Consultation statistique avec le logiciel 

Corrélations canoniques et taux d'inertie

D. Chessel

10 juillet 2006

Un message de Karine Jacquet demande d'expliciter pourquoi l'usage de l'analyse des correspondances en écologie n'est pas toujours simple. La fiche donne des éléments de réponse.

Table des matières

1 Introduction

Le message de Karine Jacquet (Laboratoire Biogéographie et Ecologie des Vertébrés) est particulièrement précis. Citons le entièrement :

> Bonjour,
>
> Je souhaite vous poser une question pas directement tournée vers ade4 ou R, mais sur les résultats d'analyses, à propos du lien, après une Analyse Factorielle des Correspondances *par exemple*, entre les pourcentages de variances expliqués par les axes factoriels de l'analyse et les corrélations canoniques (racines carré des valeurs propres).
>
> En fait, on m'a reproché, *pour une publication*, le fait que j'ai un faible pourcentage de variance expliqué par mon premier axe, alors qu'on voit clairement que l'ordination est correcte. Aussi je me suis penchée sur les corrélations canoniques, et j'ai trouvé une forte corrélation canonique associée à ce premier axe.
>
> Or là, je suis confronté à un autre cas qui me laisse perplexe. Après une AFC sur une matrice de structure de végétation (contenant des pourcentages de recouvrements de strates de végétation, avec 96 relevés et 8 strates), j'obtiens, pour mon premier axe factoriel, une variance expliquée de 56,3% et une corrélation canonique de . . .0.26 ! Pourquoi cette corrélation canonique est -elle si faible *en comparaison* du pourcentage de variance expliquée ?
>
> Inversement, sur l'AFC d'abondance spécifique d'oiseaux (d'une matrice à 93 relevés et 58 espèces) j'obtiens, pour le premier axe factoriel, un pourcentage de variance expliqué de 9,6 % et une corrélation canonique de 0.51... Pourquoi est-elle si élevée en comparaison du pourcentage de variance ?
>
> Je ne comprends pas le lien entre ces deux facteurs . . .pourriez-vous m'y aider ?

J'ai mis en italique quelques mots qui explicite la demande.

2 Pourquoi la question ?

Karine dit *pour une publication*. C'est souvent *pour une publication*. Les biométriciens ne voient plus guère que des collègues qui ont des ennuis avec un comité de lecture. C'est normal : l'accès logiciel s'est démocratisé. Consulter un statisticien c'est simplement s'exposer à l'ignorance du lecteur et avoir encore plus d'ennuis. La remarque du type :

> Vous avez fait une AFC mais l'ACP est bien préférable.

est ausi répandue que l'assertion :

> Vous avez fait une ACP mais l'AFC s'impose.

voire presqu'autant que :

> Vous avez fait une AFC (ou une ACP) mais les méthodes linéaires sont inadaptées. Vous avez tort.

En fait, c'est une question de définition. On peut désigner par ACP un programme (`prcomp`, `princomp`, `dudi.pca`), une théorie (estimation gaussienne,

automodélisation, analyses géométriques), un outil pour examiner rapidement un tableau de données particulier.

Pourquoi les trois fonctions citées donnent-elles les mêmes résultats quand il y a plus d'individus que de variables alors que l'une des trois refusera de faire le calcul dans la configuration inverse ? Parce que le même calcul ne recouvre pas le même modèle.

Comment ? Un calcul peut avoir plusieurs significations très différentes ? Oui, et c'est là l'origine de la question posée. On peut calculer une quantité et lui donner des significations radicalement différentes. Ah bon ! Et voilà les questions de comité de lecture qui arrivent.

Karine dit *par exemple*. Elle a bien raison. Parce que l'histoire a retenu la valeur propre (inertie), la racine de la valeur propre (corrélation canonique) mais pas le carré de la valeur propre (variance vectorielle). Ou même une puissance quelconque. Ou même une fonction monotone quelconque de cette valeur propre.

Mais méfions nous. La racine de la valeur propre n'est une corrélation canonique qu'en AFC, mais pas en ACP tandis que le carré de la valeur propre est une composante de la structure (le dénominateur du RV d'Yves Escoufier) dans les deux. Donc il faut être précis.

Karine dit encore *en comparaison*. La vraie question est là. Pourquoi comparer ce qui n'est peut-être pas comparable ? La valeur propre comme taux d'inertie projetée est un concept imposé par J.P. Benzécri et très populaire en France. On peut lire souvent que l'AFC est de J.P. Benzécri [3], ou bien que l'AFC est née avec la thèse de B. Escofier [4]. C'est vrai pour la valeur propre mais pas pour sa racine !

L'algorithme de l'AFC est de H.O. Hirschfeld[7], inventeur du modèle des codages de double régression linéaire.

La racine de la valeur propre comme corrélation canonique est de E.J. Williams[11].

Croyez vous que les praticiens de l'école de l'inertie utilisent la corrélation canonique ? Jamais. J'ai entendu un illustre analyste de données traiter l'article de Williams de *chiure inférentielle*. La *chiure inférentielle* est pour l'analyste des données le sommet de l'injure. Mais ne vous faites pas de soucis. Des injures, les analystes de données, et les illustres en premier, en ont reçues leur comptant.

Karine a mis le doigt sur un bouton prodigieusement douloureux. Il y a un algorithme commun (ou presque, mais c'est déjà assez compliqué comme ça) *mais* il y a plusieurs manières de donner un sens au calcul. Les différents modèles, sans s'exclure sont très différents les uns des autres. Rares sont ceux qui les connaissent tous. Un jour j'ai expliqué à C.J.F. Ter Braak que j'avais trouvé une nouvelle interprétation de l'AFC. Il m'a répondu : perdu, c'est dans l'article de W.J. Heiser[5]. Nous avons alors ouvert le livre de P. Legendre pour vérifier.

A ce jeu, j'avais déjà vu dans un colloque P. Dagnélie demander à un autre analyste des données ce que les benzécristes ajoutaient au chapitre du célèbre Kendall et Stuart [8]. Quand deux personnes (qui ont une formation minimale !) parlent d'analyse des correspondances, elles sont toujours d'accord sur l'algorithme, plus rarement sur ce que ça veut dire. Ceci pour dire que valeur propre et racine de valeur propre,*en comparaison*, c'est plus plaisant qu'il n'y paraît. En fait, la question tient sur un abus d'extension.

La première valeur propre est une variance projetée. La somme des variances projetées est l'inertie totale. Le rapport de la première à la somme a un sens.

La comparaison de la première avec la seconde a un sens. Et surtout la somme de la première et de la seconde a un sens. Les valeurs propres s'additionnent et décomposent l'inertie totale. C'est totalement sans objet pour les corrélations canoniques qui ne s'additionnent pas, ne font pas de pourcentages et ne décomposent rien.

Il faut alors clairement reconnaître, que sur des tableaux écologiques en présence-absence, en classes d'abondance, en recouvrement, en échelles logarithmique, ... la notion de pourcentage d'inertie projetée n'a pratiquement jamais de sens. L'inertie en AFC a un sens si on a affaire à une table de contingence.

3 Corrélation canonique

Utilisons pour expliquer ces éléments une fiche de `ade4` sur des données de J.M. Legay et D. Pontier [9]. Voir `table.cont`. Tout est dans la figure 1.

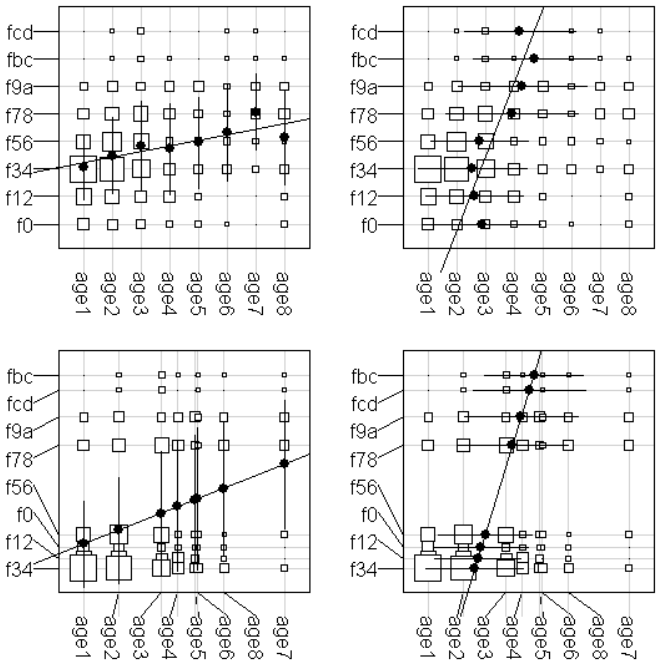


FIG. 1 – 354 chattes ont un âge (1 à 8 ans) et une fécondité annuelle (0 à 14 chatons). En haut, à gauche le graphe classique et la droite de régression fécondité fonction de l'âge. On s'arrête là en général. A droite, la droite de régression âge fonction de fécondité, un point de vue idiot d'un matheux. En dessous, à gauche le matheux aggrave son cas. Il jette les données et remplace les valeurs par des scores artificiels qui donne une double régression linéaire et une corrélation maximum. A côté l'autre droite de régression sur la même configuration. On voit qu'on conserve l'âge mais qu'on fait deux classes de fécondité. Le lien âge fécondité passe par le nombre de portées, pas par le nombre de chatons par portée. L'analyse montre la structure des données [2].

```
library(ade4)
data(chats)
chatsw <- data.frame(t(chats))
chatscoa <- dudi.coa(chatsw, scann = FALSE)
par(mfrow = c(2, 2))
table.cont(chatsw, abmean.x = TRUE, csi = 2, abline.x = TRUE, clabel.r = 1.5,
    clabel.c = 1.5)
table.cont(chatsw, abmean.y = TRUE, csi = 2, abline.y = TRUE, clabel.r = 1.5,
    clabel.c = 1.5)
table.cont(chatsw, x = chatscoa$c1[, 1], y = chatscoa$l1[, 1], abmean.x = TRUE,
    csi = 2, abline.x = TRUE, clabel.r = 1.5, clabel.c = 1.5)
table.cont(chatsw, x = chatscoa$c1[, 1], y = chatscoa$l1[, 1], abmean.y = TRUE,
    csi = 2, abline.y = TRUE, clabel.r = 1.5, clabel.c = 1.5)
par(mfrow = c(1, 1))
```

Nous sommes dans une situation parfaite pour expliquer ce qui se passe. Nous avons 354 échantillons d'un couple de mesures. l'édition des données :

```
chats
```

```
     f0 f12 f34 f56 f78 f9a fbc fcd
age1  8  15  44  11   7   4   0   0
age2  6  12  36  21  11   6   1   1
age3  4   7  18  13  12   4   2   2
age4  2   8   7   3   7   5   1   0
age5  2   3   5   3   4   6   0   0
age6  1   0   5   3   2   2   1   1
age7  0   0   3   2   5   4   1   1
age8  2   2   5   1   7   4   1   0
```

suffit. Il y a, par exemple, 12 chattes de 3 ans qui ont ou ont eu 7 ou 8 chatons. La table de contingence est une manière commode de remplacer un fichier à 365 lignes par un tableau permet de le voir. Refaisons le fichier d'origine :

```
xfac <- rep(factor(row.names(chats)[row(as.matrix(chats))]), as.numeric(as.matrix(chats)))
xnum <- rep((1:8)[row(as.matrix(chats))], as.numeric(as.matrix(chats)))
yfac <- rep(factor(names(chats)[col(as.matrix(chats))]), as.numeric(as.matrix(chats)))
ynum <- rep(c(0, seq(1.5, 13.5, by = 2))[col(as.matrix(chats))],
    as.numeric(as.matrix(chats)))
chats.df <- cbind.data.frame(xfac, xnum, yfac, ynum)
chats.df[(seq(1, 354, by = 20)), ]
```

```
    xfac xnum yfac ynum
1   age1    1   f0  0.0
21  age5    5   f0  0.0
41  age2    2  f12  1.5
61  age4    4  f12  1.5
81  age1    1  f34  3.5
101 age1    1  f34  3.5
121 age2    2  f34  3.5
141 age2    2  f34  3.5
161 age3    3  f34  3.5
181 age5    5  f34  3.5
201 age1    1  f56  5.5
221 age2    2  f56  5.5
241 age4    4  f56  5.5
261 age2    2  f78  7.5
281 age3    3  f78  7.5
301 age8    8  f78  7.5
321 age3    3  f9a  9.5
341 age8    8  f9a  9.5
```

La chatte `321` est dans la classe `age3`, elle a 3 ans, elle est dans la classe de fécondité `f9a`, elle a 9.5 chatons (le centre de la classe, par approximation, évidemment !). On se trouve alors avec quatre cas :

1. **x** et **y** sont des variables numériques ;
2. **x** est une variable qualitative, **y** est numérique ;
3. **x** est numérique et **y** est qualitative ;
4. **x** et **y** sont qualitatives (facteurs) ;

3.1 x et y sont des variables numériques

```
sunflowerplot(chats.df$xnum, chats.df$ynum)
abline(lm(chats.df$ynum ~ chats.df$xnum), lwd = 2)
cor.test(chats.df$xnum, chats.df$ynum)
```

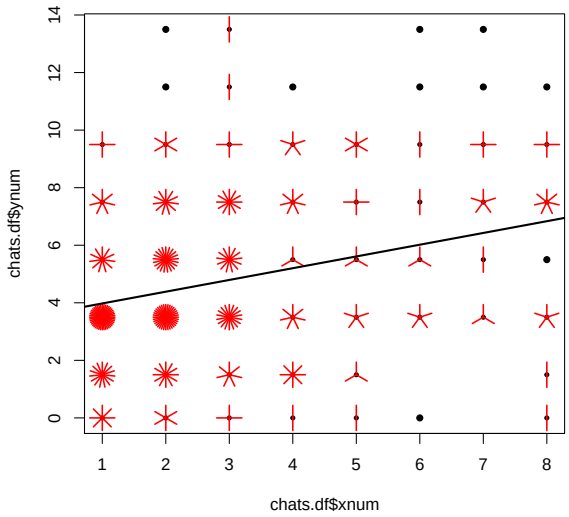
```
        Pearson's product-moment correlation
data:  chats.df$xnum and chats.df$ynum
t = 5.4983, df = 352, p-value = 7.38e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1823430 0.3744901
sample estimates:
      cor
0.2812329
```

```
anova(lm1 <- lm(chats.df$ynum ~ chats.df$xnum))
```

```
Analysis of Variance Table
Response: chats.df$ynum
              Df  Sum Sq Mean Sq F value   Pr(>F)
chats.df$xnum   1  252.32  252.32  30.231 7.38e-08 ***
Residuals     352 2937.90    8.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cor(chats.df$xnum, chats.df$ynum)^2
```

```
[1] 0.07909193
```



C'est le point de vue le plus classique. C'est significatif, c'est linéaire et on n'a rien vu. Les puristes voudront un test de linéarité.

3.2 x est une variable qualitative, y est numérique

On compare donc deux modèles. C'est le cas 2.

```
plot(jitter(chats.df$xnum), jitter(chats.df$ynum))
abline(lm(chats.df$ynum ~ chats.df$xnum), lwd = 2)
lines(1:8, tapply(chats.df$ynum, chats.df$xfac, mean), type = "b",
    cex = 2, pch = 20, lwd = 2, col = "red")
anova(lm2 <- lm(chats.df$ynum ~ chats.df$xfac))
```

```
Analysis of Variance Table
Response: chats.df$ynum
              Df  Sum Sq Mean Sq F value    Pr(>F)
chats.df$xfac   7  320.71   45.82  5.5243 4.858e-06 ***
Residuals     346 2869.51    8.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm1, lm2)
```

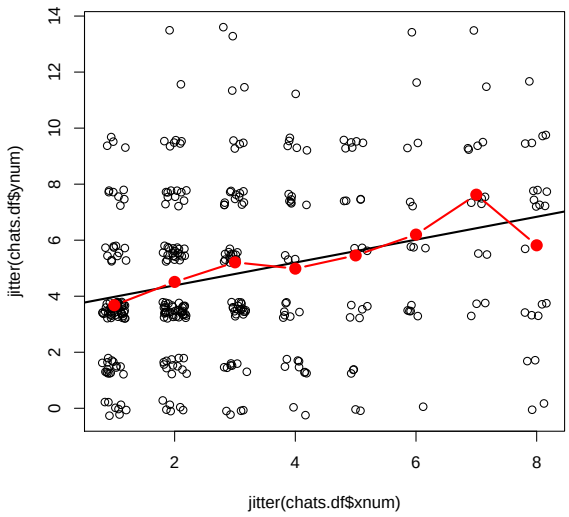
```
Analysis of Variance Table
Model 1: chats.df$ynum ~ chats.df$xnum
Model 2: chats.df$ynum ~ chats.df$xfac
  Res.Df     RSS  Df Sum of Sq      F Pr(>F)
1    352 2937.90
2    346 2869.51   6     68.39 1.3743 0.2241
```

```
summary(lm1)$r.squared
```

```
[1] 0.07909193
```

```
summary(lm2)$r.squared
```

```
[1] 0.1005282
```



On passe de 8 % de variance expliquée à 10 % mais le gain n'est pas significatif. Le test est non significatif et on dit souvent : le lien est linéaire. Voyons, on vous dit que c'est linéaire. Notez que le carré de corrélation est toujours inférieur au rapport de corrélation.

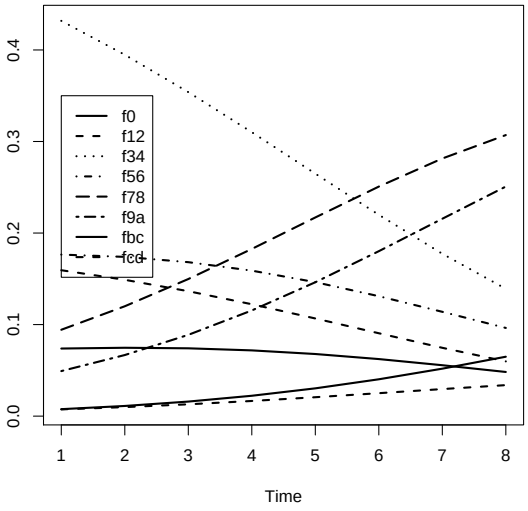
3.3 x est numérique et y est qualitative

C'est techniquement beaucoup plus difficile. A un âge donné, on a une distribution de fréquence observée des classes de fécondité. On peut supposer que la distribution de fréquence, par essence multinomiale, se déforme avec l'âge. C'est le domaine des modèles linéaires généralisés multinomiaux (*multicategory logit models*[1, Chapitre 8]).

```
library(nnet)
chamat <- as.matrix(chats)
x <- 1:8
m0 <- multinom(chamat ~ x)
```

```
# weights:  24 (14 variable)
initial  value 736.122306
iter  10 value 626.000729
iter  20 value 608.880512
final  value 608.880469
converged
```

```
ts.plot(m0$fitted.values, lwd = 2, lty = 1:8)
legend(1, 0.35, lty = 1:8, legend = dimnames(chamat)[[2]], lwd = 2)
```



3.4 x et y sont qualitatives

C'est ici que commence l'analyse des correspondances. Comment mesurer le lien entre deux variables qualitatives ? Par la racine de la première valeur propre, dirons certains. Ce point de vue est totalement légitime en écologie. Il a été illustré par R. Prodon et J.D. Lebreton [10].

```
data(rpjdl)
w <- data.frame(t(rpjdl$fau))
wcoa <- dudi.coa(w, scann = FALSE, nf = 4)
table.cont(w, abmean.y = TRUE, x = wcoa$c1[, 1], y = rank(wcoa$l1[,
    1]), csi = 0.2, clabel.c = 0, row.labels = rpjdl$lalab, clabel.r = 0.75)
```

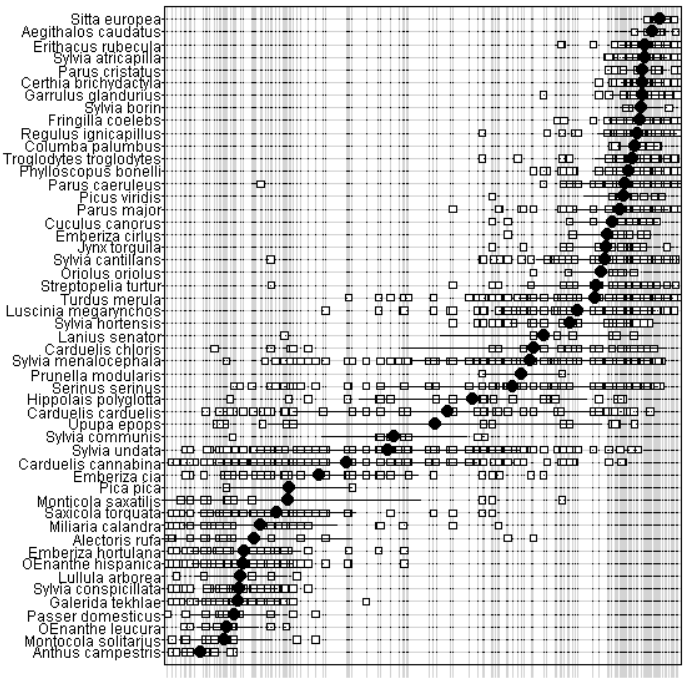
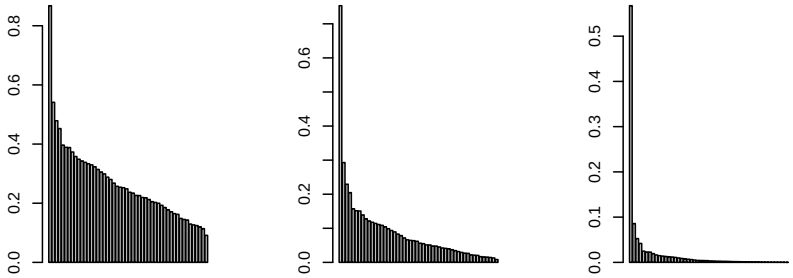


FIG. 2 – Ordination sur un gradient des espèces d'un cortège avifaunistique. En abscisse relevés positionnés par le premier score de variance unité. En ordonnée, espèce à la moyenne des sites occupés. On maximise ainsi la variance des positions des espèces [6]. Il y a deux ensembles distincts (milieux ouverts et milieux forestiers) et non un gradient.

La corrélation canonique mesure la possibilité de faire de telles ordinations. Rien de plus et rien de moins. Si on désire une nouvelle ordination, avec un nouveau score non corrélé au premier, on aura une nouvelle corrélation canonique forcément inférieure à la précédente. C'est tout ce qu'on peut dire. Le graphe des valeurs propres se lit alors comme le suivi de cette opération.

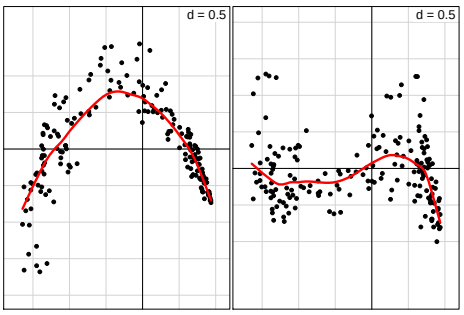
```
par(mfrow = c(1, 3))
barplot(sqrt(wcoa$eig))
barplot(wcoa$eig)
barplot(wcoa$eig^2)
```



La suite des racines des valeurs propres, celle des valeurs propres et celle des carrés des valeurs sont liées et pourtant si différentes !

Autre remarque d'importance : les écologues ont parlé de diagonalisation du tableau (et confondu avec diagonalisation d'une matrice). Il s'agit d'un outil et non d'un modèle. On ne peut rien voir dans un tableau. On peut le voir seulement à deux dimensions (lignes et colonnes), donc on ne peut regarder qu'une ordination simultanée des lignes et des colonnes. Le score est un outil exploratoire. Enfin, la possibilité de l'Arch effect, qui a fait couler beaucoup d'encre, indique que, si le premier score est indiscutable, les suivants peuvent éventuellement être des artifices numériques.

```
par(mfrow = c(1, 2))
par(mar = rep(0, 4))
s.label(wcoa$co, clab = 0, xax = 1, yax = 2)
lines(loess.smooth(wcoa$co[, 1], (wcoa$co[, 2]), span = 0.33), lwd = 3,
    col = "red")
s.label(wcoa$co, clab = 0, xax = 1, yax = 3)
lines(loess.smooth(wcoa$co[, 1], (wcoa$co[, 3]), span = 0.33), lwd = 3,
    col = "red")
```



Peut-être deux ou trois. Oui, mais on voit pointer les polynômes successifs de la première coordonnée dans les suivantes.

4 Statistiques d'inertie

Comment mesurer le lien entre deux variables qualitatives ? Par le χ^2 de la table de contingence dirons d'autres. On se réfère ici à la métrique du χ^2, celle

de la somme chère à tout biologiste de première année :

$$\sum_i \frac{(obs_i - cal_i)^2}{cal_i}$$

```
chisq.test(chatsw)$statistic
```

```
X-squared
 74.53507
```

```
sum(chatscoa$eig) * sum(chamat)
```

```
[1] 74.53507
```

Autant le dire clairement : pour une vraie table de contingence, ce calcul est légitime. Il fournit le test d'indépendance. Pour un tableau en présence-absence, il est plus ou moins discutable, et la décomposition en valeurs propres donnant des taux d'inertie projetée l'est autant.

4.1 Reconstitution des données

On peut aussi voir la suite des valeurs propres dans l'automodélisation du tableau. Ajoutons la fonction :

```
"reconst.coa" <- function(dudi, nf = 1, ...) {
    if (!inherits(dudi, "dudi"))
        stop("Object of class 'dudi' expected")
    if (nf > dudi$nf)
        stop(paste(nf, "factors need >", dudi$nf, "factors available\n"))
    if (!inherits(dudi, "coa"))
        stop("Object of class 'dudi' expected")
    pl <- dudi$lw
    pc <- dudi$cw
    n <- dudi$N
    res0 <- outer(pl, pc) * n
    res <- data.frame(res0)
    names(res) <- names(dudi$tab)
    row.names(res) <- row.names(dudi$tab)
    if (nf == 0)
        return(res)
    for (i in 1:nf) {
        xli <- dudi$li[, i]
        yc1 <- dudi$c1[, i]
        res <- res + outer(xli, yc1) * res0
    }
    return(res)
}
```

Elle applique ce qu'on appelle la formule de reconstitution des données :

$$n_{ij} = \frac{n_{i.} n_{.j}}{n_{..}} + \sum_{k=1}^{k=r} \frac{L_{ik} C_{jk}}{\sqrt{\lambda_k}}$$

k est le nombre de facteurs utilisés. Il peut être nul : c'est l'hypothèse d'indépendance.

```
mod0 <- reconst(chatscoa, 0)
round(mod0, 1)
```

```
    age1 age2 age3 age4 age5 age6 age7 age8
f0   6.3  6.6  4.4  2.3  1.6  1.1  1.1  1.6
f12 11.8 12.5  8.2  4.4  3.1  2.0  2.1  2.9
f34 30.9 32.7 21.5 11.5  8.0  5.2  5.6  7.6
f56 14.3 15.1 10.0  5.3  3.7  2.4  2.6  3.5
f78 13.8 14.6  9.6  5.1  3.6  2.3  2.5  3.4
f9a  8.8  9.3  6.1  3.3  2.3  1.5  1.6  2.2
fbc  1.8  1.9  1.2  0.7  0.5  0.3  0.3  0.4
fcd  1.3  1.3  0.9  0.5  0.3  0.2  0.2  0.3
```

```
sum((chatsw - mod0)^2/mod0)
```

```
[1] 74.53507
```

```
n <- sum(chatsw)
n * sum(chatscoa$eig)
```

```
[1] 74.53507
```

L'erreur totale est le χ^2 soit n fois la somme des valeurs propres (n est le nombre total d'individus, ici 354). Si on utilise un facteur :

```
mod <- reconst(chatscoa, 1)
round(mod, 1)
```

```
    age1 age2 age3 age4 age5 age6 age7 age8
f0   7.5  7.2  4.2  2.1  1.4  0.9  0.7  1.2
f12 14.9 13.9  7.7  3.8  2.4  1.5  0.9  2.0
f34 40.9 37.2 19.7  9.5  5.9  3.8  1.5  4.5
f56 15.8 15.8  9.7  5.0  3.4  2.2  2.0  3.1
f78  7.0 11.5 10.9  6.5  5.0  3.3  5.3  5.6
f9a  2.8  6.6  7.2  4.4  3.5  2.4  4.0  4.1
fbc  0.1  1.1  1.5  1.0  0.8  0.5  1.0  1.0
fcd  0.2  0.8  1.1  0.7  0.6  0.4  0.7  0.7
```

```
sum((chatsw - mod)^2/mod0)
```

```
[1] 31.68349
```

```
n * sum(chatscoa$eig[-1])
```

```
[1] 31.68349
```

Si on utilise deux facteurs :

```
mod <- reconst(chatscoa, 2)
round(mod, 2)
```

```
     age1  age2  age3  age4  age5 age6 age7 age8
f0   7.75  6.76  3.66  2.57  1.70 0.60 0.52 1.43
f12 16.22 12.12  5.46  5.84  3.80 0.26 0.29 3.01
f34 40.38 37.92 20.61  8.65  5.33 4.30 1.75 4.05
f56 14.24 17.81 12.26  2.67  1.80 3.69 2.67 1.87
f78  7.30 11.04 10.34  7.00  5.36 2.99 5.12 5.85
f9a  3.67  5.37  5.75  5.84  4.47 1.48 3.63 4.79
fbc -0.19  1.43  1.96  0.59  0.54 0.79 1.12 0.77
fcd -0.36  1.54  1.97 -0.15 -0.01 0.89 0.90 0.23
```

```
sum((chatsw - mod)^2/mod0)
```

```
[1] 13.06132
```

```
n * sum(chatscoa$eig[-(1:2)])
```

```
[1] 13.06132
```

On voit donc bien que la suite des valeurs propres représente la capacité à automodéliser le tableau de données. Dire que 57% de l'inertie est projetée sur le premier axe, c'est également dire que l'erreur de reconstitution diminue de 57% en utilisant le premier score. Un score peut donc être bon (grande valeur propre ou grande racine de la valeur propre) et interprétable (il range les objets dans un ordre qui a du sens) mais être relativement peu modélisateur. Et ceci est un résultat écologique très important. On peut avoir un gradient de synthèse pertinent et un taux de modélisation de 3%. Dire qu'une analyse est bonne ou pas bonne parce qu'elle donne des résultats recherchés ou non est une erreur fondamentale.

```
round(wcoa$eig[1:10]/sum(wcoa$eig), 2)
```

```
 [1] 0.17 0.07 0.05 0.05 0.04 0.03 0.03 0.03 0.03 0.03
```

```
round(chatscoa$eig[1:7]/sum(chatscoa$eig), 2)
```

```
[1] 0.57 0.25 0.07 0.06 0.03 0.01 0.00
```

4.2 Reconstitution de tableaux écologiques

Un cortège faunistique qui présente 90% d'erreur non modélisable (c'est-à-dire modélisable par petit morceau les uns derrière les autres, modélisable comme un ensemble d'aléas indépendants) ce n'est ni bon ni mauvais. C'est un fait. Et si c'était 100% ce serait un fait extraordinaire !

```
par(mfrow = c(1, 3))
data(mafragh)
eig1 <- dudi.coa(mafragh$flo, scannf = F)$eig
round(eig1[1:7]/sum(eig1), 2)
```

```
[1] 0.10 0.08 0.07 0.06 0.05 0.05 0.04
```

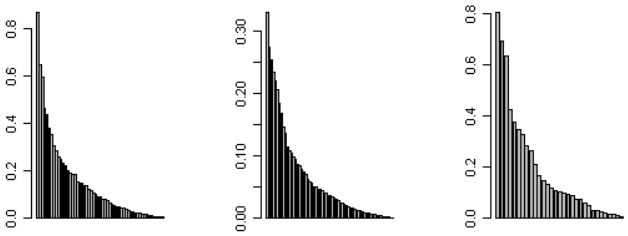
```
barplot(eig1)
data(avijons)
eig2 <- dudi.coa(avijons$fau, scannf = F)$eig
round(eig2[1:7]/sum(eig1), 2)
```

```
[1] 0.04 0.03 0.03 0.03 0.03 0.02 0.02
```

```
barplot(eig2)
data(mollusc)
eig3 <- dudi.coa(mollusc$fau, scannf = F)$eig
round(eig3[1:7]/sum(eig1), 2)
```

```
[1] 0.10 0.08 0.08 0.05 0.05 0.04 0.04
```

```
barplot(eig3)
```



5 Conclusion

Il est donc clair que la corrélation canonique ou son carré, donc la valeur propre a un sens en elle-même. Ceci est typique de l'analyse des correspondances qui a des valeurs propres comprises entre 0 et 1, qui sont donc des pourcentages. Ce pourcentage est un pourcentage de variance inter-lignes pour une variance inter-colonnes donnée (par averaging) ou un pourcentage de variance inter-colonnes pour une variance inter-lignes donnée (par averaging). Les scores maximisent cette quantité. Un tableau écologique qui ne contiendrait aucune valeur franchement plus grande que les autres serait une boule d'aléas.

La suite de ces valeurs rapportées à leur somme représente la possibilité de modéliser le tableau. La plupart du temps, l'ensemble des éléments conjoncturels ou historiques qui emplissent un tableau écologique, génèrent une impossibilité qui se traduit par des taux faibles. L'intérêt d'une valeur pour le premier point de vue et l'intérêt d'une valeur pour le second point de vue ne se comprennent que par leur définition et ne se comparent pas.

Cette valeur propre a un sens complètement différent quand on la compare aux autres. Elle mesure la possibilité de modéliser le tableau par la somme d'un petit nombre de matrices de rang 1, c'est-à-dire du type :

$$x_{ij} = a_i b_j$$

Les deux approches ne peuvent être contradictoires. Elles disent chacune quelque chose sur le tableau de données.

On a *reproché* à Karine Jacquet *pour une publication*, le fait d'avoir un faible pourcentage de variance expliqué sur un premier axe, alors qu'on y voyait une ordination correcte. Il est particulièrement absurde de reprocher un résultat descriptif. On ne peut contester que l'interprétation qui en est faite. Un tableau écologique peut être ordonné ET recouvert d'aléas, parfois en proportion énorme. La question est d'importance. Doit-on cacher cette propriété très particulière des données écologiques ?

Références

[1] A. Agresti. *An introduction to categorical data analysis*. John Wiley, New York, 1996.

[2] J.P. Benzécri. Statistical analysis as a tool to make patterns emerge from data. In S. Watanabe, editor, *Methodologies of pattern recognition*, pages 35–60. Academic Press, New York, 1969.

[3] J.P. Benzécri and Coll. *L'analyse des données. II L'analyse des correspondances*. Bordas, Paris, 1973.

[4] B. Escofier-Cordier. L'analyse factorielle des correspondances. *Cahiers du Bureau Universitaire de Recherche Opérationelle, Université de Paris*, 13 :25–59, 1969.

[5] W.J. Heiser. Joint ordination of species ans sites : the unfolding technique. In L. Legendre and P. Legendre, editors, *Developments in numerical ecology*, pages 189–221. Springer-Verlag, Berlin, Ecological Sciences, Vol. 14, 1987.

[6] M.O. Hill. Use of simple discriminant functions to classify quantitative phytosociological data. In E. Diday, editor, *Proceedings of the First International Symposium on Data Analysis and Informatics*, pages 181–199. INRIA Rocquencourt, France, 1977.

[7] H.O. Hirschfeld. A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences*, 31 :520–524, 1935.

[8] D.G. Kendall and A. Stuart. *The advanced theory of statistics. Vol 2 : Inference and relationships. Cha. 33 : Categorized data*. Griffin, London, 1961.

[9] J.M. Legay and D. Pontier. Relation âge-fécondité dans les populations de chats domestiques, felis catus. *Mammalia*, 49 :395–402, 1985.

[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] E.J. Williams. Use of scores for the analysis of association in contingency tables. *Biometrika*, 39 :274–289, 1952.