Fiche TD avec le logiciel R : tdr61

Analyses en Composantes Principales

A.B. Dufour & D. Chessel

La fiche passe en revue quelques usages de l'analyse en composantes principales sur différents types de tableaux. On rencontre le non centrage, le décentrage, le double centrage autour des tableaux de pourcentages, de notes, de rangs ou de notes d'abondance. Dans cette famille, le cas le plus utilisé est celui de l'ACP normée ou ACP sur matrice de corrélation. Cette pratique est incontournable quand le tableau contient des variables de nature diverse. La variance dépendant des unités, elle n'a pratiquement que la fonction de permettre la normalisation, c'est-à-dire sa propre disparition. Les tableaux homogènes, au contraire comporte dans chaque cellule un nombre comparable au contenu des autres cellules, qu'il s'agisse d'une notation unique d'abondance, une présence-absence, un rang, un pourcentage, etc. L'usage de l'ACP normée peut alors être sans inconvénient ou au contraire obscurcir définitivement l'information. A l'aide d'exemples, la fiche regroupe des cas typiques qui permettra de faire des choix pertinents.

Table des matières

1 Décathlon : le signe des corrélations **2**

2 Examen : l'origine est un paramètre libre **5**

3 Cohérence d'un jury : compromis **8**

4 Reconstitution de données : auto-modélisation **12**

5 Pourcentages : représenter des moyennes **14**

6 Morphométrie et non-centrage **19**

7 ACP et classification **21**

8 Truites et valeurs propres **26**

9 Voir le tableau **35**

10 Notes d'abondance **37**

11 Ne pas se tromper de centrage **42**

12 Information supplémentaire **44**

Références **46**

1 Décathlon : le signe des corrélations

Les données (exemple n°357 dans [6] d'après Lunn, A. D. & McNeil, D.R. (1991) *Computer-Interactive Data Analysis*, Wiley, New York) sont dans la librairie. Examiner l'objet `olympic` :

```
library(ade4)
data(olympic)
names(olympic)
```

```
[1] "tab"   "score"
```

```
is.list(olympic)
```

```
[1] TRUE
```

```
olympic$score
```

```
 [1] 8488 8399 8328 8306 8286 8272 8216 8189 8180 8167 8143 8114 8093 8083 8036 8021
[17] 7869 7860 7859 7781 7753 7745 7743 7623 7579 7517 7505 7422 7310 7237 7231 7016
[33] 6907
```

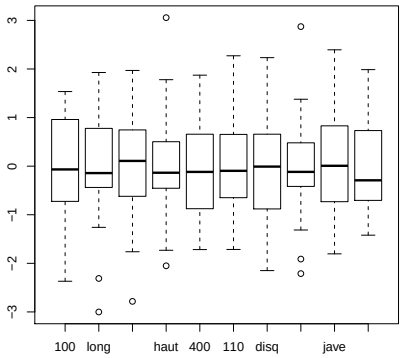
```
head(olympic$tab)
```

```
   100 long  poid haut   400   110  disq perc  jave   1500
1 11.25 7.43 15.48 2.27 48.90 15.13 49.28  4.7 61.32 268.95
2 10.87 7.45 14.97 1.97 47.71 14.46 44.36  5.1 61.76 273.02
3 11.18 7.44 14.20 1.97 48.29 14.81 43.66  5.2 64.16 263.20
4 10.62 7.38 15.02 2.03 49.06 14.72 44.80  4.9 64.04 285.11
5 11.02 7.43 12.92 1.97 47.44 14.40 41.20  5.2 57.46 256.64
6 10.83 7.72 13.58 2.12 48.34 14.18 43.06  4.9 52.18 274.07
```

```
dim(olympic$tab)
```

```
[1] 33 10
```

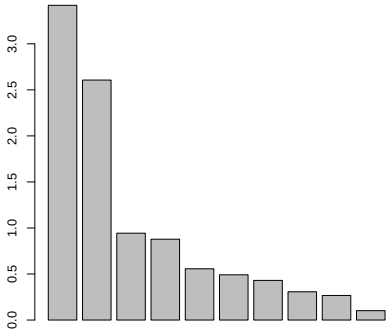
```
boxplot(as.data.frame(scale(olympic$tab)))
```



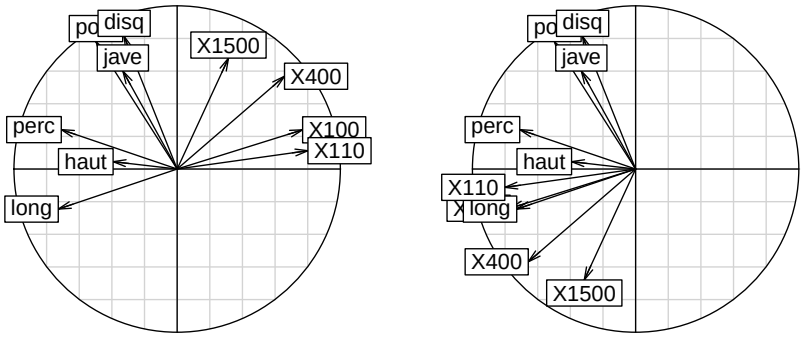

```
pca1 <- dudi.pca(olympic$tab, scannf = FALSE)
```

On sélectionne le nombre d'axes à partir du graphe des valeurs propres

```
barplot(pca1$eig)
```

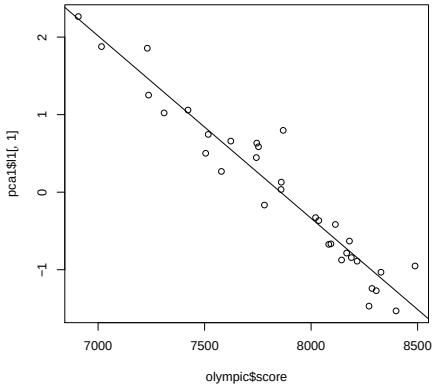



```
par(mfrow = c(1, 2))
s.corcircle(pca1$co)
olympic2 <- olympic$tab
olympic2[, c(1, 5, 6, 10)] = -olympic2[, c(1, 5, 6, 10)]
pca2 <- dudi.pca(olympic2, scan = F)
s.corcircle(pca2$co)
```



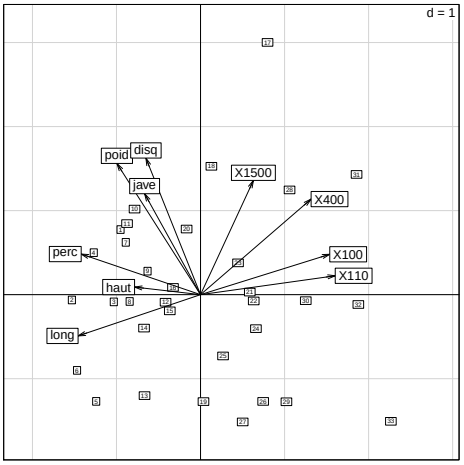
L'image de gauche est particulièrement trompeuse ! Elle est statistiquement juste et expérimentalement fausse. La performance des athlètes augmente avec les distances des lancers, la hauteur et la longueur des sauts, elle décroît avec le temps des courses. L'image de droite est mathématiquement équivalente et expérimentalement correcte.

```
plot(olympic$score, pca1$li[, 1])
abline(lm(pca1$li[, 1] ~ olympic$score))
```



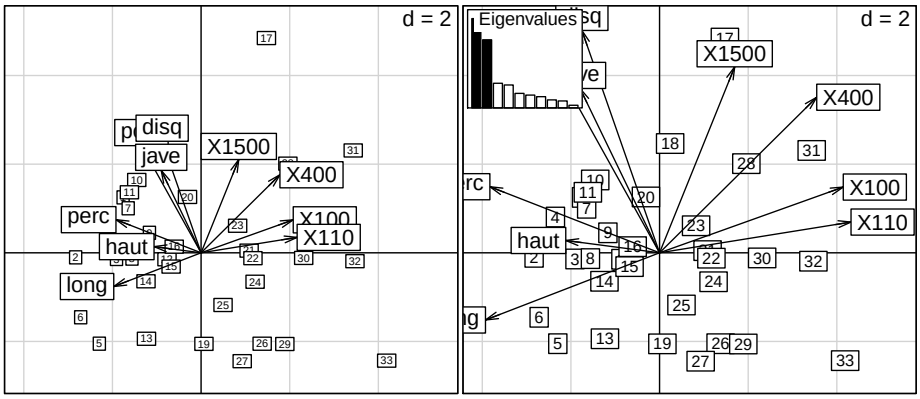
Commenter.

```
s.label(pca1$l1, clab = 0.5)
s.arrow(2 * pca1$co, add.p = T)
```



Ces deux figures sont des doubles représentations euclidiennes des nuages et des bases canoniques. La fonction générique `scatter` pour ce type d'analyse retient cette propriété.

```
par(mfrow = c(1, 2))
s.label(pca1$li, clab = 0.5)
s.arrow(5 * pca1$c1, add.p = T)
scatter(pca1)
scatter(pca2)
```



2 Examen : l'origine est un paramètre libre

`deug` est une liste à trois composantes :

`$tab` un data frame avec 104 lignes-étudiants et 9 colonnes-disciplines académiques

`$result` un facteur donnant une synthèse du résultat à l'examen (A+, A, B, B-, C-, D) pour les 104 étudiants

`$cent` un vecteur contenant la moyenne théorique pour chacune des disciplines étudiées, ceci en fonction de leur coefficient.

```
data(deug)
names(deug)
```

```
[1] "tab"    "result" "cent"
```

```
deug$result
```

```
  [1] C- B  B  A  B  C- B  B  D  A  B  B  A  C- B  B  A  D  B  A+ A  B  B  B  B  B
 [27] C- B  B  B  B  B  B  B  B  C- D  D  D  B  B  B  A  B  C- B  B  B  B  B  B  D
 [53] A+ A  A  C- B  A  C- A  B  B  B  C- C- D  B- C- C- A  B  D  A  B  B  C- C- C-
 [79] A  C- C- C- B  B  B  A  B  A  A  D  D  B- C- C- D  A  B  B  B  B  B  B  B  B
Levels: D A A+ C- B- B
```

```
names(deug$tab)
```

```
[1] "Algebra"    "Analysis"   "Proba"      "Informatic" "Economy"    "Option1"
[7] "Option2"    "English"    "Sport"
```

Exécuter et interpréter une analyse en composantes principales normée et une analyse en composantes principales décentrée.

```
args(dudi.pca)
```

```
function (df, row.w = rep(1, nrow(df))/nrow(df), col.w = rep(1,
    ncol(df)), center = TRUE, scale = TRUE, scannf = TRUE, nf = 2)
NULL
```

```
pcanor <- dudi.pca(deug$tab, scann = F)
```

Définir et interpréter le graphe canonique.

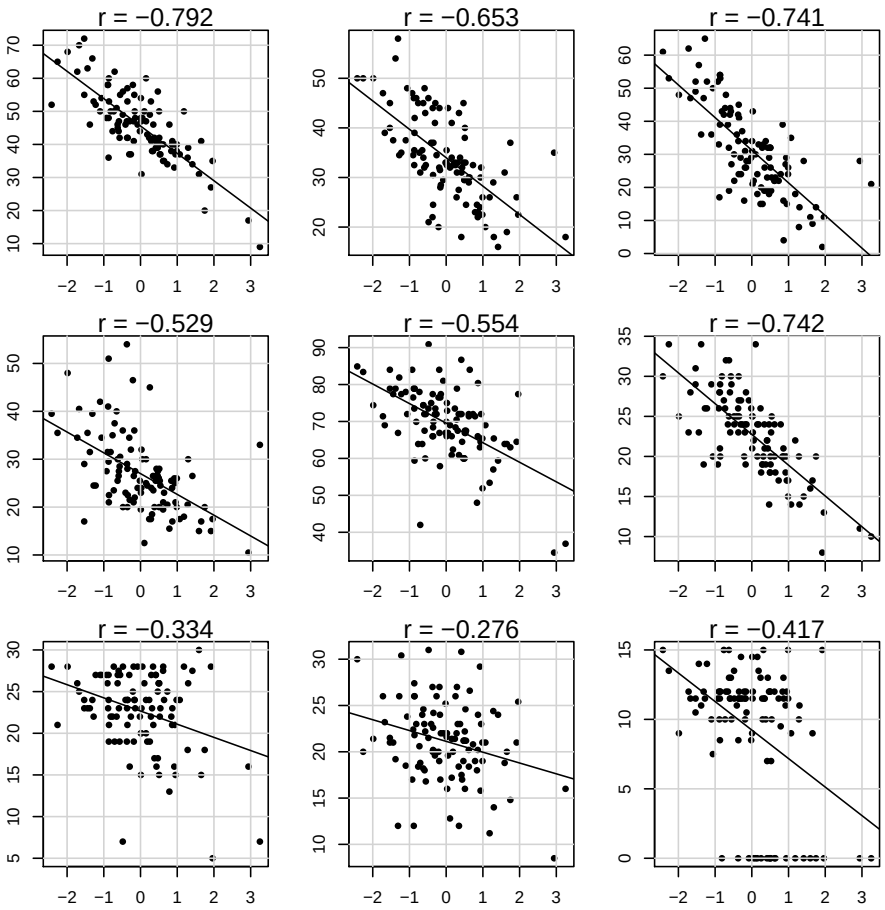
```
plotreg <- function(x, y) {
    par(mar = c(2, 2, 2, 2))
    plot(y, x, pch = 20)
    grid(lty = 1)
    abline(lm(x ~ y))
    mtext(paste("r = ", round(cor(x, y), digits = 3), sep = ""))
}
```

```
par(mfrow = c(3, 3))
apply(deug$tab, 2, plotreg, y = pcanor$l1[, 1])
```

```
NULL
```

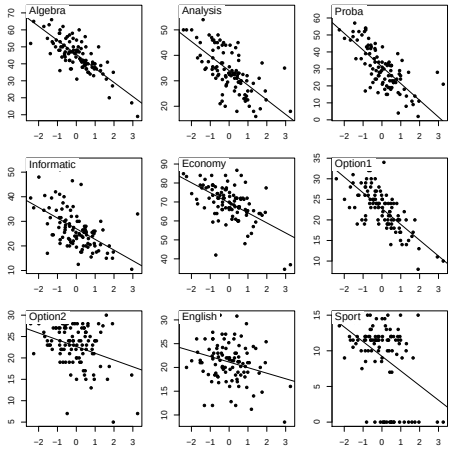
```
pcanor$co
```

```
                Comp1       Comp2
Algebra    -0.7924753  0.09239958
Analysis   -0.6531896  0.46263515
Proba      -0.7410261  0.24213427
Informatic -0.5287294  0.28126036
Economy    -0.5538660 -0.62678201
Option1    -0.7416171  0.01601705
Option2    -0.3336153 -0.37599247
English    -0.2755026 -0.66982537
Sport      -0.4171874 -0.13974793
```

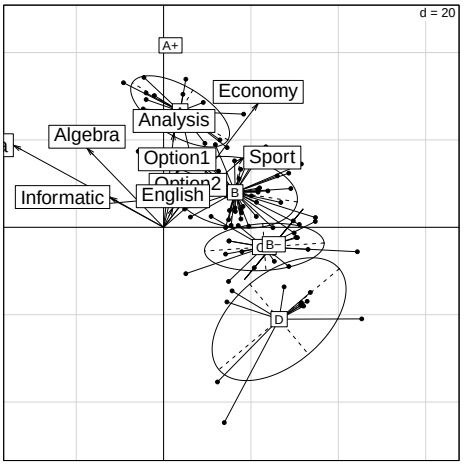


La fonction générique `score` utilise ce point de vue :

```
score(pcanor)
```

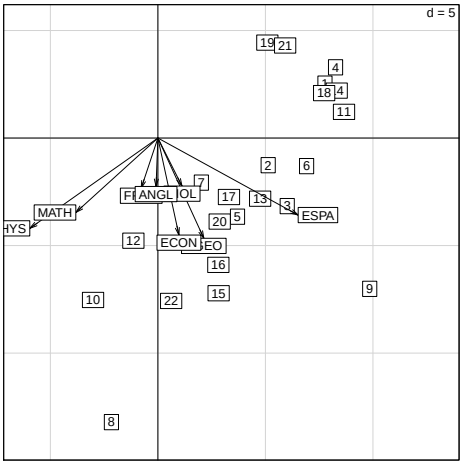



```
pca1 <- dudi.pca(deug$tab, scal = FALSE, center = deug$cent, scan = FALSE)
s.class(pca1$li, deug$result)
s.arrow(50 * pca1$c1, add.plot = T, clab = 1.5)
```



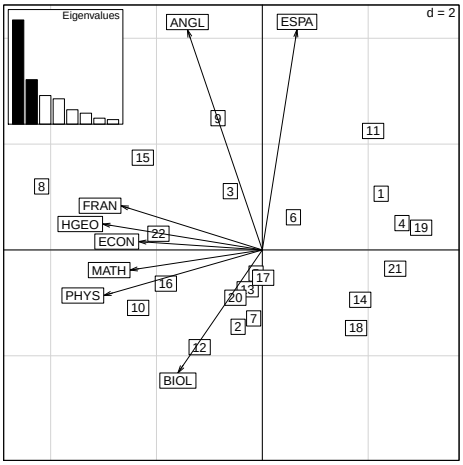
L'origine n'est plus le centre de gravité du nuage. On y gagne la distinction entre les matières qui font la forme du nuage et celles qui donnent sa position. On retrouve cet aspect dans les données 'seconde'.

```
data(seconde)
pca2 <- dudi.pca(seconde, center = rep(10, 8), scale = F, scan = FALSE)
s.label(pca2$li)
s.arrow(10 * pca2$c1, add.plot = T)
```



Cette figure ne modifie pas la notion de compromis (toutes les matières valorisent peu ou prou les bons élèves) mais souligne que Maths et Physique excluent d'abord (les élèves en haut à droite) et que le professeur d'espagnol est d'une bienveillance évidente. Ce qui diffère sensiblement de l'ACP normée qui mettait en évidence l'originalité et la solidarité des professeurs de langue :

```
scatter(dudi.pca(seconde, scan = F), clab.r = 1)
```



Quand un vecteur de valeurs définit un point signifiant de l'espace devant servir de référence, il convient d'en faire l'origine.

3 Cohérence d'un jury : compromis

S'il s'agit de mesurer la cohérence du jury, d'exprimer un compromis entre jugements, un choix collectif (ou plusieurs tendances regroupant des parties du jury), de mettre en évidence la ressemblance entre juges, *les juges sont en colonnes dans une ACP normée.* Les moyennes sont toutes égales, les variances aussi, la normalisation n'est ni nécessaire, ni nuisible. On peut l'utiliser pour tracer les cercles de corrélation.

25 juges classent 8 bouteilles dans le tour final du concours d'une grande foire :

```
data(macon)
macon
```

```
  a b c d e f g h i j k l m n o p q r s t u v w x y
A 5 5 4 3 3 4 7 2 1 3 5 4 4 5 4 8 5 7 8 5 4 6 7 2 8
B 4 8 2 4 1 5 2 7 8 8 1 6 3 7 8 5 7 8 1 4 1 5 4 4 6
C 2 6 1 1 6 2 1 5 5 4 3 7 2 2 6 2 1 6 2 1 2 1 2 5 1
D 6 7 5 8 2 6 8 8 6 6 6 5 6 6 3 6 8 1 7 6 7 4 1 6 7
E 1 4 3 2 7 1 6 4 3 1 2 8 1 1 1 3 2 2 6 2 8 2 8 1 2
F 3 2 8 6 5 8 3 3 4 7 8 1 5 8 7 4 4 3 3 8 6 8 6 7 3
G 7 1 6 5 4 7 4 1 7 5 7 3 8 3 2 7 3 5 4 7 3 7 3 8 5
H 8 3 7 7 8 3 5 6 2 2 4 2 7 4 5 1 6 4 5 3 5 3 5 3 4
```

16 juges ont rangé par ordre de préférence 28 lots de fruits [8].

```
data(fruits)
names(fruits)
```

```
[1] "type" "jug"  "var"
```

```
fruits$jug
```

```
      J1 J2 J3 J4 J5 J6 J7 J8 J9 J10 J11 J12 J13 J14 J15 J16
1.nec 10  5  8  3  1 18  5 17  3   4   1   2   5   3   1   1
2.nec  3  1  9  8  6 16  8 10  2   1   8   8   9   5   6   4
3.pea  5 11  5  2  8  8 18  3  4  15  14   4   7   1   3  13
4.pea  6 12  3  4  4  7 17  2  1  16  13   7   3   8   4  14
5.pea  4  2  4 14 17 10 16  1  5  19  21  13   6   2   5  15
6.nec  2  6 16 10 13  2 11  5 13   8  10  14  10   9  15   8
7.pea 17  9  1  5  9 24 23 19 23  18   6   3   8  14   7   5
8.nec 22 20 12  6 26 23  2  7  8  14   3  11   1  11  22   2
9.nec  7  7 14 16 23 17  6 21  9   2   2  16  20  13  11  17
10.nec 14 18 24 27 15  1 14  6 11   3  11   6  19   6  10  10
11.nec  1  8 20 15 28 15 13 11 10  11   7  22  15   7   8  11
12.nec  9  3 25 12 11 13  7 12 14   7   9  17  16  10  14   9
13.nec 12 10 19 13 25 22  4 18  6   5  17  15  11  12  18   3
14.pea 11 14  7  7  3 21 15 23 17  17  12  10   4   4   9  24
15.nec 24 22 13  1  2 28  1 16  7  13  15   1  27  16   2  20
16.nec 15  4 21 19 18  3  9  8 12  10  16  19  21  20  26   7
17.pea 20 17 10 17 21  9 20  9 20  22   4  12   2  17  20  18
18.nec 13 15 15 23 27 14 12 13 16   9   5  23  14  15  25  12
19.pea 19 19  2  9 19  6 28 14 21  21  18   9  12  22  21  19
20.pea 23 27  6 11 24 25 19 20 25  25  23   5  17  19  12   6
21.pea 16 16 17 22 16  5 21  4 19  20  19  20  13  18  23  16
22.nec 18 21 26 18 10 20  3 27 15  12  22  18  25  21  27  21
23.nec  8 13 28 28 20 19 10 24 18   6  20  24  24  27  28  22
24.pea 21 24 23 20 14  4 22 15 22  23  25  21  18  28  16  25
25.pea 28 25 11 26  7 11 24 26 24  26  26  27  22  25  17  27
26.pea 27 23 18 21 22 27 27 22 28  24  28  25  23  23  13  23
27.pea 26 28 27 24  5 12 25 28 27  28  24  28  28  24  19  28
28.pea 25 26 22 25 12 26 26 25 26  27  27  26  26  26  24  26
```

51 étudiants de la filière biomathématique ont exprimé leur préférence sur 10 groupes de musique. Les données sont dans l'objet `rankrock`.

```
data(rankrock)
rankrock
```

```
           X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 X20
Metallica   6  7  4  7  7 10  9  3 10   1   7   3   7   1   7   6   8   8  10   6
Guns.n.Roses 10  9  9  8  5  8 10  2  5   9   3   6   2   7   9   4   5   6   8   9
Nirvana     5  8  8  4  2  6  6  9  3   2   8   2   6   6   3   5   4   4   9   5
AC.DC       9  4  2  6  6  9  7  8  4   6   6   7   5   2   8   7   9   9   5   8
Noir.Desir  2  6  3  1  1  3  2  1  2  10   5   4   9   5   1   2   1   2   3   3
U2          7  2  7  3  8  4  5  4  1   4   2   1   4   8   2   1   2   7   6   7
Pink.Floyd  1  3  5  5  4  1  1  6  6   5   4  10   3   4   4   3   3   1   1   1
Led.Zeppelin  3  1  1  9  9  2  4  5  8   8   9   9   1   3   5   8   6   5   4   4
Deep.Purple  4  5  6 10 10  5  3  7  9   7  10   8   8   9   6   9   7   3   2   2
Bon.Jovi    8 10 10  2  3  7  8 10  7   3   1   5  10  10  10  10  10  10   7  10
           X21 X22 X23 X24 X25 X26 X27 X28 X29 X30 X31 X32 X33 X34 X35 X36 X37 X38
Metallica    9   9  10   9   9   5  10   9   4   6   9   1   1   8   9   6   6   6
Guns.n.Roses  5   6   9   8   8   7   8   8   7   5   5   2   5   1   6   5   9   4
Nirvana      4   7   2   6   7   3   7   2   8   4   3   3   4   6   5   7   4   5
AC.DC        8  10   8  10  10   6   9   7   1   7  10   6   8   9  10  10   7   7
Noir.Desir   3   3   1   2   2   8   2  10   9  10   4   4   2   4   4   3   1   3
U2           1   2   3   1   1   1   1   5   6   2   1   9   3   2   1   2   3   1
Pink.Floyd   2   1   4   5   6   2   3   1   3   1   2   7   6   5   2   1   2   2
Led.Zeppelin  6   8   5   3   5   9   6   6   2   3   8   5   7   7   7   8  10  10
Deep.Purple  7   5   7   4   4  10   5   3   5   8   6  10   9  10   3   4   5   9
```

http://www.metallica.com/flashport/index.html

http://www.gnronline.com/

http://www.thehighwaystar.com/lang/fr/index.thtml

http://www.ac-dc.net/

http://www.noirdesir.fr.fm/

http://www.bonjovi.com/

http://www.pinkfloyd.com/home/10.html

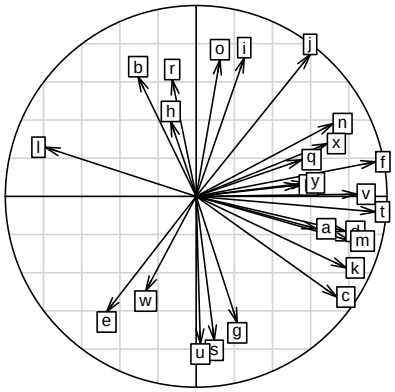
http://www.led-zeppelin.com/

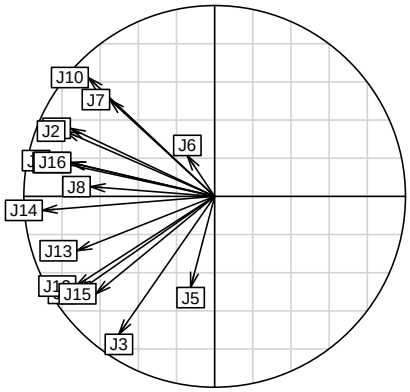
http://www.geocities.com/Hollywood/Movie/6821/thecobain.html

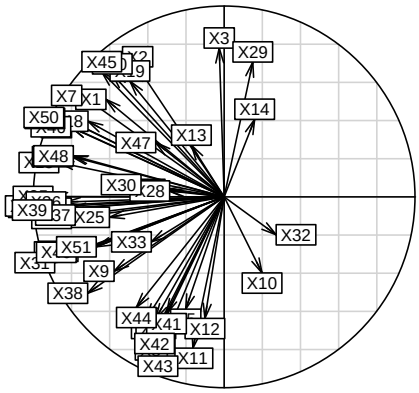
```
Bon.Jovi      10   4   6   7   3   4   4   4  10   9   7   8  10   3   8   9   8   8
              X39 X40 X41 X42 X43 X44 X45 X46 X47 X48 X49 X50 X51
Metallica     10   6   6   4   6   9   7   8   9   9   8   8   4
Guns.n.Roses   6   5   2   5   5   4   8   9   8   6   9   9   6
Nirvana        5   3   5   7   4   1   6   3   3   5   7   6   2
AC.DC          9   7  10   8   8  10   9   7   1   7   6   7   8
Noir.Desir     2   4   4   6   7   6   5   1   4   4   3   4   5
U2             3   1   7   1   1   5   4   5   2   1   1   1   1
Pink.Floyd     1   2   1   2   3   2   2   2   7   3   2   2   3
Led.Zeppelin   8   8   8   9   9   8   1   4   5   2   5   5   7
Deep.Purple    4   9   9  10  10   7   3   6   6   8   4   3   9
Bon.Jovi       7  10   3   3   2   3  10  10  10  10  10  10  10
```

Chacune des colonnes donne le rang (1 pour le préféré, ..., 10 pour le moins apprécié) qu'un étudiant attribue à chaque groupe.

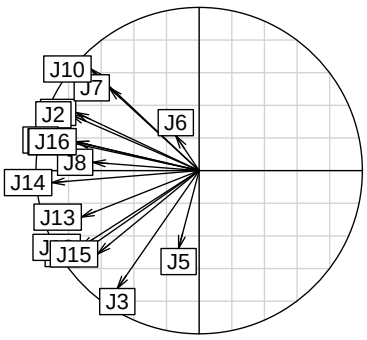
```
par(mfrow = c(2, 2))
data(macon)
s.corcircle(dudi.pca(macon, scan = F)$co)
data(fruits)
s.corcircle(dudi.pca(fruits$jug, scan = F)$co)
data(rankrock)
s.corcircle(dudi.pca(rankrock, scan = F)$co)
```

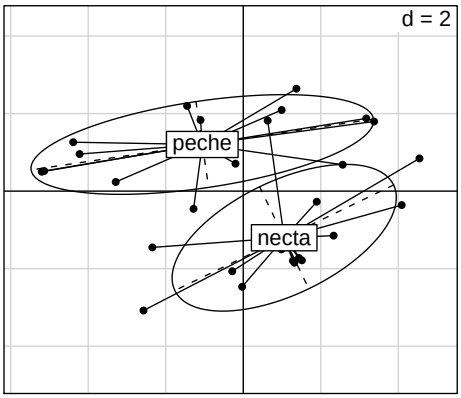







```
par(mfrow = c(1, 2))
s.corcircle(dudi.pca(fruits$jug, scan = F)$co)
s.class(dudi.pca(fruits$jug, scan = F)$li, fruits$type)
```

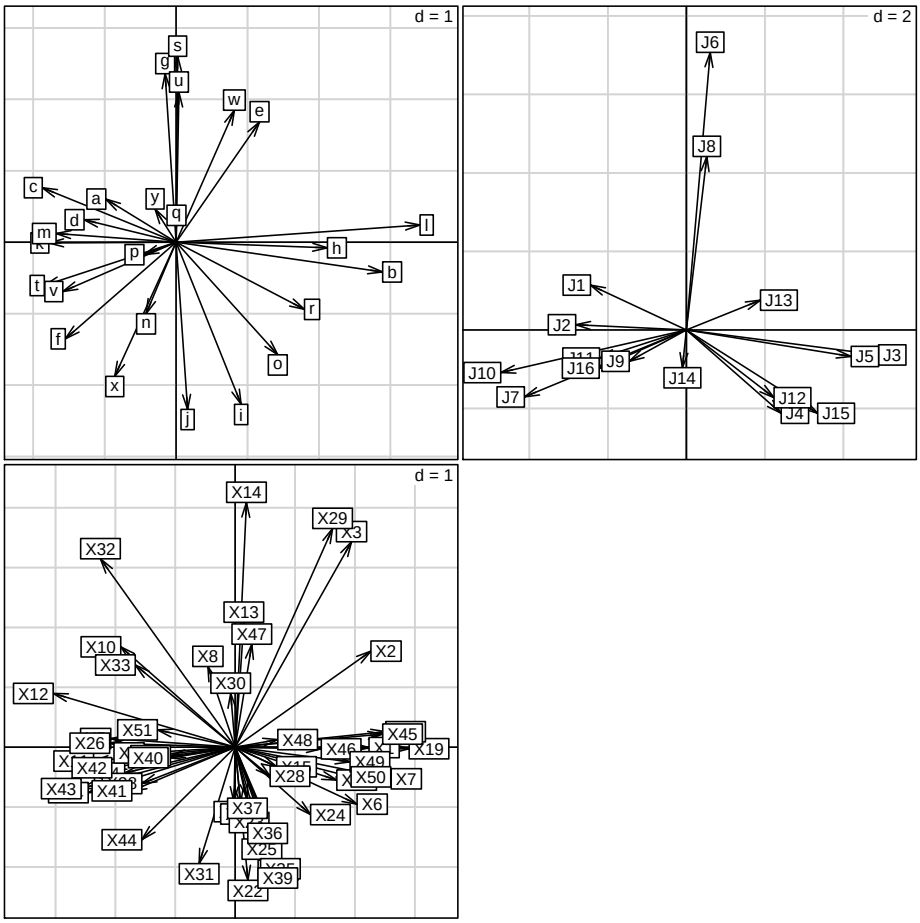




Quelle est la particularité des juges 7 et 10 ? Indiquer comment on relie ces deux figures.

S'il s'agit de faire une typologie des juges, de mettre en évidence ce qui les opposent, de montrer qu'il existe plusieurs types de jugements, *les juges sont en lignes dans une ACP centrée.* On laissera ainsi dominer dans l'analyse les produits qui ont reçu les appréciations les plus variables. Les deux approches sont antinomiques.

```
par(mfrow = c(2, 2))
s.arrow(dudi.pca(t(macon), scan = F)$li)
s.arrow(dudi.pca(t(fruits$jug), scan = F)$li)
s.arrow(dudi.pca(t(rankrock), scan = F)$li)
```



On remarquera, que derrière le compromis, s'il est suffisant pour définir le premier axe, les premières analyses définissent aussi les contradictions entre jugements, mais les secondes sont plus claires.

4 Reconstitution de données : auto-modélisation

Dans la thèse [1] de G. Carrel, les données portent sur 15 variables physico-chimiques mesurées en une station au cours de l'année 1983-1984 à 39 reprises. Ces 15 variables sont :

```
data(rhone)
names(rhone$tab)
```

```
 [1] "air.temp"   "wat.temp"   "conduc"     "pH"         "oxygen"     "secchi"
 [7] "caco3"      "totca"      "mg"         "so4"        "no2"        "hco3"
[13] "suspension" "organique"  "chloro"
```

`air.temp` Température de l'air (C) mg Magnésium (mg/l Mg++)

`wat.temp` Température de l'eau (C) so4 Sulfates (mg/l x10)

`conduc` Conductivité (mS/cm) no2 Azote nitrique (mb/l)

`pH` potentiel Hyrogène (pH) hco3 TAC (mg/l HCO3-)

`oxygen` Saturation en oxygène (%) suspension Mat. en suspension (mg/l)

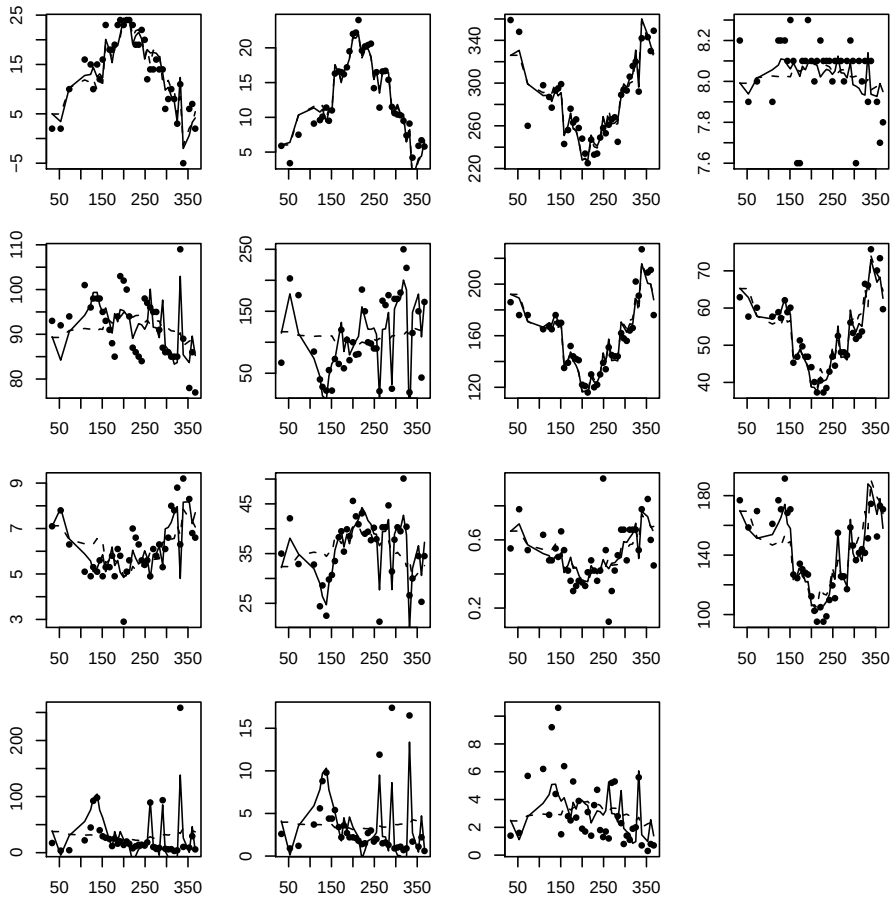
`secchi` Transparence (cm) organique Mat. organique (mg/l)

`caco3` Dureté totale (mg/l CaCO3) chloro Chlorophyle a (mg/l)

`totca` Dureté calcique (mg/l Ca++)

Automodéliser la chronique :

```
dd1 <- dudi.pca(rhone$tab, nf = 2, scann = F)
rh1 <- reconst(dd1, 1)
rh2 <- reconst(dd1, 2)
par(mfrow = c(4, 4))
par(mar = c(2.6, 2.6, 1.1, 1.1))
for (i in 1:15) {
    plot(rhone$date, rhone$tab[, i], pch = 20)
    lines(rhone$date, rh1[, i], lty = 2)
    lines(rhone$date, rh2[, i], lty = 1)
}
```



5 Pourcentages : représenter des moyennes

Un tableau **X** est un tableau de fréquences si la somme des valeurs par ligne ou la somme des valeurs par colonne vaut l'unité. On notera alors $x_{ij} = f_{j/i}$ dans le premier cas et $x_{ij} = f_{i/j}$ dans le second. Si on hésite, c'est qu'on est sur un problème d'analyse des correspondances (AFC).

```
data(granulo)
names(granulo)
```

```
[1] "tab"  "born"
```

```
head(granulo$tab)
```

```
  V1 V2  V3  V4   V5   V6   V7   V8   V9
1  0  2   1  15  260  810 1815 1990    0
2  0  0  12 340 1220 2095 2950 2050 1160
3  0  0   0  12 1505 1960 4415 1265  260
4  0  0 185 440  330  335  725 1195 2210
5  0  2   1   8  400 1480 4935 2900    0
6  0  0   0   0  115 1810 6845 4030  280
```

```
granulo$born
```

```
[1]   0.3   0.5   1.0   2.0   4.0   8.0  16.0  32.0  64.0 128.0
```

Ce tableau comporte 49 lignes - échantillons et 9 colonnes - classes de diamètres [4]. L'échantillon 32 a fourni 2 grammes de grains ayant un diamètre compris entre 0.3 et 0.5 mm (sable fin), . . . , 293 grammes de grains compris entre 64 et 128 mm (gros

galets). Le poids total de sédiments récoltés par la drague n'est que le résultat du hasard et permet de calculer un profil par lignes :

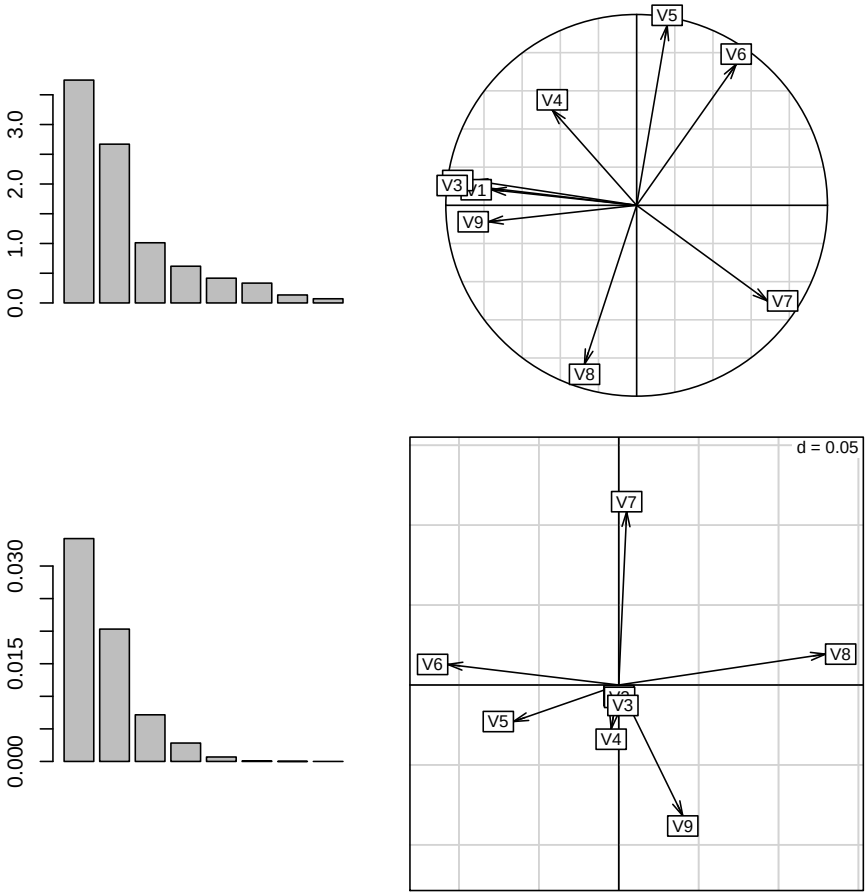
```
grapc <- t(apply(granulo$tab, 1, function(x) x/sum(x)))
round(head(grapc), dig = 2)
```

```
  V1 V2   V3   V4   V5   V6   V7   V8   V9
1  0  0 0.00 0.00 0.05 0.17 0.37 0.41 0.00
2  0  0 0.00 0.03 0.12 0.21 0.30 0.21 0.12
3  0  0 0.00 0.00 0.16 0.21 0.47 0.13 0.03
4  0  0 0.03 0.08 0.06 0.06 0.13 0.22 0.41
5  0  0 0.00 0.00 0.04 0.15 0.51 0.30 0.00
6  0  0 0.00 0.00 0.01 0.14 0.52 0.31 0.02
```

```
grapc = data.frame(grapc)
```

Chaque cellule contient une fréquence et le tableau est homogène.

```
par(mfrow = c(2, 2))
pcaa <- dudi.pca(grapc, scann = F)
barplot(pcaa$eig)
s.corcircle(pcaa$co)
pcab <- dudi.pca(grapc, scal = F, scann = F)
barplot(pcab$eig)
s.arrow(pcab$co)
```



Les variances par colonne sont très différentes :

```
apply(grapc, 2, var)
```

```
          V1           V2           V3           V4           V5           V6
5.047118e-07 6.827175e-06 9.624694e-05 2.187413e-03 6.208129e-03 1.333202e-02
          V7           V8           V9
1.382115e-02 1.952539e-02 1.143872e-02
```

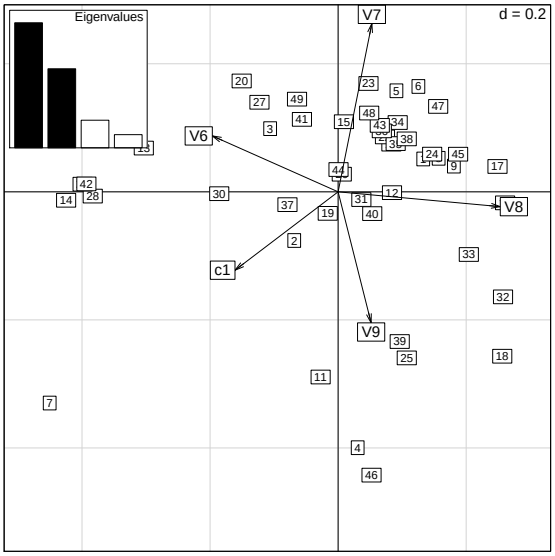
mais ce n'est pas une raison pour s'en débarrasser. La première classe joue dans l'ACP normée un rôle disproportionné à son importance expérimentale. On peut regrouper les cinq premières classes.

```
c1 <- apply(grapc[, 1:5], 1, sum)
granou <- cbind.data.frame(c1, grapc[, 6:9])
round(head(granou), dig = 2)
```

```
   c1   V6   V7   V8   V9
1 0.06 0.17 0.37 0.41 0.00
2 0.16 0.21 0.30 0.21 0.12
3 0.16 0.21 0.47 0.13 0.03
4 0.18 0.06 0.13 0.22 0.41
5 0.04 0.15 0.51 0.30 0.00
6 0.01 0.14 0.52 0.31 0.02
```

Pour une première approche, rapide et approximative :

```
par(mfrow = c(1, 1))
scatter(dudi.pca(granou, scal = F, scan = F))
```

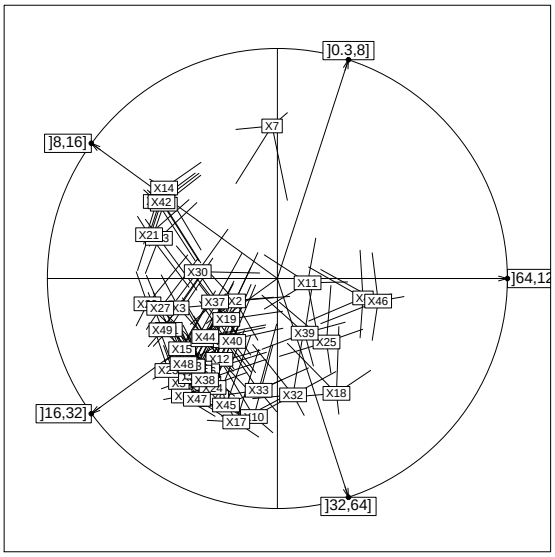


On a perdu une valeur propre (démontrer que la dernière est toujours nulle). L'analyse indique que la position des variables est celle de la représentation triangulaire étendue à 5 dimensions.

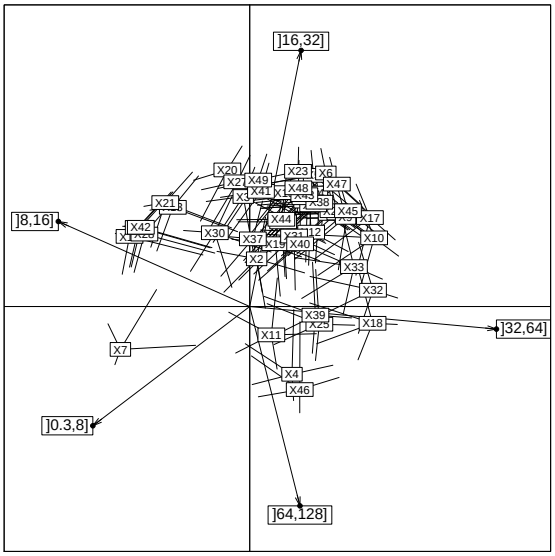
```
x = cos((1:5) * 2 * pi/5)
y = sin((1:5) * 2 * pi/5)
xy = cbind.data.frame(x, y)
row.names(xy) = c("]0.3,8]", "]8,16]", "]16,32]", "]32,64]", "]64,128]")
xy
```

```
                x            y
]0.3,8]    0.309017  9.510565e-01
]8,16]    -0.809017  5.877853e-01
]16,32]   -0.809017 -5.877853e-01
]32,64]    0.309017 -9.510565e-01
]64,128]   1.000000 -2.449294e-16
```

```
s.corcircle(xy, grid = F)
s.distri(xy, data.frame(t(granou)), add.p = T, axesell = T, csta = 0.2,
    clab = 0.75, cell = 0)
```




```
names(granou) <- row.names(xy)
xy.pca <- dudi.pca(granou, scal = F, scan = F)$c1
s.arrow(xy.pca, grid = F, lab = names(granou))
s.distri(xy.pca, data.frame(t(granou)), add.p = T, axesell = T,
    csta = 0.2, clab = 0.75, cell = 0)
```

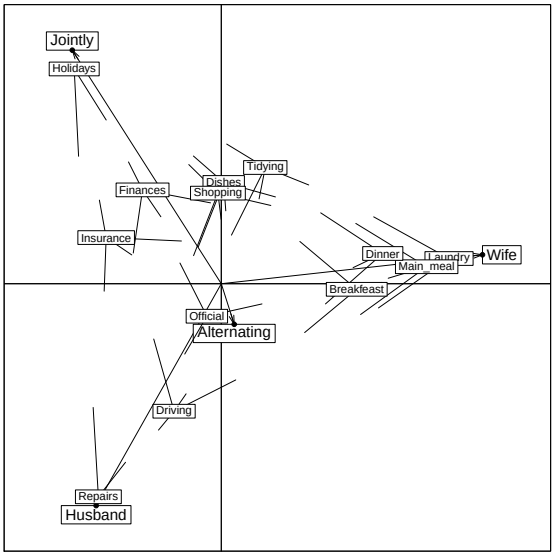


L'image *naïve* (pas tant que ça ?) contient presque la même information que l'image optimale. Les structures des tableaux de pourcentages sont très particulières et supportent mal une normalisation indésirable. Il est logique de privilégier l'expression d'un point au centre de gravité de sa distribution plutôt que de laisser faire le centrage. Mais ce centrage est nécessaire pour éviter l'expression absurde de l'évidence "les données sont positives". La représentation triangulaire s'impose pour trois variables.

```
data(housetasks)
hc <- t(apply(housetasks, 1, function(x) x/sum(x)))
round(hc, dig = 2)
```

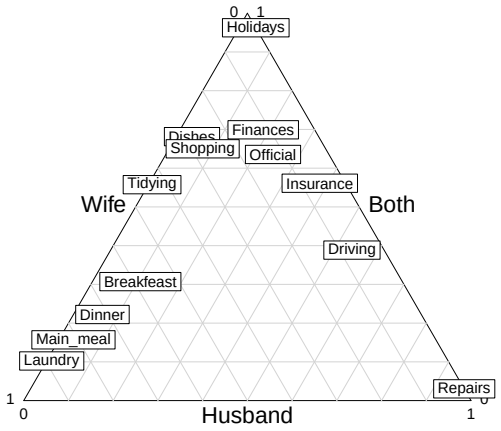
```
          Wife Alternating Husband Jointly
Laundry   0.89        0.08    0.01    0.02
Main_meal 0.81        0.13    0.03    0.03
Dinner    0.71        0.10    0.06    0.12
Breakfeast 0.59       0.26    0.11    0.05
Tidying   0.43        0.09    0.01    0.47
Dishes    0.28        0.21    0.04    0.47
Shopping  0.28        0.19    0.08    0.46
Official  0.12        0.48    0.24    0.16
Driving   0.07        0.37    0.54    0.02
Finances  0.12        0.12    0.19    0.58
Insurance 0.06        0.01    0.38    0.55
Repairs   0.00        0.02    0.97    0.01
Holidays  0.00        0.01    0.04    0.96
```

```
hc <- data.frame(hc)
hc.pca <- dudi.pca(hc, scal = F, scan = F)$c1
s.arrow(hc.pca, grid = F, lab = names(hc))
s.distri(hc.pca, data.frame(t(hc)), add.p = T, axesell = T, csta = 0.2,
    clab = 0.75, cell = 0)
```



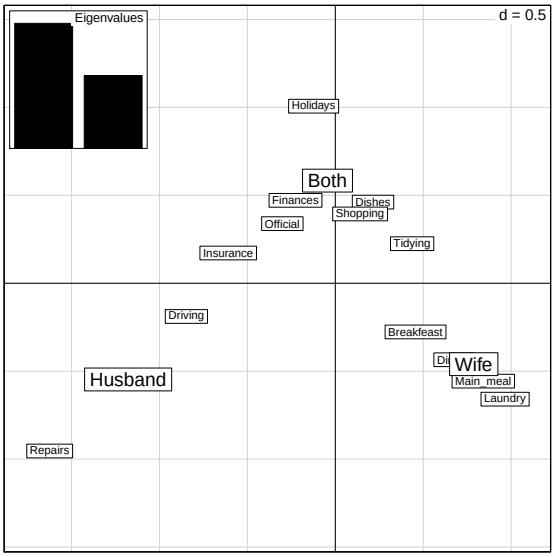
A Madame, la cuisine, à Monsieur la voiture, au couple le reste. Groupons "ensemble" et "chacun à son tour", la représentation triangulaire suffira :

```
tri <- cbind.data.frame(hc$Wife, hc$Husband, hc$Jointly + hc$Alternating)
names(tri) <- c("Wife", "Husband", "Both")
triangle.plot(tri, show = F, clab = 1, label = row.names(hc))
```



Une dissymétrie certaine. Ci-après, l'analyse des correspondances.

```
row.names(tri) <- row.names(hc)
scatter(dudi.coa(tri, scann = F))
```



En indiquant comment, pour une tâche donnée, les couples se répartissent, on indique clairement une intention.

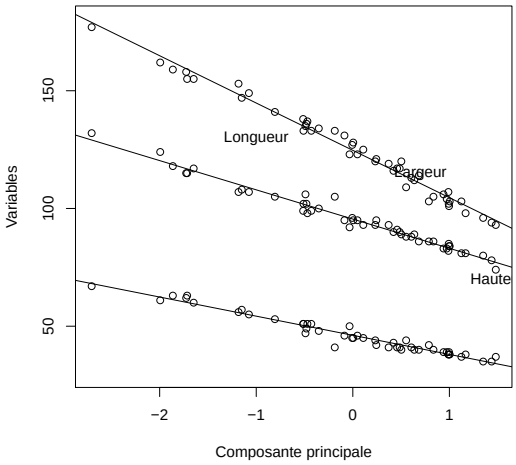
6 Morphométrie et non-centrage

C'est souvent un problème de morphométrie. Dans `tortues`, les variables sont les trois dimensions de carapaces de tortues mesurées en mm [7].

```
data(tortues)
ttaille <- tortues[, 1:3]
tsexe <- tortues[, 4]
names(ttaille)
```

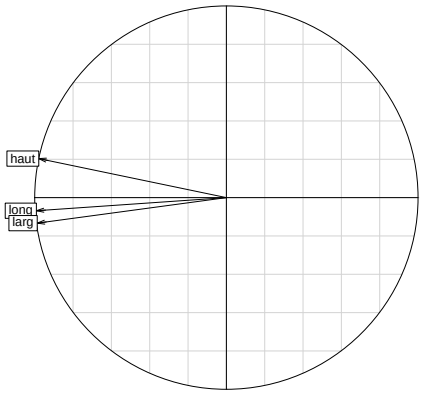
```
[1] "long" "larg" "haut"
```

```
dudi1 <- dudi.pca(ttaille, scan = F)
plot(dudi1$l1[, 1], tortues$long, ylim = c(30, 180), xlab = "Composante principale",
    ylab = "Variables")
text(-1, 130, "Longueur")
points(dudi1$l1[, 1], tortues$haut)
text(1.5, 70, "Hauteur")
points(dudi1$l1[, 1], tortues$larg)
text(0.7, 115, "Largeur")
abline(lm(tortues$larg ~ dudi1$l1[, 1]))
abline(lm(tortues$long ~ dudi1$l1[, 1]))
abline(lm(tortues$haut ~ dudi1$l1[, 1]))
```



L'ACP joue ici son rôle de recherche de *variable latente*. La composante principale prédit les trois variables. C'est une *explicative cachée*. Elle représente la taille théorique de la tortue. C'est la variable cachée qui prédit au mieux les autres, c'est aussi la variable cachée qui est le mieux prédite par toutes les autres.

```
s.corcircle(dudi1$co)
```



On dit que c'est un effet "taille".

```
coefficients(lm(tortues$larg ~ dudi1$l1[, 1]))
```

```
 (Intercept) dudi1$l1[, 1]
    95.50000     -12.41816
```

```
coefficients(lm(tortues$long ~ dudi1$l1[, 1]))
```

```
 (Intercept) dudi1$l1[, 1]
   124.68750     -20.09136
```

```
coefficients(lm(tortues$haut ~ dudi1$l1[, 1]))
```

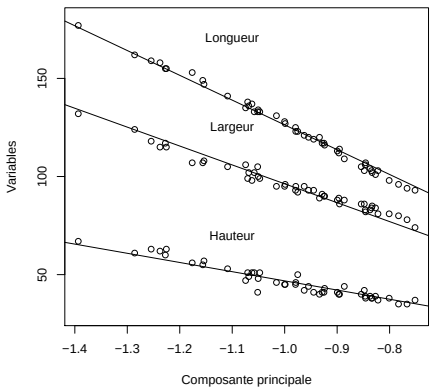
```
 (Intercept) dudi1$l1[, 1]
   46.145833     -8.120671
```

```
apply(ttaille, 2, mean)
```

```
    long      larg      haut
124.68750  95.50000  46.14583
```

Si on veut une variable latente qui fait des prédictions nulles pour une valeur nulle (régression par l'origine), on fait une ACP non centrée.

```
dudi2 <- dudi.pca(ttaille, scan = F, cent = F, scal = F)
plot(dudi2$l1[, 1], tortues$long, ylim = c(30, 180), xlab = "Composante principale",
    ylab = "Variables")
text(-1.1, 170, "Longueur")
points(dudi2$l1[, 1], tortues$haut)
text(-1.1, 70, "Hauteur")
points(dudi2$l1[, 1], tortues$larg)
text(-1.1, 125, "Largeur")
abline(lm(tortues$larg ~ -1 + dudi2$l1[, 1]))
abline(lm(tortues$long ~ -1 + dudi2$l1[, 1]))
abline(lm(tortues$haut ~ -1 + dudi2$l1[, 1]))
```



Cet *auto-modèle* (créé par les données pour modéliser les données) est plus réaliste mais moins bon (les erreurs ont une organisation, liée à la présence des deux sexes).

7 ACP et classification

La source des données est dans Prodon et Lebreton [10]. Le pourcentage de recouvrement de la végétation est mesuré pour 8 strates dans 182 sites :

1) Rocher
2) 0 m / 0.25 m
3) 0.25 m / 0.50 m

4) 0.50 m - 1 m

5) 1 m / 2 m

6) 2 m / 4 m

7) 4 m / 8 m

8) 8 m / 16 m

```
data(rpjdl)
names(rpjdl)
```

```
[1] "fau"   "mil"   "frlab" "lab"   "lalab"
```

```
mil = rpjdl$mil
names(mil)
```

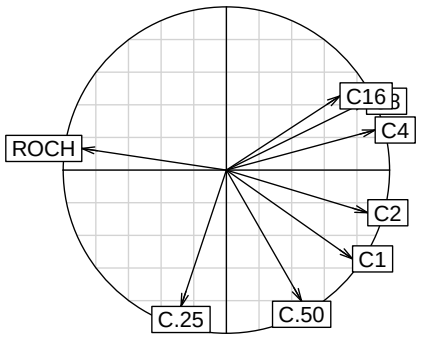
```
[1] "ROCH" "C.25" "C.50" "C1"   "C2"   "C4"   "C8"   "C16"
```

```
dim(mil)
```

```
[1] 182   8
```

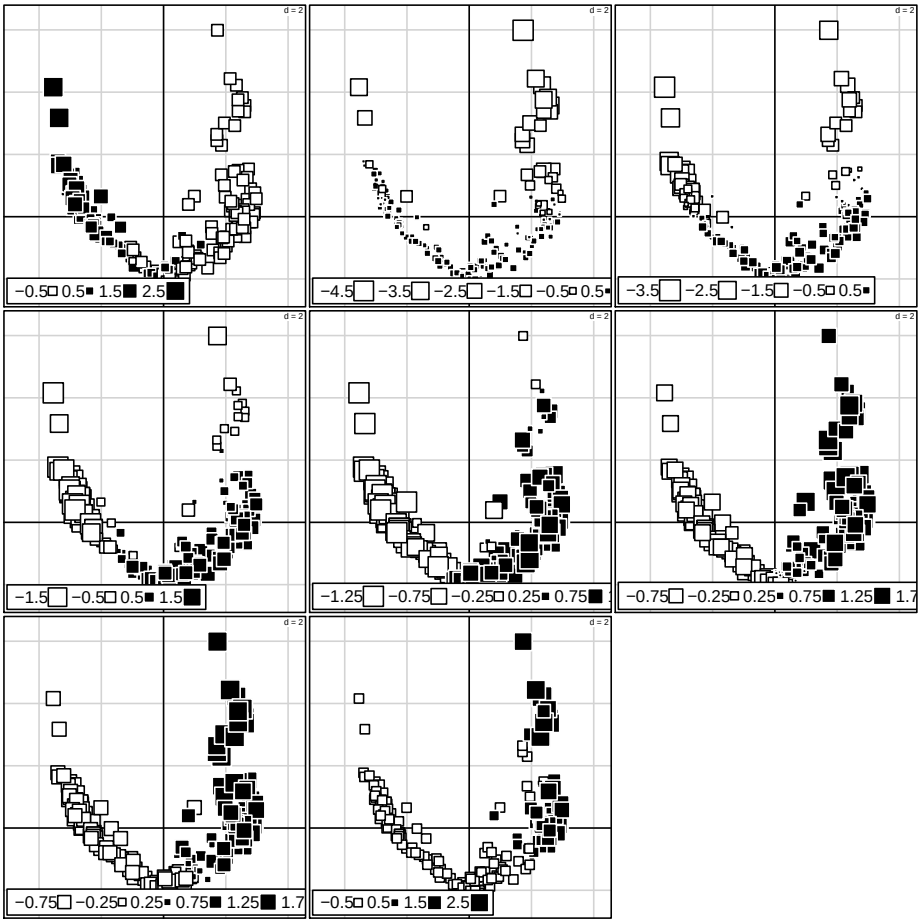
```
millog <- log(rpjdl$mil + 1)
pcamil <- dudi.pca(millog, scan = F)
s.corcircle(pcamil$co)
s.label(pcamil$li, clab = 0.75)
round(cor(millog), dig = 3)
```

```
       ROCH   C.25   C.50     C1     C2     C4     C8    C16
ROCH  1.000  0.132 -0.484 -0.733 -0.725 -0.732 -0.686 -0.533
C.25  0.132  1.000  0.504  0.157 -0.076 -0.419 -0.523 -0.441
C.50 -0.484  0.504  1.000  0.752  0.532  0.212  0.079  0.017
C1   -0.733  0.157  0.752  1.000  0.840  0.539  0.395  0.256
C2   -0.725 -0.076  0.532  0.840  1.000  0.737  0.590  0.399
C4   -0.732 -0.419  0.212  0.539  0.737  1.000  0.920  0.665
C8   -0.686 -0.523  0.079  0.395  0.590  0.920  1.000  0.779
C16  -0.533 -0.441  0.017  0.256  0.399  0.665  0.779  1.000
```

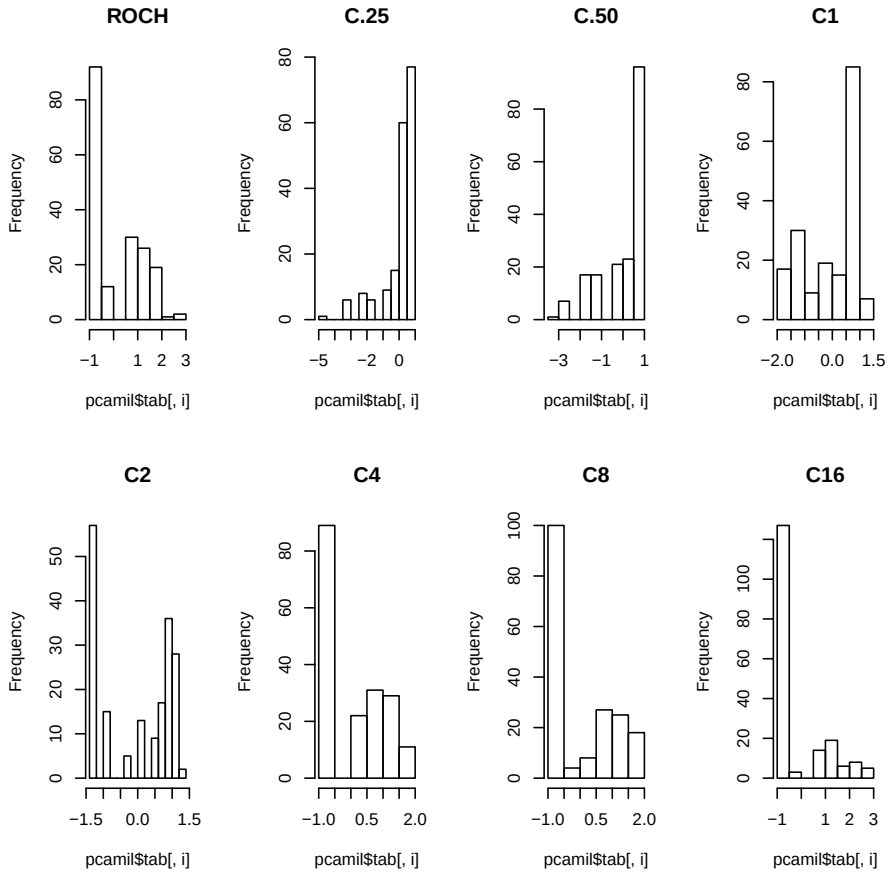


Un artifice ? Retourner aux données :

```
par(mfrow = c(3, 3))
for (i in 1:8) s.value(pcamil$li, pcamil$tab[, i], cleg = 1.5)
```

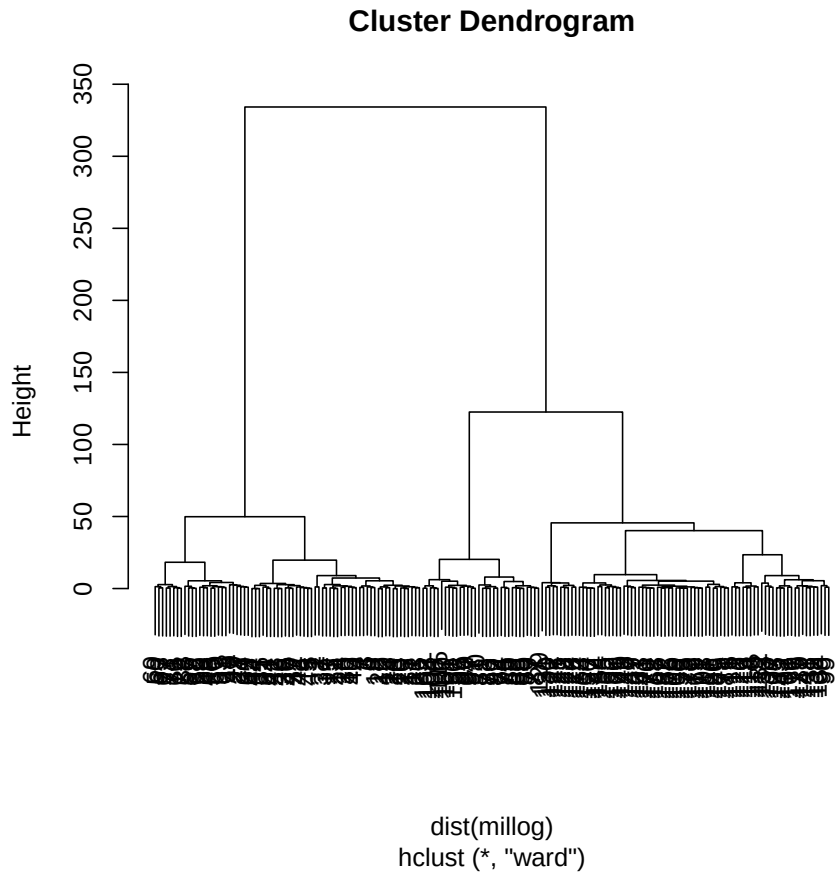



```
par(mfrow = c(2, 4))
for (i in 1:8) hist(pcamil$tab[, i], main = names(pcamil$tab)[i])
```

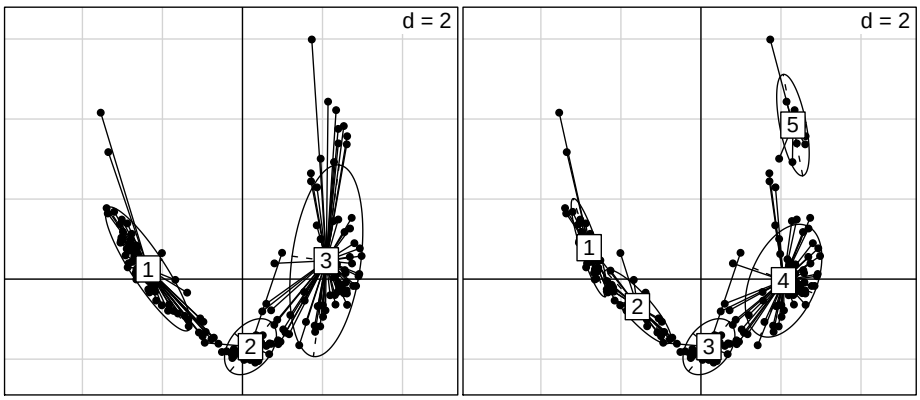


Variables à seuil ? Classement des stations ?
Deux gradients ? Un gradient et une partition ?

```
plot(hclust(dist(millog), "ward"))
```




```
par(mfrow = c(1, 2))
s.class(pcamil$li, as.factor(cutree(hclust(dist(millog), "ward"),
    3)))
s.class(pcamil$li, as.factor(cutree(hclust(dist(millog), "ward"),
    5)))
par(mfrow = c(1, 1))
```




```
parti <- as.factor(cutree(hclust(dist(millog), "ward"), 5))
summary(parti)
```

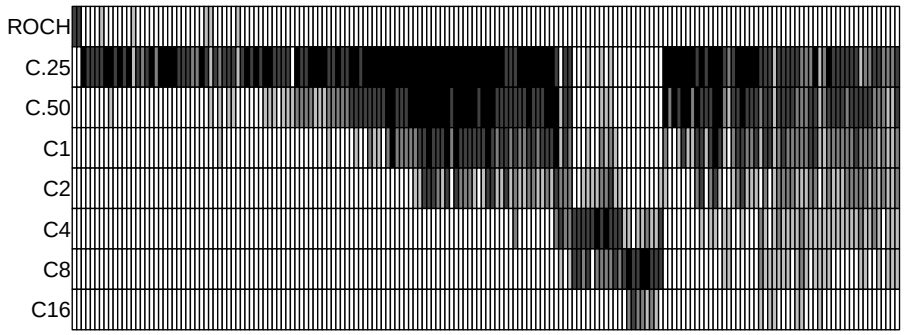
```
 1  2  3  4  5
46 26 32 68 10
```

```
round(data.frame(lapply(split(mil, parti), function(x) apply(x,
    2, mean))), dig = 3)
```

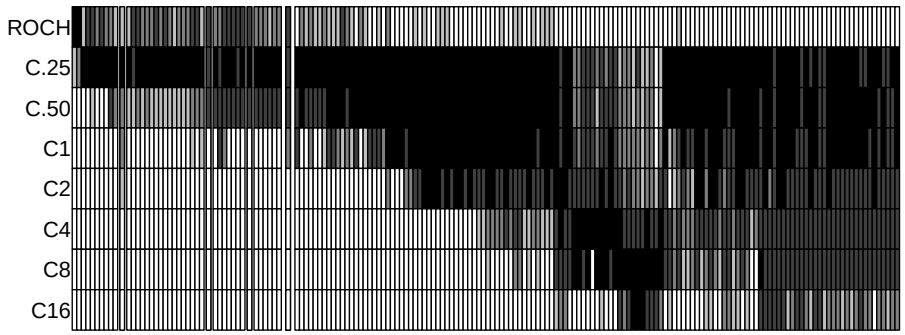
```
         X1     X2     X3     X4   X5
ROCH 16.848  7.000  2.156  0.235  0.0
C.25 74.130 83.462 90.938 67.353  7.6
C.50 17.413 52.308 83.438 63.603  7.6
C1    0.630 22.115 70.312 53.015  8.6
C2    0.087  0.577 45.000 41.191 14.0
C4    0.000  0.038  6.125 38.162 43.0
C8    0.000  0.000  0.125 28.824 78.0
C16   0.000  0.000  0.000  7.691 41.0
```

L'ordination suggère la classification. La classification renvoie une ordination. Il s'agit de modèles. Et si on regardait les données ?

```
table.paint(t(mil), clabel.c = 0, cleg = 0)
```




```
table.paint(t(millog), x = sort(pcamil$li[, 1]), clabel.c = 0, cleg = 0)
```



8 Truites et valeurs propres

J.M. Lascaux a étudié 306 truites [9]

```
data(lascaux)
names(lascaux$meris)
```

```
[1] "rd"    "ra"    "rpelg" "rpecg" "caec"
```

J.M. Lascaux a mesuré 5 variables méristiques :

`rd` Nombre de rayons à la dorsale

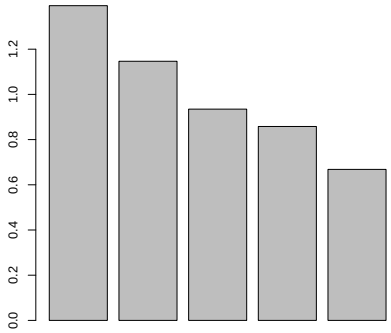
`ra` Nombre de rayons à l'anale

`rpelg` Nombre de rayons à la pelvienne gauche

`rpecg` Nombre de rayons à la pectorale gauche

`caec` Nombre de caeca pyloriques

```
pcmeris <- dudi.pca(lascaux$meris, scan = F)
barplot(pcmeris$eig)
```




```
cor(lascaux$meris)
```

```
             rd          ra       rpelg      rpecg        caec
rd     1.00000000  0.31108004  0.02633742 0.07528987 -0.06038263
ra     0.31108004  1.00000000 -0.05677408 0.12549442 -0.11191258
rpelg  0.02633742 -0.05677408  1.00000000 0.04819827  0.10350996
rpecg  0.07528987  0.12549442  0.04819827 1.00000000  0.06108539
caec  -0.06038263 -0.11191258  0.10350996 0.06108539  1.00000000
```

```
cor.test(lascaux$meris$ra, lascaux$meris$rpecg)
```

```
        Pearson's product-moment correlation
data:  lascaux$meris$ra and lascaux$meris$rpecg
t = 2.2055, df = 304, p-value = 0.02817
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01356170 0.23432087
sample estimates:
      cor
0.1254944
```

Conclure sur la nature de ce type de variables.
J.M. Lascaux a mesuré 35 variables morphologiques :

```
names(lascaux$morpho)
```

```
 [1] "LS"     "MD"     "MAD"    "MAN"    "MPEL"   "MPEC"   "DAD"    "DC"     "DAN"
[10] "DPEL"   "DPEC"   "ADC"    "ADAN"   "ADPEL"  "ADPEC"  "PECPEL" "PECAN"  "PECC"
[19] "PELAN"  "PELC"   "ANC"    "LPRO"   "DO"     "LPOO"   "LTET"   "HTET"   "LMAX"
[28] "LAD"    "LD"     "HD"     "LC"     "LAN"    "HAN"    "LPELG"  "LPECG"  "HPED"
[37] "ETET"
```

`LS` Longueur standard

`MD` Distance bout du museau - insertion de la dorsale

`MAD` Distance bout du museau - insertion de l'adipeuse

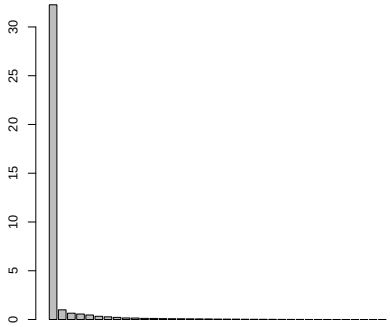
`MAN` Distance bout du museau - insertion de l'anale

`MPEL` Distance bout du museau - insertion de la pelvienne
`MPEC` Distance bout du museau - insertion de la pectorale
`DAD` Distance insertion de la dorsale - insertion de l'adipeuse
`DC` Distance insertion de la dorsale - départ de la caudale
`DAN` Distance insertion de la dorsale - insertion de l'anale
`DPEL` Distance insertion de la dorsale - insertion de la pelvienne
`DPEC` Distance insertion de la dorsale - insertion de la pectorale
`ADC` Distance insertion de l'adipeuse - départ de la caudale
`ADAN` Distance insertion de l'adipeuse - insertion de l'anale
`ADPEL` Distance insertion de l'adipeuse - insertion de la pelvienne
`ADPEC` Distance insertion de l'adipeuse - insertion de la pectorale
`PECPEL` Distance insertion de la pectorale - insertion de la pelvienne
`PECAN` Distance insertion de la pectorale - insertion de l'anale
`PECC` Distance insertion de la pectorale - départ de la caudale
`PELAN` Distance insertion de la pelvienne - insertion de l'anale
`PELC` Distance insertion de la pelvienne - départ de la caudale
`ANC` Distance insertion de l'anale - départ de la caudale
`LPRO` Longueur préorbitale
`DO` Diamètre de l'orbite
`LPOO` Longueur postorbitale
`LTET` Longueur de la tête
`HTET` Hauteur de la tête (en passant par au milieu de l'orbite)
`LMAX` Longueur de la mâchoire supérieure
`LAD` Longueur de l'adipeuse
`LD` Longueur de la dorsale
`HD` Hauteur de la dorsale
`LC` Longueur de la caudale
`LAN` Longueur de l'anale
`HAN` Hauteur de l'anale
`LPELG` Longueur de la pelvienne gauche
`LPECG` Longueur de la pectorale gauche
`HPED` Hauteur du corps au niveau du pédoncule caudal
`ETET` Largeur de la tête (au niveau des orbites)

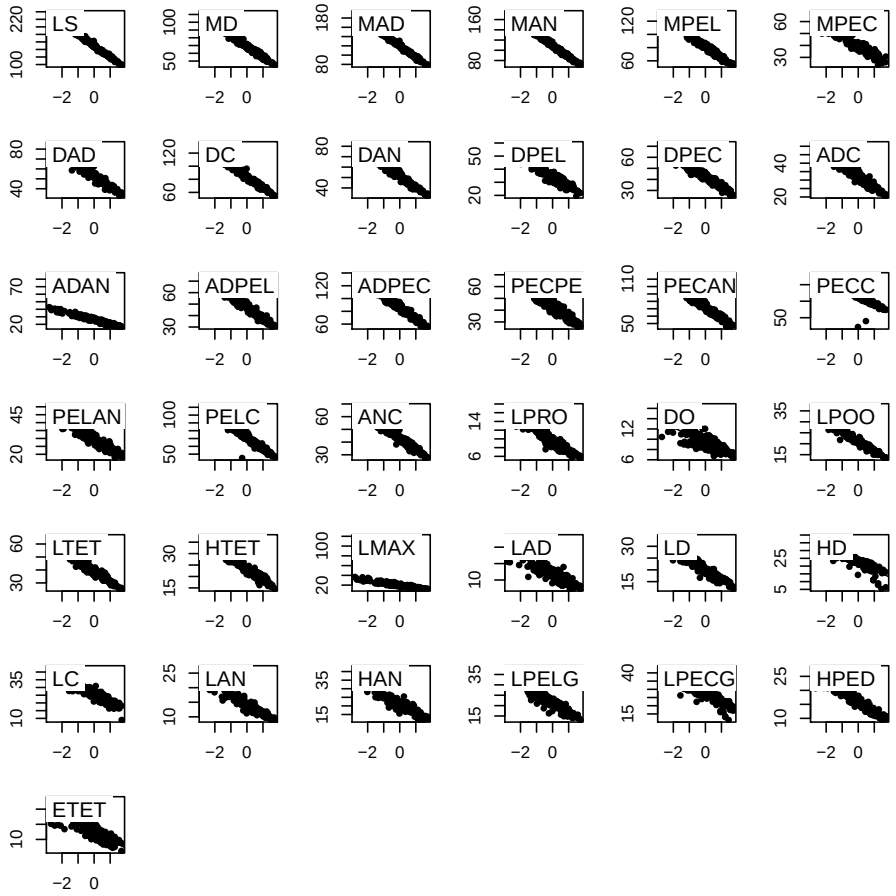
```
m <- na.omit(lascaux$morpho)
m.acp <- dudi.pca(m, scann = F)
```

L'effet taille est écrasant :

```
barplot(m.acp$eig)
```

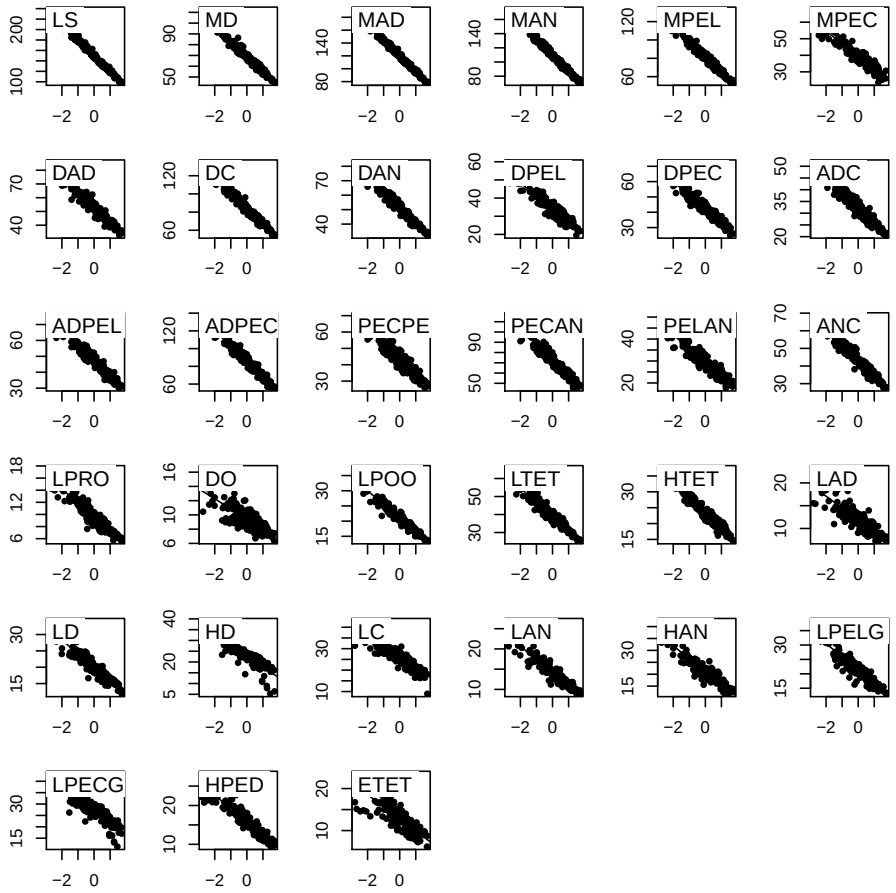


```
score(m.acp)
```



Une truite qui a perdu un morceau de nageoire dans une bataille est facilement repérée ! Ce qui reste quand on enlève l'effet taille concerne, grossièrement parlant, *la forme*. Les variables 13, 18, 20 et 27 présentent des *outliers*. Elles sont exclues :

```
m <- na.omit(lascaux$morpho)[, -c(13, 18, 20, 27)]
m.acp <- dudi.pca(m, scan = F, nf = 3)
score(m.acp)
```

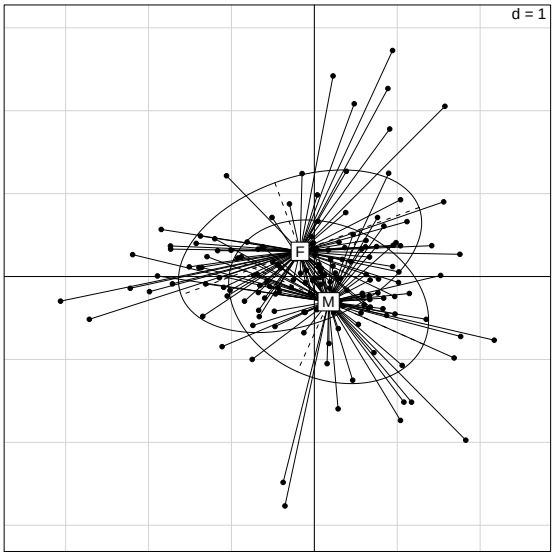


Noter que le problème esr résolu, avec brutalité, certes !

```
names(lascaux$sex) <- row.names(lascaux$morpho)
names(lascaux$gen) <- row.names(lascaux$morpho)
```

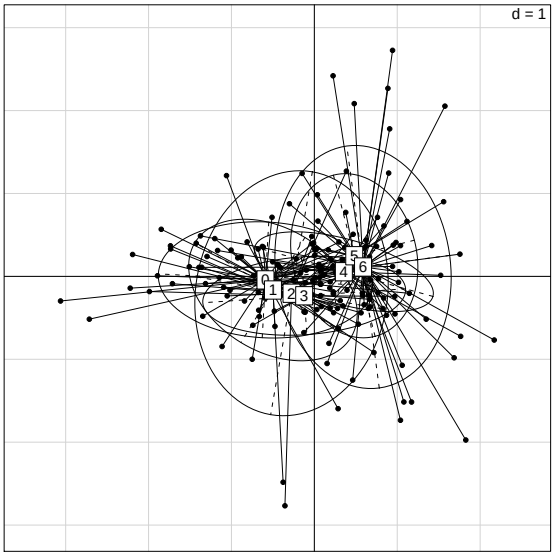
Enlever des données le modèle issu de l'effet taille ou dépouiller l'analyse à partir du facteur 2 donne strictement le même résultat. C'est l'illustration qu'une ACP peut faire deux choses de nature radicalement différente. Le plan 2-3 est directement une analyse de la forme :

```
s.class(m.acp$li[, 2:3], lascaux$sex[as.numeric(row.names(m.acp$li))])
```



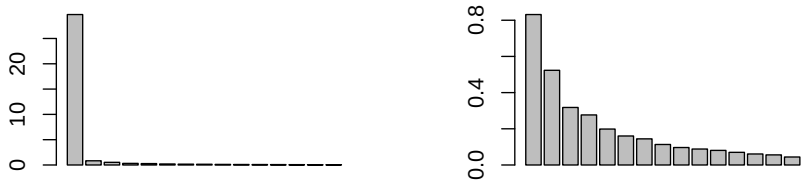
Dans la forme, il y a certainement une composante sexuelle.

```
s.class(m.acp$li[, 2:3], lascaux$gen[as.numeric(row.names(m.acp$li))])
```



Dans la forme, il y a certainement une composante génétique.

```
par(mfrow = c(1, 2))
barplot(m.acp$eig[1:15])
barplot(m.acp$eig[2:16])
```

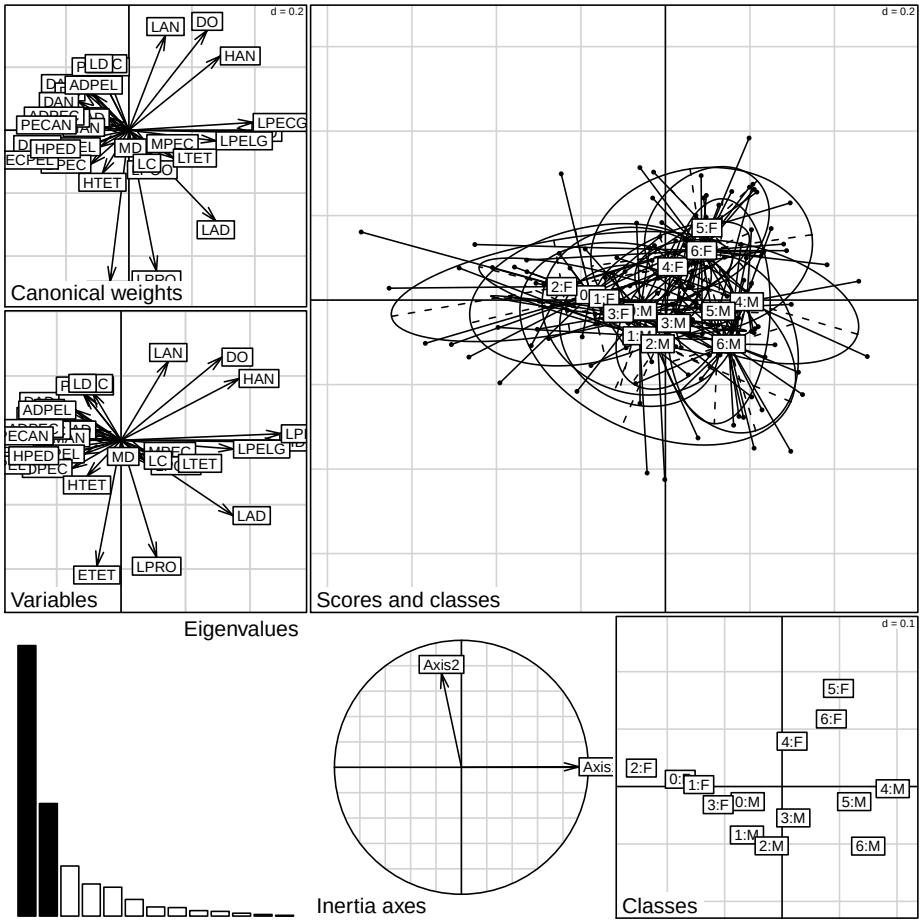


Une ACP peut ainsi en cacher une autre. On peut analyser l'effet taille par une ACP non centrée sur *le tableau des log doublement centré* :

```
mlog <- log(m)
mlog0 <- bicenter.wt(mlog)
mlog0 <- data.frame(mlog0)
names(mlog0) <- names(m)
mlog0.acp <- dudi.pca(mlog0, scan = F, scal = F, cent = F)
```

Ceci indique que l'introduction des deux facteurs directement dans l'analyse est pertinente. L'ACP inter-classe est celle des centres de gravité des sous-nuages.

```
croi <- lascaux$gen[as.numeric(row.names(m.acp$li))]:lascaux$sex[as.numeric(row.names(m.acp$li))]
plot(between(mlog0.acp, croi, scan = F))
```



Ces truites sont classées en 7 groupes génétiques par le nombre d'allèles méditerranéens. La variable génétique prend les modalités 0 (0 allèle méditerranéen, homozygotes atlantiques, truites dites modernes), 1 à 5 (respectivement 1 à 5 allèles méditerranéens) et 6 (6 allèles méditerranéens, homozygotes méditerranéens, truites dites ancestrales). Les modernes sont à gauche, les ancestrales à droite, les femelles en haut et les mâles en bas. Le code des variables permettra de poursuivre.

J.M. Lascaux a mesuré 15 variables de coloration de la robe :

```
names(lascaux$colo)
 [1] "PRAD"     "PNAD"     "PRAA"     "PNAA"     "PRDA"     "PNDA"     "PRLI"
 [8] "PRINF"    "PNINF"    "PRSUP"    "PNSUP"    "PNO"      "PNPERIOP" "PRD"
[15] "PND"
```

`PRAD` Nombre de points rouges avant l'aplomb de la Dorsale

`PNAD` Nombre de points noirs avant l'aplomb de la Dorsale

`PRAA` Nombre de points rouges après l'aplomb de l'Anale

`PNAA` Nombre de points noirs après l'aplomb de l'Anale

`PRDA` Nombre de points rouges entre aplomb Dorsale et aplomb Anale

`PNDA` Nombre de points noirs entre aplomb Dorsale et aplomb Anale

`PRLI` Nombre de points rouges sur la ligne latérale

`PRINF` Nombre de points rouges au dessous de la ligne latérale

`PNINF` Nombre de points noirs au dessous de la ligne latérale

`PRSUP` Nombre de points rouges au dessus de la ligne latérale

`PNSUP` Nombre de points noirs au dessus de la ligne latérale

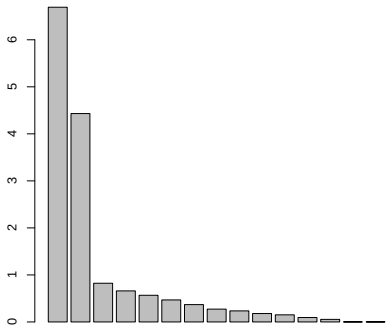
`PNO` Nombre de points noirs sur l'opercule

`PNPERIOP` Nombre de tâches noires à la périphérie de l'opercule

`PRD` Nombre de points rouges sur la Dorsale

`PND` Nombre de points noirs sur la Dorsale

```
pcacol <- dudi.pca(lascaux$colo, scann = F)
barplot(pcacol$eig)
```



Questions

1) Trouver en quoi une partie de la corrélation est arte-factuelle dans ce tableau.
2) Interpréter les composantes.
3) Trouve-t-on dans la coloration de la robe une composante génétique ?
4) Trouve-t-on dans la coloration de la robe une composante environnementale ?

J.M. Lascaux a mesuré 15 variables ornementales qualitatives :

```
names(lascaux$ornem)
 [1] "ocpr"   "ocpn"   "maju"   "taop"   "pntet"  "frd"    "ptsad"  "frad"   "fran"
[10] "frpel"  "frc"    "ptsdos" "coufl"  "coupn"  "zeb"
```

`ocpr` Ocelles autour des points rouges (1 nulles, 2 faibles, 3 marquées)

`ocpn` Ocelles autour des points noirs (1 nulles, 2 faibles, 3 marquées)

`maju` Marques juvéniles (1 absence, 2 présence)

`taop` Tache operculaire (1 absence, 2 présence)

`pntet` Points noirs sur la tête (1 absence, 2 présence)

`frd` Frange de la dorsale (1 aucune, 2 blanche, 3 blanche et noire)

`ptsad` Points sur l'adipeuse (1 absence, 2 présence)

`frad` Frange de l'adipeuse (1 aucune, 2 rouge, 3 très rouge)

fran Frange de l'anale (1 aucune, 2 blanche, 3 blanche et noire)

frpel Frange des pelviennes (1 aucune, 2 blanche, 3 blanche et noire)

frc Frange de la caudale (1 aucune, 2 plus ou moins rouge)

ptsdos Points sur le dos (1 absence, 2 présence)

coufl Couleur des flancs (1 brun-jaune, 2 gris, 3 blanc argenté)

coupn Contour des points noirs du flan (1 net, 2 flou)

zeb Zébrures sur les flancs (1 absence, 2 présence)

```
summary(lascaux$ornem)
```

```
ocpr    ocpn    maju    taop    pntet   frd     ptsad   frad    fran    frpel
1:203   1:238   1:111   1: 57   1: 89   1:178   1:218   1: 51   1: 27   1:107
2: 66   2: 48   2:195   2:249   2:217   2: 14   2: 88   2:251   2:173   2:146
3: 37   3: 20                           3:114           3:  4   3:106   3: 53
frc     ptsdos  coufl   coupn   zeb
1:249   1:248   1:213   1:239   1:278
2: 57   2: 58   2: 74   2: 67   2: 28
                3: 19
```

On retrouvera ces données plus tard.

9 Voir le tableau

Source : Devillers, J., J. Thioulouse, and W. Karcher. 1993 [2].

```
data(toxicity)
head(toxicity$tab)
```

```
    V1    V2    V3    V4    V5    V6    V7
1 1.224 1.824 3.004 3.323 2.579 5.241 6.264
2 1.438 1.540 2.630 2.801 2.833 5.284 4.024
3 0.679 1.893 2.855 2.293 2.547 4.943 3.163
4 1.001 1.390 2.807 2.939 3.636 4.926 6.103
5 1.097 1.747 2.815 3.086 3.170 5.303 6.364
6 0.967 1.666 2.870 3.198 3.358 5.441 6.120
```

```
toxicity$species
```

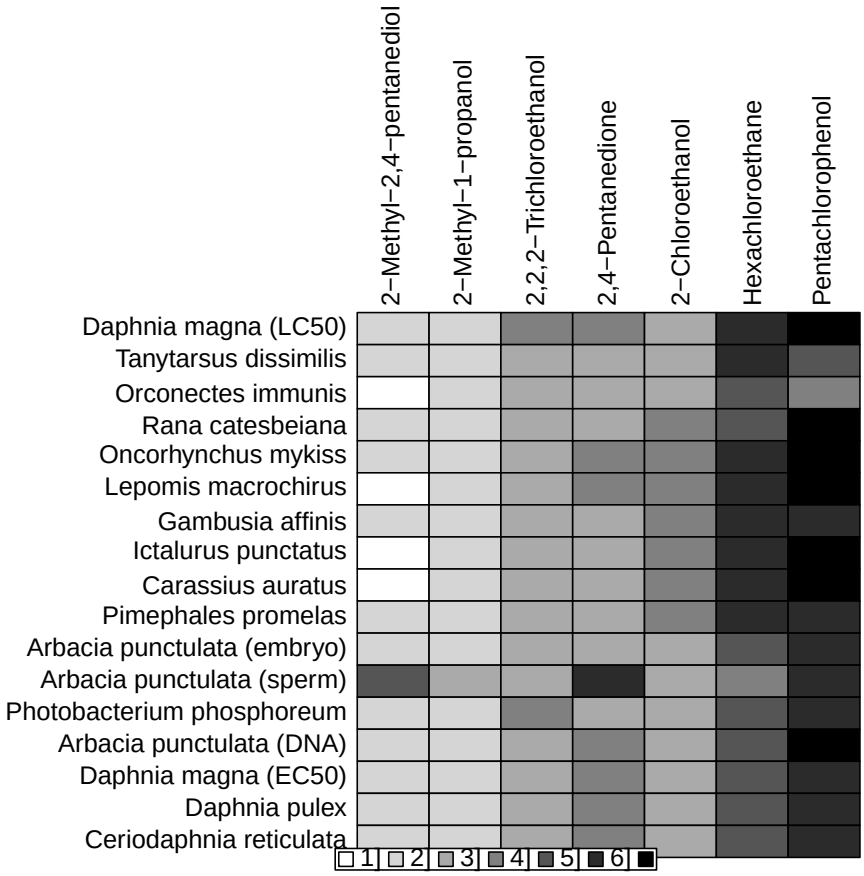
```
 [1] "Daphnia magna (LC50)"        "Tanytarsus dissimilis"
 [3] "Orconectes immunis"          "Rana catesbeiana"
 [5] "Oncorhynchus mykiss"         "Lepomis macrochirus"
 [7] "Gambusia affinis"            "Ictalurus punctatus"
 [9] "Carassius auratus"           "Pimephales promelas"
[11] "Arbacia punctulata (embryo)" "Arbacia punctulata (sperm)"
[13] "Photobacterium phosphoreum"  "Arbacia punctulata (DNA)"
[15] "Daphnia magna (EC50)"        "Daphnia pulex"
[17] "Ceriodaphnia reticulata"
```

```
toxicity$chemicals
```

```
[1] "2-Methyl-2,4-pentanediol" "2-Methyl-1-propanol"      "2,2,2-Trichloroethanol"
[4] "2,4-Pentanedione"         "2-Chloroethanol"          "Hexachloroethane"
[7] "Pentachlorophenol"
```

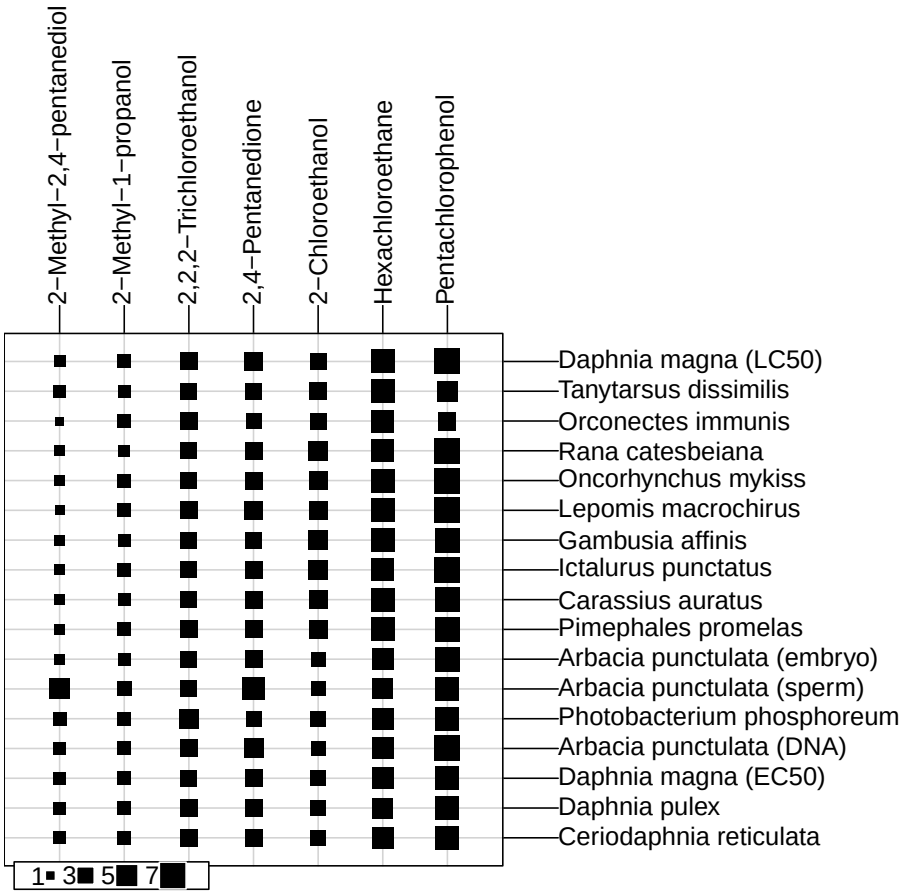
On a la toxicité de 7 molécules (colonnes) sur 16 cibles (lignes) exprimée en : -log(mol/litre)

```
data(toxicity)
table.paint(toxicity$tab, row.lab = toxicity$species, col.lab = toxicity$chemicals)
```




```
table.value(toxicity$tab, row.lab = toxicity$species, col.lab = toxicity$chemicals)
f1 <- as.factor(row(as.matrix(toxicity$tab)))
f2 <- as.factor(col(as.matrix(toxicity$tab)))
tox <- unlist(toxicity$tab)
anova(lm(tox ~ f1 + f2))
```

```
Analysis of Variance Table
Response: tox
          Df  Sum Sq Mean Sq  F value Pr(>F)
f1        16   5.933   0.371   1.1217 0.3467
f2         6 244.098  40.683 123.0630 <2e-16 ***
Residuals 96  31.736   0.331
---
Signif. codes:  0
```

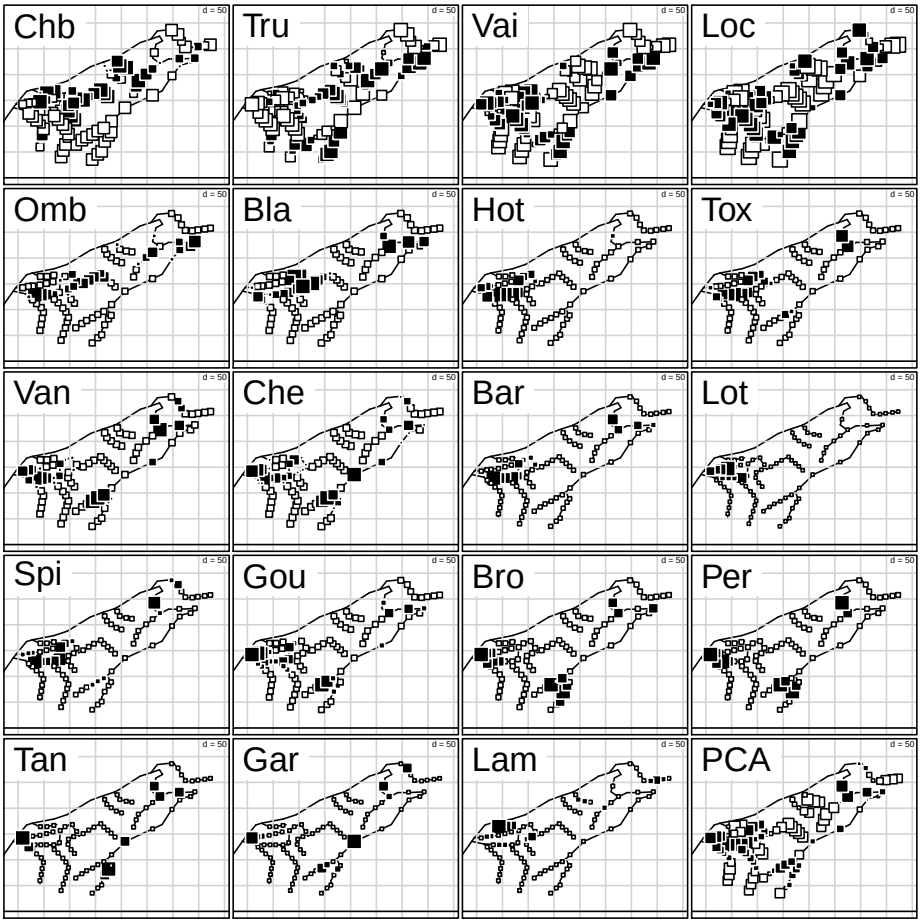


On a fait le tour. Commenter.

10 Notes d'abondance

Quand on pense l'ACP comme approche d'un échantillon d'une loi normale multivariée, son usage sur un tableau de notes binaires (0-1) est sans objet. On dit à l'utilisateur "*qu'il n'a pas le droit*". Quand on pense l'ACP comme une opération géométrique ou la maximisation d'une forme quadratique, on dit à l'utilisateur "qu'il a le droit". Quand un lecteur formé à une école juge le manuscrit d'un auteur formé à l'autre école, le dialogue est sommaire et tous les coups sont permis. Le même problème se pose pour les notes d'abondance (souvent de 0 à 7 en phytosociologie). L'ACP d'un tableau floro-faunistique est valide comme première approche d'objectifs précis. Elle a marqué les pionniers [5] par la facilité avec laquelle elle permet des cartes de synthèse.

```
data(jv73)
par(mfrow = c(5, 4))
for (m in 1:19) {
    s.value(jv73$xy, jv73$poi[, m] - mean(jv73$poi[, m]), incl = F,
        contour = jv73$contour, sub = names(jv73$poi)[m], possub = "topleft",
        csub = 3, cleg = 0)
}
s.value(jv73$xy, dudi.pca(jv73$poi, scal = F, scan = F)$li[, 1],
    incl = F, contour = jv73$contour, sub = "PCA", possub = "topleft",
    csub = 3, cleg = 0)
```



L'ACP est en fait souvent discutée sur ce genre de tableau parce que c'est une méthode qui introduit une grande dissymétrie entre le traitement des lignes et des colonnes. Elle est centrée sur la différence entre les lignes et la redondance entre colonnes. Dans un tableau sites-espèces, on s'appuie sur la covariance entre espèces pour discriminer les sites ; dans un tableau espèces-sites, on fait l'inverse. Seule l'analyse des correspondances fait les deux, c'est-à-dire, donne un compromis des deux objectifs. L'important est de *ne pas utiliser la procédure à l'aveugle*. La difficulté duale est la grande diversité des structures de ce genre de tableaux et l'objectif est de trouver le particulier par le biais de procédure générale.

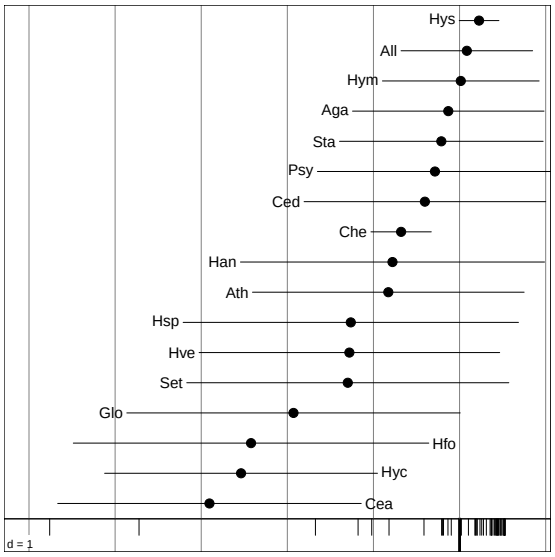
```
data(trichometeo)
```

`trichometeo$fau` comporte 49 pièges et 17 espèces.

```
t1 <- data.frame(apply(log(trichometeo$fau + 1), 2, function(x) x/sum(x)))
s.distri(dudi.pca(t1, scan = F, scale = F)$l1, t1, cell = 0, csta = 0.3,
    clab = 1)
```

Les dénombrements sont généralement transformés par $x \mapsto \log(x+1)$. Le tableau [11] est passé en pourcentage par colonnes (espèces). Un taxon est une distribution de fréquence entre sites sur une colonne. La moyenne est $1/n$ et la colonne centrée est l'écart entre la distribution de l'espèce et la distribution uniforme. Une composante principale est un score normalisé des sites et la coordonnée de l'espèce est exactement la position moyenne sur ce score. L'analyse maximise la somme des carrés des écarts des positions moyennes à l'origine qui est la position du taxon indifférent :

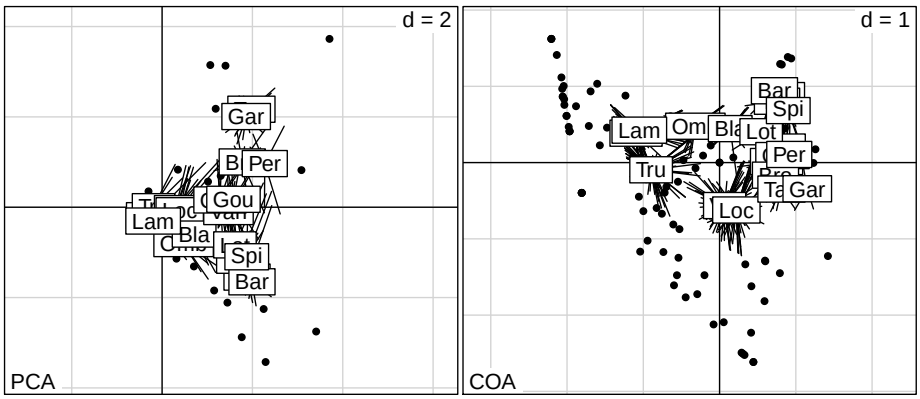
```
sco.distri(dudi.pca(t1, scan = F, scale = F)$l1[, 1], t1)
```



L'essentiel est que certains relevés ont décalé la majorité des espèces dans le même sens (une configuration météorologique particulière provoque l'émergence simultanée des trichoptères). Autre ambiguïté : la covariance de deux variables est une mesure de ressemblance même quand elle est négative (une corrélation de -1 indique qu'on a

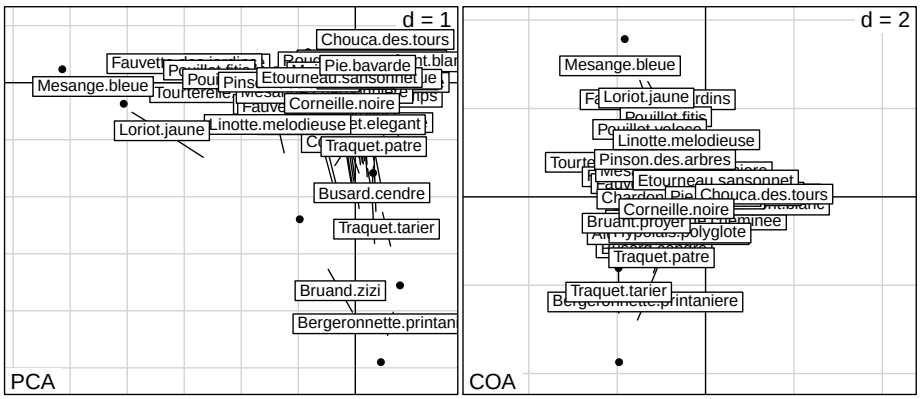
mesuré exactement la même chose au signe près). Pour deux espèces, c'est au contraire une mesure de différence : une espèce disparaît quand l'autre apparaît. L'ACP a alors le statut de méthode qui fait une double typologie. Le même calcul n'a pas le même sens expérimental.

```
t1 <- data.frame(apply(jv73$poi, 2, function(x) x/sum(x)))
par(mfrow = c(1, 2))
s.distri(dudi.pca(t1, scan = F, scale = F)$l1, t1, cstar = 0.3,
    clab = 1, cell = 0, axesel = F, sub = "PCA")
s.distri(dudi.coa(jv73$poi, scan = F)$l1, jv73$poi, cstar = 0.3,
    clab = 1, cell = 0, axesel = F, sub = "COA")
```



Les deux analyses parlent de la typologie de distributions des espèces. La carte de l'ACP est bien meilleure en ce sens qu'elle indique le gradient amont-aval de richesse croissante (salmonidés-cyprinidés) et dans la seconde partie le gradient de vitesse de courant (lénitique-lentique). Les contraintes de l'AFC déforment sans gain précis cette vision juste du réseau hydrographique.

```
data(aviurba)
names(aviurba$fau) <- aviurba$species.names.fr
t1 <- data.frame(apply(aviurba$fau, 2, function(x) x/sum(x)))
par(mfrow = c(1, 2))
s.distri(dudi.pca(t1, scan = F, scale = F)$l1, t1, cstar = 0.3,
    clab = 0.75, cell = 0, axesel = F, sub = "PCA")
s.distri(dudi.coa(aviurba$fau, scan = F)$l1, aviurba$fau, cstar = 0.3,
    clab = 0.75, cell = 0, axesel = F, sub = "COA")
```

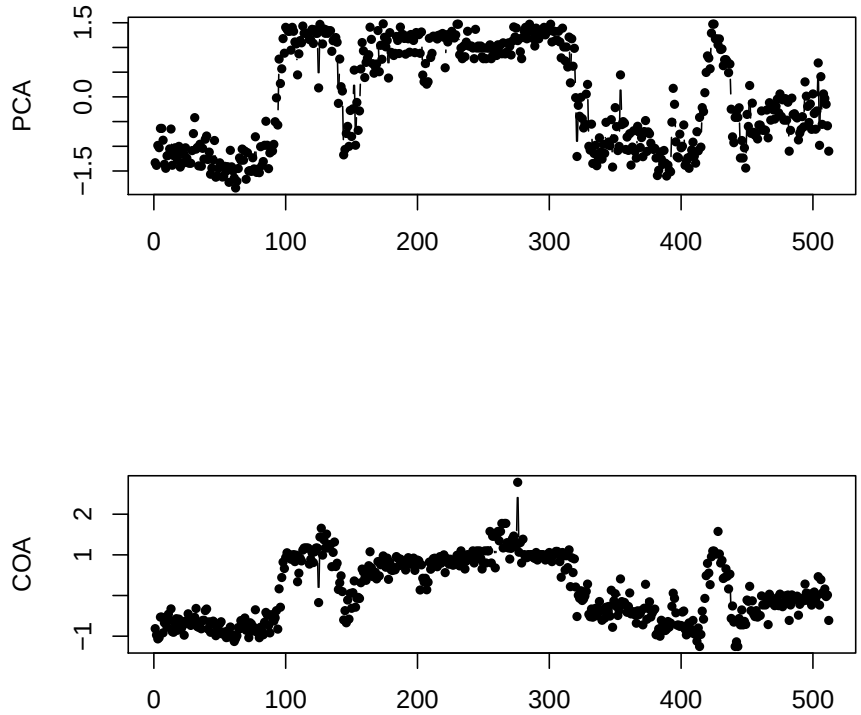


Deux gradients centre-ville vers rural fermé et centre-ville vers rural ouvert pour l'ACP correspondent aux deux gradients urbain-rural et ouvert-fermé pour l'AFC. L'avantage est à la seconde. Il n'y a pas de meilleures méthodes, il n'y a que de meilleures interactions entre une méthode et un objet. Le même raisonnement s'applique aux tableaux espèces-relevés en pourcentage par lignes (profils espèces) centrés par lignes. *Les moyennes se calculent toujours par espèces et la normalisation par espèces est à éviter*. Si on désire d'abord éviter toute réflexion préalable, on peut préférer l'analyse des correspondances (on aura souvent un résultat acceptable mais ce peut être une erreur totale).

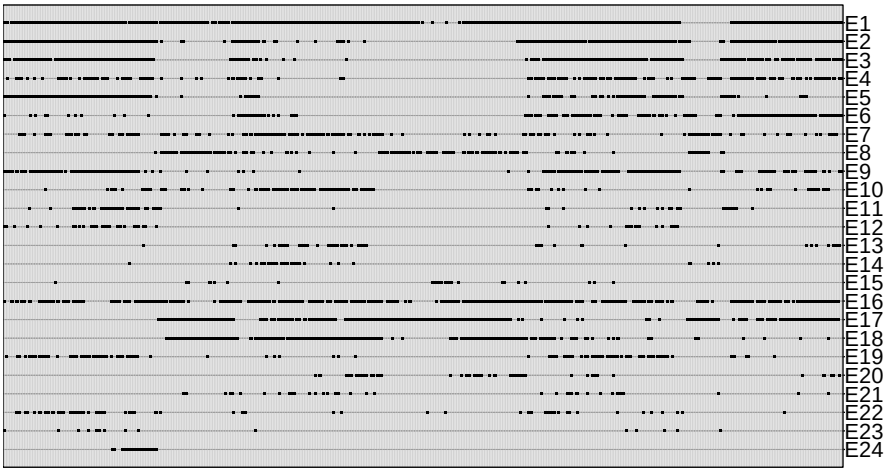
```
data(steppe)
```

512 relevés par quadrat sur un transect de 5.12 km ont enregistré la présence / absence de 37 espèces (milieu steppique). A gauche l'ACP du tableau non modifié,à droite son AFC :

```
par(mfrow = c(2, 1))
plot(dudi.pca(steppe$tab, scan = F, scale = F)$li[, 1], pch = 20,
    ylab = "PCA", xlab = "", type = "b")
plot(dudi.coa(steppe$tab, scan = F)$li[, 1], pch = 20, ylab = "COA",
    xlab = "", type = "b")
```



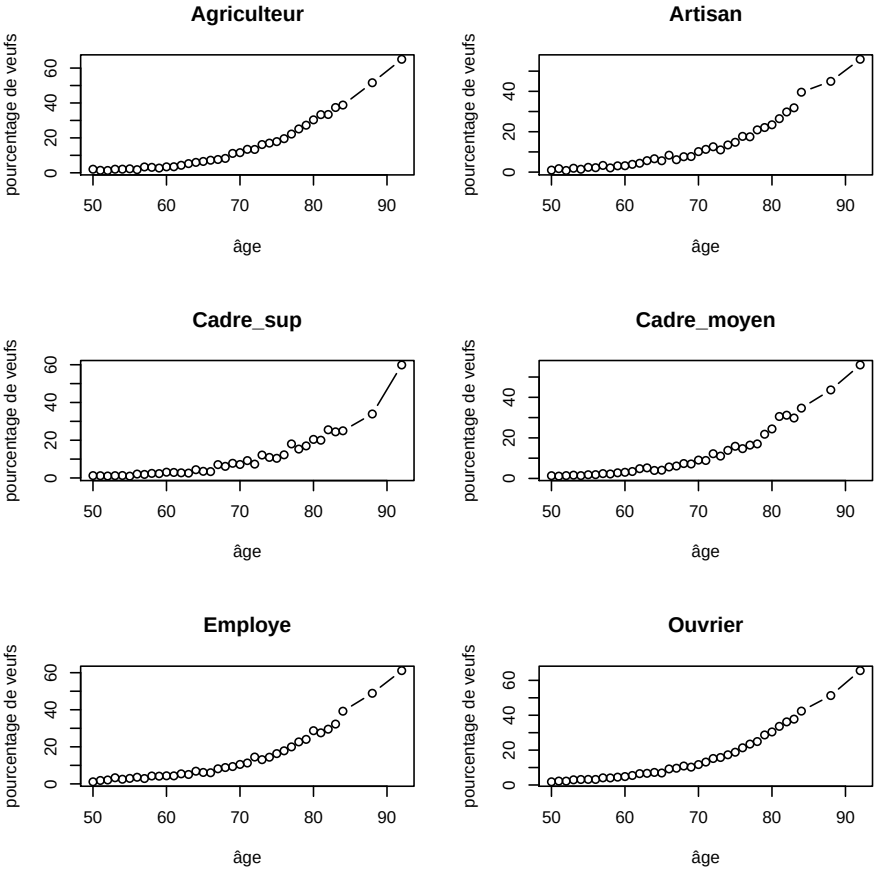

```
table.value(t(steppe$tab[1:512, 1:24]), clabel.c = 0, csize = 0.05,
    cleg = 0)
```



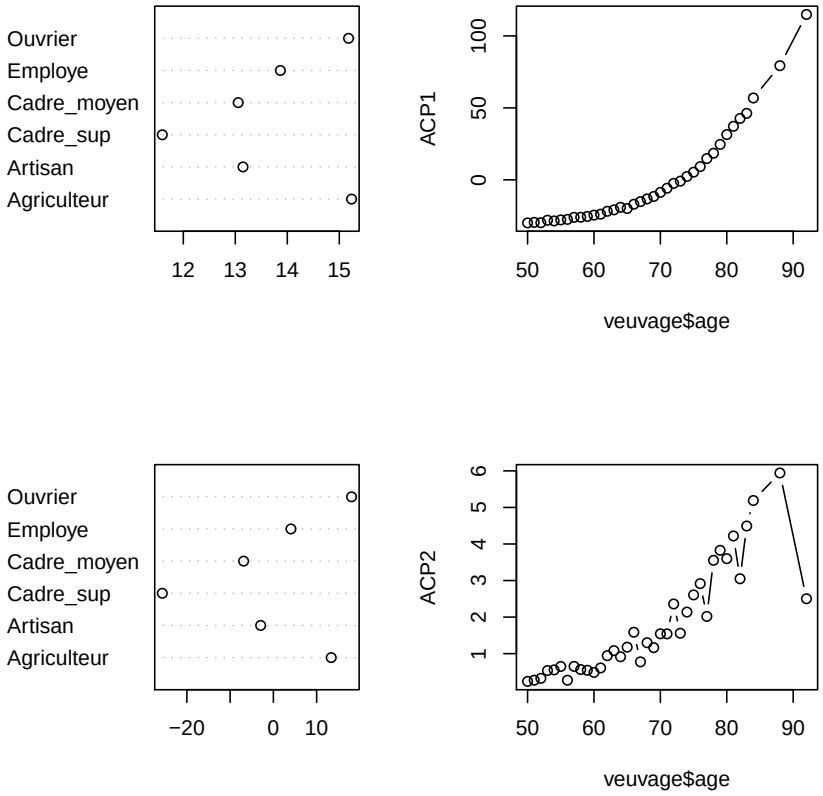
On obtient cette fois un résultat très voisin. L'ACP n'est cependant qu'un point de départ dans la question finalement difficile de l'*ordination* [3] des tableaux floro-faunistiques. Ces exemples illustrent l'insertion de la procédure dans un comportement adapté à chaque cas.

11 Ne pas se tromper de centrage

```
data(veuvage)
par(mfrow = c(3, 2))
for (j in 1:6) plot(veuvage$age, veuvage$tab[, j], xlab = "âge",
    ylab = "pourcentage de veufs", type = "b", main = names(veuvage$tab)[j])
```




```
par(mfrow = c(2, 2))
PCA1 <- dudi.pca(veuvage$tab, scal = F, scan = F)
PCA2 <- dudi.pca(t(veuvage$tab), scal = F, scan = F)
dotchart(PCA1$co[, 1], lab = names(veuvage$tab))
plot(veuvage$age, PCA1$li[, 1], type = "b", ylab = "ACP1")
dotchart(PCA2$li[, 1], lab = names(veuvage$tab))
plot(veuvage$age, PCA2$co[, 1], type = "b", ylab = "ACP2")
```



Ces deux analyses se ressemblent fort et disent deux choses radicalement différentes. Lesquelles ?

12 Information supplémentaire

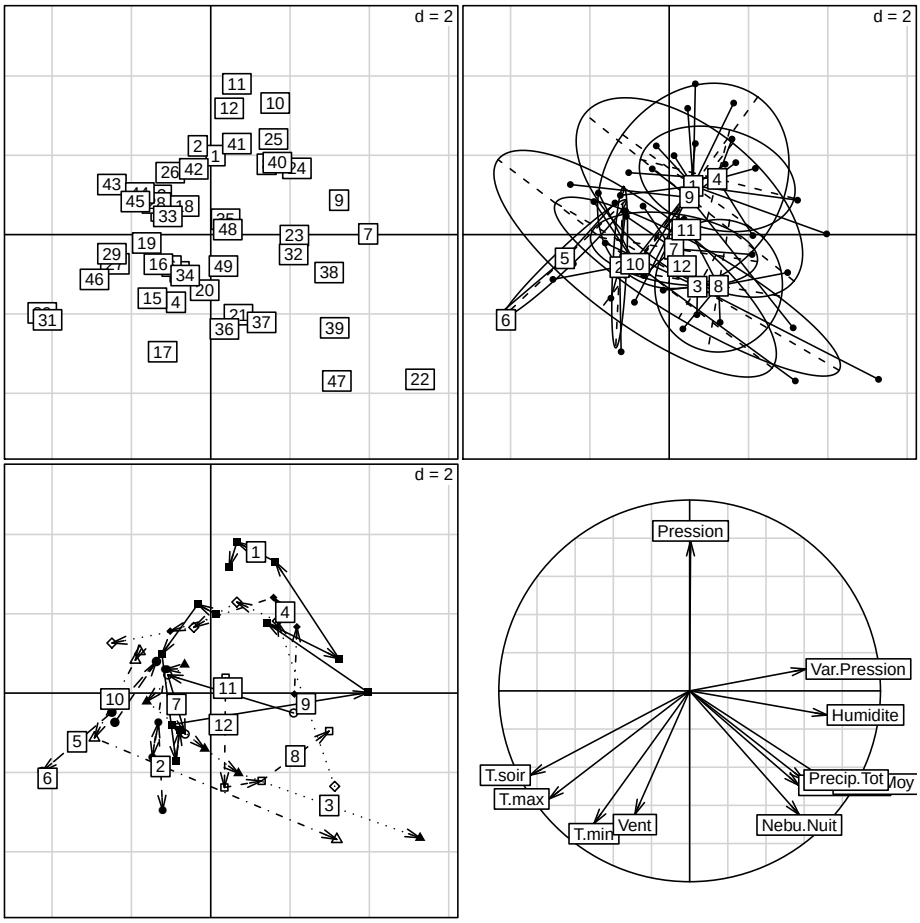
Plusieurs auteurs ont souligné que le terme *supplémentaire* s'applique souvent de manière abusive à tout ce qui ne fait pas partie du tableau des données alors qu'on devrait bien réserver le terme projection en individus supplémentaires à une opération géométrique précise.

```
data(trichometeo)
names(trichometeo)
```

```
[1] "fau"    "meteo" "cla"
```

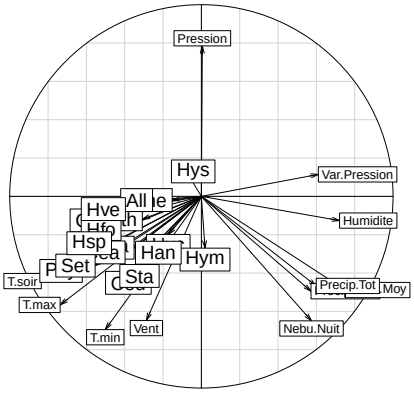
On a un tableau de 11 variables météorologiques [11]. Chaque ligne est une journée d'été.

```
meteo.pca <- dudi.pca(trichometeo$meteo, scan = F)
par(mfrow = c(2, 2))
s.label(meteo.pca$li)
s.class(meteo.pca$li, trichometeo$cla)
s.traject(meteo.pca$li, trichometeo$cla)
s.corcircle(meteo.pca$co)
```



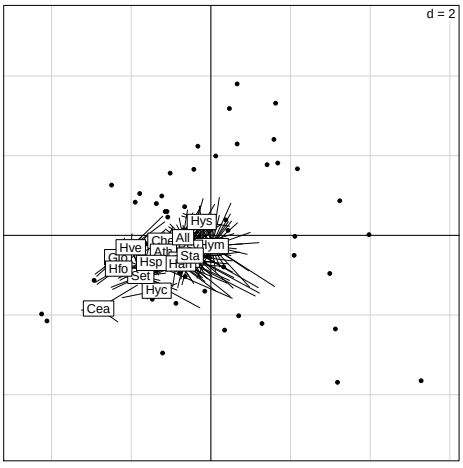
Il y a beaucoup de redondance dans les mesures. La carte des lignes supporte un premier type d'information supplémentaire. Chaque journée est suivie d'une nuit pendant laquelle fonctionne un piège lumineux destiné à la capture des trichoptères émergeant de la rivière. Ce piège est récolté régulièrement mais non chaque jour (week-end ?). On a gardé le résultat obtenu pour une nuit unique de piégeage et le facteur 'cla' donne les groupes de nuits consécutives. Ces groupes sont représentés sur la carte factorielle : il s'agit simplement d'*information complémentaire*. Sur les trajectoires, on voit qu'il faut lire une sorte de mouvement circulaire. On note la succession haute pression (beau temps) puis fortes températures puis précipitations (orages d'été) caractéristiques du temps estival de la région.
Le tableau 'fau' associé contient les abondances d'animaux capturés dans le piège triés par espèce. La question porte sur l'influence des variables météorologiques sur l'abondance des piégeages lumineux. Le tableau faunistique a 17 espèces (variables). Les variables faunistiques sont supplémentaires :

```
tabsup <- scalewt(log(trichometeo$fau + 1))
w <- supcol(meteo.pca, tabsup)
s.corcircle(meteo.pca$co)
s.corcircle(w$cosup, add.p = T, clab = 1.5)
```



Les projections des variables supplémentaires normalisées (vecteurs de norme 1) donnent des coordonnées qui sont des coefficients de corrélation avec les coordonnées factorielles. Ces corrélation sont faibles mais de même signe. Les nouvelles variables ont été projetées sur les deux composantes principales. Il s'agit d'une *projection de variables supplémentaires* au sens euclidien. Mais on peut considérer également que chaque espèce définit de l'information supplémentaire pour le plan des individus :

```
s.distri(meteo.pca$li, log(trichometeo$fau + 1), csta = 0.3, cell = 0,
    clab = 1)
```



On a ainsi superposé les moyennes des positions des espèces. On pourra aussi représenter l'abondance des espèces sur le plan. Ici domine l'idée d'une combinaison de variables météorologiques ayant une influence commune sur les émergences de tous les taxons. Notons enfin qu'il arrive que de véritables projections euclidiennes soient également des représentations par moyennes de distribution et que les notions d'individus supplémentaires et d'information supplémentaire se confondent.
Quoiqu'il en soit le graphique appliqué à la statistique multidimensionnelle est un moyen d'expression. Cela suppose quelques libertés dans les choix et la référence à un comportement "conforme à la règle" peut être le signe d'une certaine absence d'imagination. Ce n'est évidemment pas une raison pour faire n'importe quoi.
En conclusion, l'analyse en composantes principales est une procédure simple utilisable de multiples façons. Elle donne à voir dans $\mathbb{R}^p$.

Références

[1] G. Carrel. *Caractérisation physico-chimique du Haut-Rhône français et de ses annexes : incidences sur la croissance des populations d'alevins*. PhD thesis, Université Claude Bernard, Lyon 1, 1986.

[2] J. Devillers, J. Thioulouse, and W. Karcher. Chemometrical evaluation of multispecies-multichemical data by means of graphical techniques combined with multivariate analyses. *Ecotoxicology and Environnemental Safety*, 26 :333–345, 1993.

[3] J. Estève. Les méthodes d'ordination : éléments pour une discussion. In J.M. Legay and R. Tomassone, editors, *Biométrie et Ecologie*, pages 223–250. Société Française de Biométrie, Paris, 1978.

[4] O. Gaschignard-Fossati. *Répartition spatiale des macroinvertébrés benthiques d'un bras vif du Rhône. Rôle des crues et dynamique saisonnière*. PhD thesis, Université Claude Bernard, Lyon 1, 1986.

[5] D.W. Goodall. Objective methods for the classification of vegetation iii. an essay in the use of factor analysis. *Australian Journal of Botany*, 2 :304–324, 1954.

[6] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A handbook of small data sets*. Chapman & Hall, London, 1994.

[7] P. Jolicœur and J.E. Mosimann. Size and shape variation in the painted turtle. a principal component analysis. *Growth*, 24 :339–354, 1960.

[8] J. Kervella. Analyse de l'attrait d'un produit : exemple d'une comparaison de lots de pêches. In *2émes journées européennes Agro-Industrie et Méthodes Statistiques*, pages 103–106. Association pour la Statistique et ses Utilisations, Paris, Nantes 13-14 juin 1991, 1991.

[9] J.M. Lascaux. *Analyse de la variabilité morphologique de la truite commune (Salmo trutta L.) dans les cours d'eau du bassin pyrénéen méditerranéen*. PhD thesis, INP Toulouse, 1996.

[10] R. Prodon and J.D. Lebreton. Breeding avifauna of a mediterranean succession : the holm oak and cork oak series in the eastern pyrénées. 1 : Analysis and modelling of the structure gradient. *Oïkos*, 37 :21–38, 1981.

[11] P. Usseglio-Polatera and Y. Auda. Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie*, 23 :65–79, 1987.