abs(inert1$row.rel)
> colnames(inert1$row.abs)=paste0("Axis", 1:3)
> colnames(inert1$row.cum)=c("Ax1","Ax1:2", "Ax1:3")
```

1 Performances des athlètes au décathlon

On étudie les performances des athlètes ayant participé en 2004 au épreuves de décathlon des Jeux Olympiques et du Décastar, à partir des données ci-dessous.

```
> deca
```

	c100	long	poids	haut	c400	c110	disq	perche	javel	c1500	RANG	POINTS	COMPET
SEBRLE	11.0	7.58	14.8	2.07	49.8	14.7	43.8	5.02	63.2	292	1	8217	DS
CLAY	10.8	7.40	14.3	1.86	49.4	14.0	50.7	4.92	60.2	302	2	8122	DS
KARPOV	11.0	7.30	14.8	2.04	48.4	14.1	49.0	4.92	50.3	300	3	8099	DS
BERNARD	11.0	7.23	14.2	1.92	48.9	15.0	40.9	5.32	62.8	280	4	8067	DS
YURKOV	11.3	7.09	15.2	2.10	50.4	15.3	46.3	4.72	63.4	276	5	8036	DS
WARNERS	11.1	7.60	14.3	1.98	48.7	14.2	41.1	4.92	51.8	278	6	8030	DS
ZSIVOCZKY	11.1	7.30	13.5	2.01	48.6	14.2	45.7	4.42	55.4	268	7	8004	DS
McMULLEN	10.8	7.31	13.8	2.13	49.9	14.4	44.4	4.42	56.4	285	8	7995	DS
MARTINEAU	11.6	6.81	14.6	1.95	50.1	14.9	47.6	4.92	52.3	262	9	7802	DS
HERNU	11.4	7.56	14.4	1.86	51.1	15.1	45.0	4.82	57.2	285	10	7733	DS
BARRAS	11.3	6.97	14.1	1.95	49.5	14.5	42.1	4.72	55.4	282	11	7708	DS
NOOL	11.3	7.27	12.7	1.98	49.2	15.3	37.9	4.62	57.4	267	12	7651	DS
BOURGUIGNON	11.4	6.80	13.5	1.86	51.2	15.7	40.5	5.02	54.7	292	13	7313	DS
Sebrle	10.8	7.84	16.4	2.12	48.4	14.0	48.7	5.00	70.5	280	1	8893	JO
Clay	10.4	7.96	15.2	2.06	49.2	14.1	50.1	4.90	69.7	282	2	8820	JO
Karpov	10.5	7.81	15.9	2.09	46.8	14.0	51.6	4.60	55.5	278	3	8725	JO
Macey	10.9	7.47	15.7	2.15	49.0	14.6	48.3	4.40	58.5	265	4	8414	JO
Warners	10.6	7.74	14.5	1.97	48.0	14.0	43.7	4.90	55.4	278	5	8343	JO
Zsivoczky	10.9	7.14	15.3	2.12	49.4	15.0	45.6	4.70	63.4	270	6	8287	JO
Hernu	11.0	7.19	14.6	2.03	48.7	14.2	44.7	4.80	57.8	264	7	8237	JO
Nool	10.8	7.53	14.3	1.88	48.8	14.8	42.0	5.40	61.3	276	8	8235	JO
Bernard	10.7	7.48	14.8	2.12	49.1	14.2	44.8	4.40	55.3	276	9	8225	JO
Schwarzl	11.0	7.49	14.0	1.94	49.8	14.2	42.4	5.10	56.3	274	10	8102	JO
Pogorelov	11.0	7.31	15.1	2.06	50.8	14.2	44.6	5.00	53.4	288	11	8084	JO
Schoenbeck	10.9	7.30	14.8	1.88	50.3	14.3	44.4	5.00	60.9	279	12	8077	JO
Barras	11.1	6.99	14.9	1.94	49.4	14.4	44.8	4.60	64.6	267	13	8067	JO
Smith	10.8	6.81	15.2	1.91	49.3	14.0	49.0	4.20	61.5	273	14	8023	JO
Averyanov	10.6	7.34	14.4	1.94	49.7	14.4	39.9	4.80	54.5	271	15	8021	JO
Ojaniemi	10.7	7.50	15.0	1.94	49.1	15.0	40.4	4.60	59.3	276	16	8006	JO
Smirnov	10.9	7.07	13.9	1.94	49.1	14.8	42.5	4.70	60.9	263	17	7993	JO
Qi	11.1	7.34	13.6	1.97	49.6	14.8	45.1	4.50	60.8	273	18	7934	JO
Drews	10.9	7.38	13.1	1.88	48.5	14.0	40.1	5.00	51.5	274	19	7926	JO
Parkhomenko	11.1	6.61	15.7	2.03	51.0	14.9	41.9	4.80	65.8	278	20	7918	JO
Terek	10.9	6.94	15.2	1.94	49.6	15.1	45.6	5.30	50.6	290	21	7893	JO
Gomez	11.1	7.26	14.6	1.85	48.6	14.4	41.0	4.40	60.7	270	22	7865	JO
Turi	11.1	6.91	13.6	2.03	51.7	14.3	39.8	4.80	59.3	290	23	7708	JO
Lorenzo	11.1	7.03	13.2	1.85	49.3	15.4	40.2	4.50	58.4	263	24	7592	JO
Karlivans	11.3	7.26	13.3	1.97	50.5	15.0	43.3	4.50	52.9	279	25	7583	JO
Korkizoglou	10.9	7.07	14.8	1.94	51.2	15.0	46.1	4.70	53.0	317	26	7573	JO
Uldal	11.2	6.99	13.5	1.85	51.0	15.1	43.0	4.50	60.0	282	27	7495	JO
Casarsa	11.4	6.68	14.9	1.94	53.2	15.4	48.7	4.40	58.6	296	28	7404	JO

Les dix épreuves du décathlon :
- course sur 100 m (`c100`),
- saut en longueur (`long`),
- lancer de poids (`poids`),
- saut en hauteur (`haut`),
- course sur 400 m (`c400`),
- course de haies sur 110 m (`c110`),
- lancer de disque (`disq`),
- saut à la perche (`perch`),
- lancer de javelot (`javel`),
- course sur 1500 m (`c1500`)

Les performances de course sont mesurées en secondes, les autres en mètres.

Autres variables
- rang de classement (`RANG`),
- nombre de points (`POINTS`),
- compétition (`COMPET`),
  - Jeux Olympiques (`JO`),
  - Décastar (`DS`)

Attention ! Les noms des participants sont en majuscule pour le Décastar, afin de permettre de différencier les participations d'un même athlète aux deux épreuves.
(exemple : `SERBLE`/`Serble`).

1.1 Analyse rapide des variables

On donne ci-dessous la matrice de corrélation des variables quantitatives.

```
> round(cor(deca[,1:12]), 2)

        c100  long poids  haut  c400  c110  disq perche javel c1500  RANG POINTS
c100    1.00 -0.61 -0.38 -0.28  0.55  0.59 -0.24  -0.08 -0.20 -0.03  0.33  -0.72
long   -0.61  1.00  0.18  0.29 -0.60 -0.51  0.19   0.20  0.12 -0.03 -0.60   0.73
poids  -0.38  0.18  1.00  0.49 -0.14 -0.26  0.62   0.07  0.37  0.11 -0.37   0.63
haut   -0.28  0.29  0.49  1.00 -0.19 -0.27  0.37  -0.16  0.17 -0.05 -0.49   0.58
c400    0.55 -0.60 -0.14 -0.19  1.00  0.54 -0.11  -0.07  0.00  0.42  0.56  -0.66
c110    0.59 -0.51 -0.26 -0.27  0.54  1.00 -0.33  -0.03  0.00  0.04  0.45  -0.65
disq   -0.24  0.19  0.62  0.37 -0.11 -0.33  1.00  -0.15  0.16  0.26 -0.39   0.48
perche -0.08  0.20  0.07 -0.16 -0.07 -0.03 -0.15   1.00 -0.03  0.24 -0.32   0.20
javel  -0.20  0.12  0.37  0.17  0.00  0.00  0.16  -0.03  1.00 -0.18 -0.21   0.42
c1500  -0.03 -0.03  0.11 -0.05  0.42  0.04  0.26   0.24 -0.18  1.00  0.09  -0.20
RANG    0.33 -0.60 -0.37 -0.49  0.56  0.45 -0.39  -0.32 -0.21  0.09  1.00  -0.74
POINTS -0.72  0.73  0.63  0.58 -0.66 -0.65  0.48   0.20  0.42 -0.20 -0.74   1.00
```

Question 1 *Quelles sont les couples de variables les plus corrélées, les moins corrélées, les plus opposées ?*

— variables les plus corrélées (r proche de 1) : $\text{cor}(\texttt{POINTS}, \texttt{long}) = 0,73$
— variables les moins corrélées (r proche de 0) : $\text{cor}(\texttt{javel}, \texttt{c110}) = \text{cor}(\texttt{javel}, \texttt{c400}) = 0$
— variables les plus opposées (r proche de -1) : $\text{cor}(\texttt{POINTS}, \texttt{RANG}) = -0,74$

Question 2 *Comment se groupent les variables du point de vue des signes de corrélation ? Expliquez pourquoi.*

Les performances de courses sont dans l'ensemble corrélées négativement avec les performances de lancer et de saut. C'est normal puisqu'une bonne performance à la course est un petit temps, alors qu'une bonne performance au lancer ou au saut est une grande longueur.

De même, les points sont corrélés positivement aux performances de saut et de lancer, et les rangs sont corrélés négativement (un petit rang correspond à un grand nombre de points).

Certains couples ne vérifient pas cette règle :
— (`c100`, `c1500`) avec $-0,03$,
— (`poids`, `c1500`) avec $0,11$,
— (`disq`, `c1500`) avec $0,26$,
— (`perche`, `c1500`) avec $0,24$.
— (`haut`, `perche`) avec $-0,16$,
— (`disq`, `perche`) avec $-0,15$,
— (`javel`, `perche`) avec $-0,03$,

Il s'agit donc principalement de corrélations avec `c1500` ou `perche`. Les valeurs des corrélations en jeu sont toutefois peu importantes (au maximum $0,26$). Ces deux variables sont les moins corrélées avec `POINTS` : les performances des athlètes en ces épreuves sont peu corrélées avec leur performance totale.

1.2 Analyse des composantes principales

On procède à une analyse en composantes principales des performances centrées-réduites, en mettant de coté pour l'instant les variables `RANG`, `POINTS` et `COMPET`. On donne ci-après les parts d'inertie suivantes associées aux 5 premiers axes, puis, pour les trois premiers axes seulement, les corrélations des variables (actives et supplémentaires), les coordonnées des individus, leurs contributions aux axes (en %) et leurs qualités de représentation par les premiers espaces principaux (en %).

Décomposition de l'inertie

```
> round(pca1$eig,2)[1:5]

[1] 3.32 1.74 1.40 1.05 0.68
```

Corrélation variables / axes

```
> round(pca1$co, 2)

       Comp1 Comp2 Comp3
c100    0.80 -0.16  0.15
long   -0.74  0.34 -0.21
poids  -0.63 -0.59  0.03
haut   -0.57 -0.34  0.27
c400    0.68 -0.58 -0.11
c110    0.74 -0.22  0.13
disq   -0.55 -0.61 -0.03
perche -0.05  0.16 -0.71
javel  -0.29 -0.29  0.39
c1500   0.07 -0.51 -0.75

> round(supvar1$cosup,2)

       Comp1 Comp2 Comp3
RANG    0.67 -0.05  0.08
POINTS -0.95  0.02  0.05
```

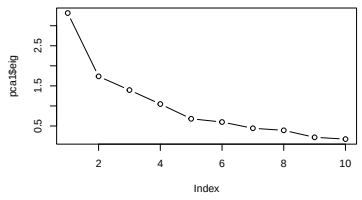
```
> round(pca1$li, 2)

            Axis1 Axis2 Axis3
SEBRLE      -0.84 -0.77 -0.85
CLAY        -1.19 -0.72 -2.13
KARPOV      -1.34 -0.59 -1.96
BERNARD      0.57  0.92 -0.93
YURKOV       0.49 -2.02  1.29
WARNERS     -0.36  1.67 -0.88
ZSIVOCZKY   -0.29  1.13  1.21
McMULLEN    -0.63 -0.24  0.47
MARTINEAU    1.90 -0.47  0.66
HERNU        1.65 -0.53 -0.82
BARRAS       1.34  0.32 -0.05
NOOL         2.29  2.03  1.23
BOURGUIGNON  4.06 -0.30 -1.27
Sebrle      -4.16 -1.33  0.23
Clay        -4.01 -0.81 -0.23
Karpov      -4.56 -0.05  0.05
Macey       -2.16 -0.98  1.93
Warners     -2.20  1.75 -0.93
Zsivoczky   -0.92 -1.12  1.51
Hernu       -0.87  0.71  0.87
Nool        -0.34  1.52 -1.40
Bernard     -1.87  0.08  0.81
Schwarzl    -0.06  1.32 -0.94
Pogorelov   -0.43 -0.86 -1.34
Schoenbeck  -0.16  0.01 -0.77
Barras      -0.05 -0.27  1.55
Smith       -0.93 -0.99  1.65
Averyanov   -0.27  1.53 -0.27
Ojaniemi    -0.41  0.75  0.39
Smirnov      0.48  1.10  1.24
Qi           0.48  0.32  1.08
Drews        0.29  3.02 -1.15
Parkhomenko  0.98 -2.01  1.07
Terek        0.62 -0.62 -2.15
Gomez        0.30  1.21  1.24
Turi         1.65 -0.46 -0.47
Lorenzo      2.39  1.67  1.51
Karlivans    1.97  0.33  0.31
Korkizoglou  1.09 -2.23 -2.40
Uldal        2.55 -0.22  0.43
Casarsa      2.94 -3.81  0.18
```

```
> round(inert1$row.abs,1)

            Axis1 Axis2 Axis3
SEBRLE        0.5   0.8   1.3
CLAY          1.0   0.7   7.9
KARPOV        1.3   0.5   6.7
BERNARD       0.2   1.2   1.5
YURKOV        0.2   5.7   2.9
WARNERS       0.1   3.9   1.4
ZSIVOCZKY     0.1   1.8   2.6
McMULLEN      0.3   0.1   0.4
MARTINEAU     2.7   0.3   0.8
HERNU         2.0   0.4   1.2
BARRAS        1.3   0.1   0.0
NOOL          3.9   5.8   2.7
BOURGUIGNON  12.1   0.1   2.8
Sebrle       12.7   2.5   0.1
Clay         11.9   0.9   0.1
Karpov       15.3   0.0   0.0
Macey         3.4   1.3   6.5
Warners       3.6   4.3   1.5
Zsivoczky     0.6   1.8   4.0
Hernu         0.6   0.7   1.3
Nool          0.1   3.2   3.4
Bernard       2.6   0.0   1.1
Schwarzl      0.0   2.4   1.5
Pogorelov     0.1   1.0   3.1
Schoenbeck    0.0   0.0   1.0
Barras        0.0   0.1   4.2
Smith         0.6   1.4   4.8
Averyanov     0.1   3.3   0.1
Ojaniemi      0.1   0.8   0.3
Smirnov       0.2   1.7   2.7
Qi            0.2   0.1   2.0
Drews         0.1  12.8   2.3
Parkhomenko   0.7   5.7   2.0
Terek         0.3   0.5   8.1
Gomez         0.1   2.1   2.7
Turi          2.0   0.3   0.4
Lorenzo       4.2   3.9   4.0
Karlivans     2.9   0.2   0.2
Korkizoglou   0.9   7.0  10.1
Uldal         4.8   0.1   0.3
Casarsa       6.4  20.3   0.1
```

```
> round(inert1$row.cum[,1:3],1)

             Ax1 Ax1:2 Ax1:3
SEBRLE      12.5  23.0  36.0
CLAY        11.5  15.7  52.3
KARPOV      15.7  18.8  52.5
BERNARD      4.2  15.2  26.4
YURKOV       2.8  50.2  69.4
WARNERS      2.2  49.4  62.5
ZSIVOCZKY    1.3  21.4  44.6
McMULLEN     6.0   6.9  10.2
MARTINEAU   27.1  28.7  32.0
HERNU       32.8  36.1  44.1
BARRAS      50.4  53.2  53.3
NOOL        38.7  69.1  80.3
BOURGUIGNON 86.4  86.9  95.3
Sebrle      71.9  79.3  79.5
Clay        72.0  75.0  75.2
Karpov      84.2  84.2  84.2
Macey       39.9  48.1  80.0
Warners     53.5  87.3  96.9
Zsivoczky   12.7  31.6  66.1
Hernu       21.1  35.2  56.6
Nool         1.2  25.2  45.5
Bernard     44.9  45.0  53.4
Schwarzl     0.1  42.9  64.6
Pogorelov    3.3  16.2  47.4
Schoenbeck   0.8   0.8  18.5
Barras       0.0   1.5  51.7
Smith        6.9  14.6  36.3
Averyanov    1.3  42.2  43.5
Ojaniemi     3.3  13.9  16.9
Smirnov      5.7  34.7  72.2
Qi           7.6  11.0  48.7
Drews        0.7  79.5  91.0
Parkhomenko  8.0  41.6  51.1
Terek        3.5   7.1  50.5
Gomez        1.3  22.7  45.3
Turi        28.5  30.7  33.0
Lorenzo     45.7  68.2  86.3
Karlivans   54.8  56.3  57.7
Korkizoglou  7.5  38.9  75.5
Uldal       74.3  74.9  77.0
Casarsa     35.3  94.1  94.3
```

Question 3 *Combien d'axes doit-on garder pour l'analyse ? Quelle part d'inertie totale sera alors représentée ? Si la 1re valeur propre 3.32 était manquante, comment pourrait-on la retrouver à partir des données disponibles ?*

```
> plot(pca1$eig, type="b")
```



Les parts d'inerties mentionnées ci-dessus sont les valeurs propres des axes (voir figure). La règle de Kaiser nous propose de garder celles qui sont supérieures à 1. On peut en théorie garder les 4 premiers axes, mais le quatrième est tangent et on propose de le rejeter (de plus l'énoncé ne nous fournit pas de données pour le quatrième axe !). En conservant 3 axes, on représente alors 64% de l'inertie totale, puisque la somme des valeurs propres est égale à 10 (nombre de variables).

On peut utiliser la formule de la contribution aux axes pour retrouver λ_1. Par exemple, la contribution de `Karpov` à l'axe 1 s'écrit

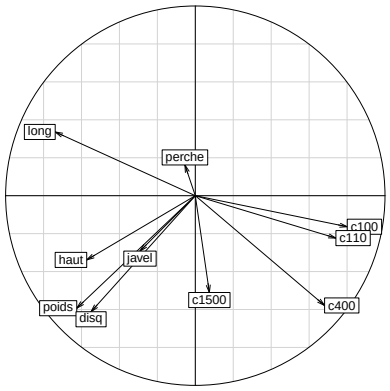
$$\frac{15,3}{100} = p_{\texttt{Karpov}} \frac{c^2_{\texttt{Karpov},1}}{\lambda_1} = \frac{1}{41}\frac{(-4,56)^2}{\lambda_1},$$

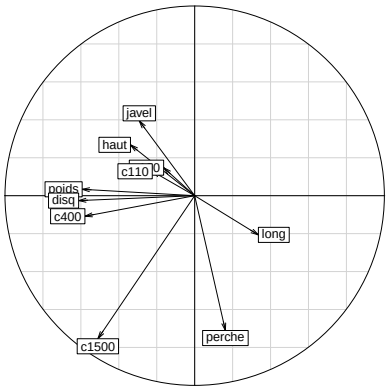
et donc

$$\lambda_1 = \frac{(-4,56)^2}{41}\frac{100}{15,3} = 3,31,$$

ce qui n'est pas trop mal vu les arrondis dans les données.

Question 4 *Quelles sont les variables qui déterminent les 3 premières composantes principales (précisez les critères utilisés) ?*





On représente ci-dessus les cercles des corrélations pour les axes $(1,2)$ et $(2,3)$ (l'axe 2 est donc représenté deux fois ici).

On se fixe un seuil au dessus duquel la corrélation est supposée pertinente. On propose ici un seuil égal à $0,60$, qui semble adapté à tous les axes. On obtient les caractérisations suivantes :

Axe 1

$\ominus$	$\oplus$
`long` $(-0,74)$	`c100` $(0,80)$
`poids` $(-0,63)$	`c110` $(0,74)$
[`haut` $(-0,57)$]	`c400` $(0,68)$

Axe 2

$\ominus$	$\oplus$
`disq` $(-0,61)$	
[`poids` $(-0,59)$]	
[`c400` $(-0,58)$]	

Axe 3

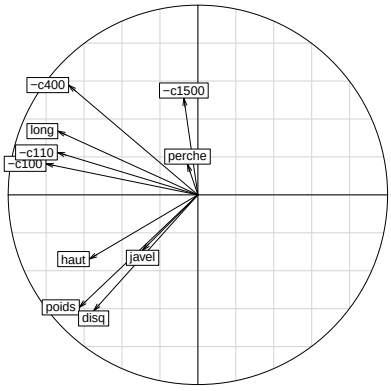
$\ominus$	$\oplus$
`c1500` $(-0,75)$	
`perche` $(-0,71)$	

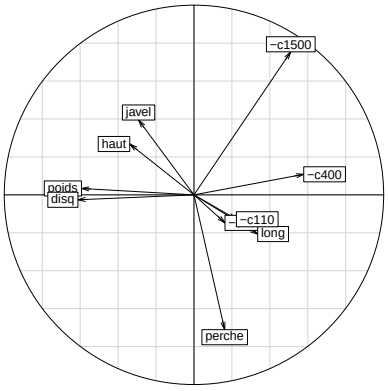
Question 5 *Expliquez comment les données peuvent être modifiées pour faire apparaître un effet de taille. Comment peut-on alors interpréter les axes principaux de la question 3 ? Est-ce que cela change ce choix d'axes ?*

```
> pca2=pca1
> courses=c(1,5,6,10)
> pca2$co[courses,]=-pca1$co[courses,]
> row.names(pca2$co)[courses]=paste("-", row.names(pca1$co)[courses], sep="")
```

En reprenant les conclusions de la question 2, on voit qu'il est souhaitable d'inverser le signe des variables de course. Si l'on fait cette opération (on renomme les variables `-c...`), on change le signe des corrélations de ces variables, qui sont maintenant corrélées positivement avec le premier axe principal.

Les cercles des corrélations deviennent :





Avec ces nouvelles variables de la question précédente, on peut récrire la description des axes comme suit

- axe 1 : facteur de taille, corrélé négativement avec la performance générale des décathloniens ;
- axe 2 : facteur de forme, opposant la force des bras (lancer d'objets lourds `poids`, `disqu`) à l'endurance (`-c400`) et plus généralement les activités où les jambes sont importantes.
- axe 3 : oppose le 1500m au saut au saut à la perche ; cet axe est difficile à expliquer, mais reprend les deux variables que nous avons remarqué en question 2 comme étant peu corrélées au score final.

On pourrait en fait se contenter de conserver 2 axes.

Question 6 *Quels sont les individus qui déterminent les trois premiers axes principaux ? (précisez les critères utilisés)*

On compare les contributions des athlètes aux axes avec leur poids, c'est à dire $100/41 = 2,44$. On ne conserve que les athlètes dont la contribution est supérieure à 3 fois le poids, c'est-à-dire ici $2,44 \times 3 = 7,32$. On garde alors

Axe 1

$\ominus$	$\oplus$
`Karpov` $(15,3)$	`BOURGUIGNON` $(12,1)$
`Serble` $(12,7)$	
`Clay` $(11,9)$	

Axe 2

$\ominus$	$\oplus$
`Casarsa` $(20,3)$	`Drews` $(12,8)$

Axe 3

$\ominus$	$\oplus$
`Korkizoglou` (10,1)	
`Terek` $(8,1)$	
`CLAY` $(7,9)$	

On aurait eu plus d'individus en prenant une limite moins contraignante, mais cela n'aurait pas aidé à l'interprétation.

Question 7 *Quels sont les 4 individus les moins bien représentés par le sous espace qu'on a décidé de conserver en question 3 ? On expliquera la signification de la qualité de la représentation des individus par un sous-espace.*

La qualité de la représentation d'un individu sur un axe principal est donnée par le cosinus carré de l'angle entre l'individu et sa projection sur cet axe (plus exactement des vecteurs partant du centre de gravité et allant vers ces points). Si on note c_{ki} la coordonnée de l'individu i sur l'axe k, on sait que la qualité de la représentation de l'individu i par l'axe k peut être calculée comme

$$\frac{c_{ki}^2}{c_{1i}^2+c_{2i}^2+\cdots+c_{10i}^2}.$$

Si on s'intéresse au sous-espace vectoriel formé par les 3 premiers axes, on additionne les qualités pour obtenir

$$\frac{c_{1i}^2+c_{2i}^2+c_{3i}^2}{c_{1i}^2+c_{2i}^2+\cdots+c_{10i}^2}.$$

Ce sont en fait ces nombres que l'on trouve dans la troisième colonne du tableau fourni.

On cherche les 4 individus les moins bien représentés par sous espace principal F_3. On trouve `McMULLEN` $(10,2)$, `Ojaniemi` $(16,9)$, `Schoenbeck` $(18,5)$ et `BERNARD` $(26,4)$.

On notera que les trois premiers athlètes sont proches de zéro sur les 3 premières composantes, ce qui rend l'interprétation plus difficile.

Question 8 *Commentez la manière dont les variables supplémentaires `RANG` et `POINTS` sont corrélées avec les axes principaux. Est-ce que cela nous apprend quelque chose ?*

Ces deux variables sont uniquement corrélées de manière significative avec l'axe 1, dont on a dit qu'il reflète le niveau des athlètes. Cela renforce l'interprétation qu'on a fait de cet axe. D'autre part, on retrouve que `POINTS` et `RANG` sont opposés (puisqu'un fort score correspond à un petit classement).

Le fait que `POINTS` et `RANG` ne soient pas corrélés avec les axes 2 et 3 est plutôt une bonne chose : s'ils ne l'étaient pas, alors on pourrait dire que le barème du décathlon privilégie certains types d'athlètes (par exemple les bons coureurs aux dépends des bons lanceurs ou sauteurs).

2 ACP sur un tableau à 2 colonnes

On se place dans le cadre de l'ACP sur 2 variables centrées réduites avec n individus. C'est bien sûr un cas où l'ACP a peu d'intérêt, mais les calculs peuvent être faits explicitement. On suppose un poids uniforme $\frac{1}{n}$ pour les individus et on note $r_{i\ell}$ la corrélation entre les variables $\mathbf{z}^i$ et $\mathbf{z}^\ell$. La matrice de corrélation s'écrit donc ici $\mathbf{R}=\begin{bmatrix}1 & r_{12}\\ r_{21} & 1\end{bmatrix}$.

Question 9 *Montrer que* $\begin{bmatrix}1\\1\end{bmatrix}$ *et* $\begin{bmatrix}1\\-1\end{bmatrix}$ *sont des vecteurs propres de* $\mathbf{R}$ *et donner les valeurs propres associées.*

La première chose à faire est de noter que $r_{12}=r_{21}$, puisque la corrélation est une notion symétrique. Du coup,

$$\mathbf{R}\begin{bmatrix}1\\1\end{bmatrix}=\begin{bmatrix}1 & r_{12}\\ r_{21} & 1\end{bmatrix}\begin{bmatrix}1\\1\end{bmatrix}=\begin{bmatrix}1+r_{12}\\ r_{21}+1\end{bmatrix}=(1+r_{12})\begin{bmatrix}1\\1\end{bmatrix}.$$

Le vecteur $\begin{bmatrix}1\\1\end{bmatrix}$est donc vecteur propre de $\mathbf{R}$ associé à la valeur propres $1+r_{12}$. De même

$$\mathbf{R}\begin{bmatrix}1\\-1\end{bmatrix}=\begin{bmatrix}1 & r_{12}\\ r_{21} & 1\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix}=\begin{bmatrix}1-r_{12}\\ r_{21}-1\end{bmatrix}=(1-r_{12})\begin{bmatrix}1\\-1\end{bmatrix},$$

ce qui termine la preuve.

Question 10 *Donner en fonction du signe de* r_{12} *l'expression des facteurs principaux* $\mathbf{u}_1$ *et* $\mathbf{u}_2$ *et des valeurs propres* λ_1 *et* λ_2*. Calculez la part d'inertie totale portée par le premier axe principal.*

On sait que l'on veut avoir $\lambda_1\geq\lambda_2$. Il faut donc définir les axes propres en fonction du signe de r_{12}. de plus, on normalise les $\mathbf{u}$ en les divisant par $\sqrt{2}$.

— $r_{12}\geq 0$: alors $\lambda_1=1+r_{12}$, $\mathbf{u}_1=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$, $\lambda_2=1-r_{12}$, $\mathbf{u}_2=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$

— $r_{12}\leq 0$: alors $\lambda_1=1-r_{12}$, $\mathbf{u}_1=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$, $\lambda_2=1+r_{12}$, $\mathbf{u}_2=\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$

Dans tous les cas, on a $I_g=2$ et la part d'inertie expliquée par le premier axe est $(1+|r_{12}|)/2$.

Question 11 *On rappelle que la corrélation entre la variable* $\mathbf{z}^j$ *et la composante principale* $\mathbf{c}_k$ *est égale à* $\sqrt{\lambda_k}u_{kj}$*. En déduire les conditions sous lesquelles l'ACP présente un effet de taille.*

Pour avoir un effet de taille, il faut que toutes les corrélations avec l'axe 1 soient de même signe. Il faut donc pour cela que $\mathbf{u}_1$ ait toutes ses coordonnées de même signe. On en déduit que $\mathbf{u}_1=\begin{bmatrix}1\\1\end{bmatrix}$ et donc $r_{12}>0$.