Fiche TD avec le logiciel R : tdr54

Extensions de l'analyse discriminante linéaire

D. Chessel

Table des matières

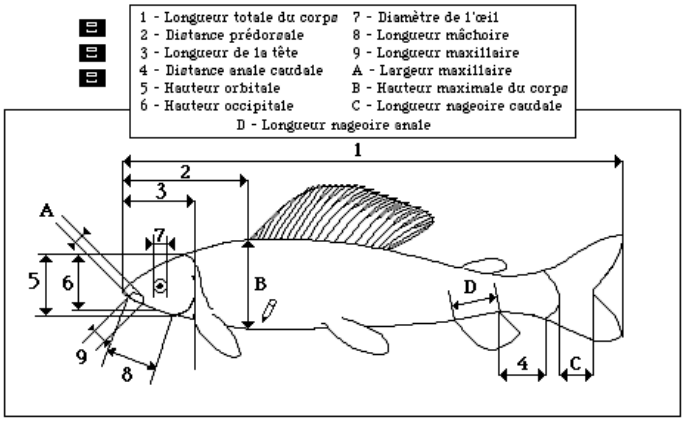
1 Introduction

Les exemples étudiés dans la fiche :

`http://pbil.univ-lyon1.fr/R/pdf/tdr63.pdf`

appartiennent à la morphométrie, discipline qui a généré l'analyse discriminante. Il s'agissait d'abord, en anthropométrie, de discriminer des races, des castes, des groupes, des populations, à une époque où cette activité n'était pas contestée. Les conséquences méthodologiques des travaux de Mahalanobis [1] sont considérables. Ils sont d'abord basés sur la loi normale. L'essentiel est dans l'homogénéité des matrices de variances covariances par sous-populations. Dans la MANOVA un groupe fournit un échantillon d'une loi normale de matrice de covariances fixée dont l'espérance est une fonction des variables de contrôle. On discrimine d'abord par les moyennes des échantillons d'une loi normale.

Considérons encore un exemple en morphométrie. Le tableau de données est rapporté dans [24] et étudié *pro parte* dans [28]. Il s'agit de 13 mensurations sur des ombres communs (*Thymallus thymallus*) décrites dans une carte de documentation [8] sous la forme :



Les individus sont rangés par lieu de capture :

pop1 [1-41] Ain amont (41 individus)

pop2 [42-59] Bienne (18 individus)

pop3 [60-79] Loue (20 individus)

pop4 [80-102] Ain aval (23 individus)

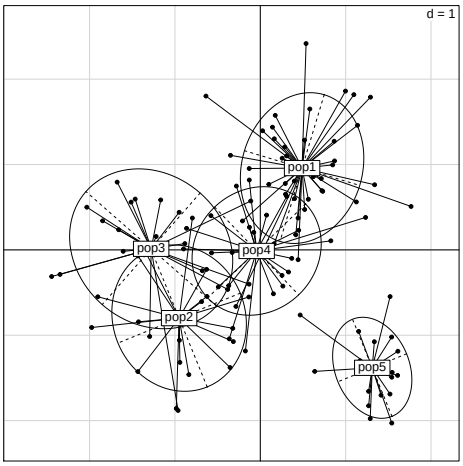
pop5 [103-120] Loire (18 individus)

```
options(digit=4)
library(ade4)
omb <- read.table(url("http://pbil.univ-lyon1.fr/R/donnees/thymallus.txt"))
pop <- factor(rep(paste("pop",1:5,sep=""),c(41,18,20,23,18)))
omb.pca <- dudi.pca(omb,scannf=F)
omb.bet <- bca(omb.pca,pop,scannf=F)
omb.dis <- discrimin(omb.pca,pop,scannf=F)
```

1. Photo : http ://www.isibang.ac.in/ISIBC/mahal.htm

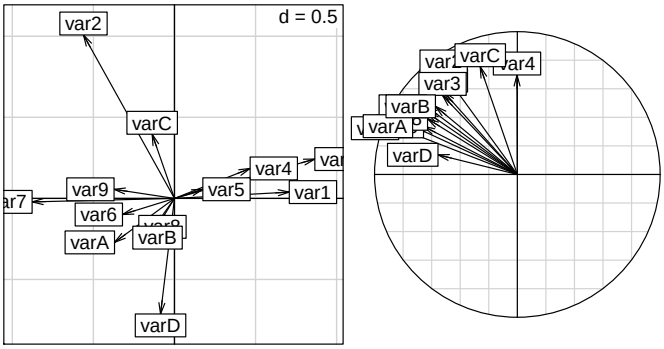
Dans la figure de référence (variables canoniques, groupes, moyennes et variances-covariances) :

```
s.class(omb.dis$li,pop)
```



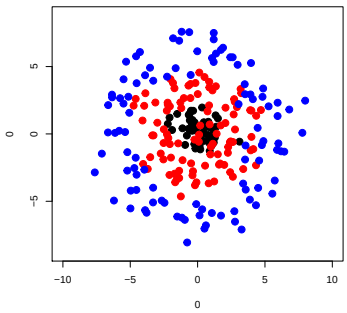
on voit d'abord un niveau de variabilité commun pour au moins quatre des populations. La cinquième est très différente. Mais l'analyse est marquée par un gros défaut :

```
par(mfrow=c(1,2))
s.arrow(omb.dis$fa)
s.corcircle(omb.dis$va)
```



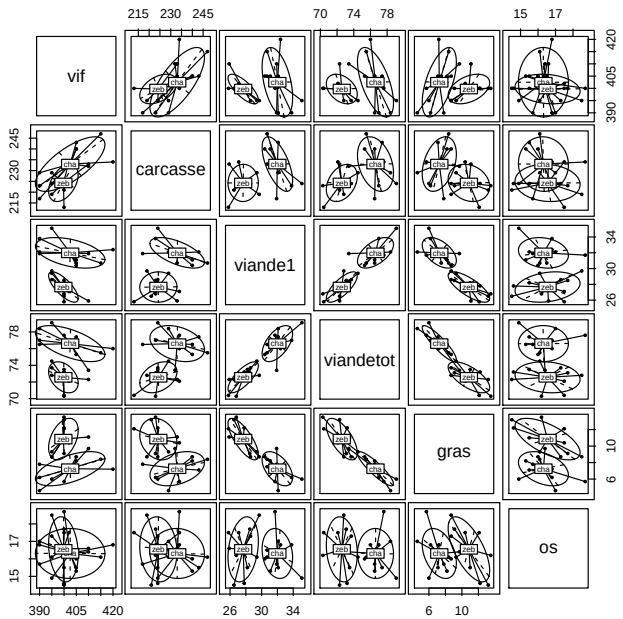
L'ACP du tableau est complètement écrasée par l'effet taille qui se reporte intégralement dans l'ACP inter-classe et perturbe gravement l'analyse discriminante. Observer, en effet, que les poids canoniques (coefficients des combinaisons linéaires discriminantes) sont incohérents avec les corrélations variables-score qui restent de signes constants. L'effet taille est encore visible et prend sa part dans la discrimination de la population 5. Si on désire s'affranchir totalement de l'effet taille (ce qui est légitime chez des organismes dont la taille croît de façon continue) il convient d'étendre la définition de l'analyse discriminante aux cas des matrices de covariances non inversibles.

Sur un autre plan, l'analyse discriminante sépare les groupes sur leur position moyenne. Il est bien évident que nombre de problèmes ne concerne pas les moyennes, du genre :



Mais avant d'en arriver à ce type de questions, il faut s'intéresser à la discrimination par position sur des variables qui ne sont jamais normales comme les facteurs, les données binaires ou distributionnelles, les mélanges. On peut garder la notion de variable canonique (combinaisons des colonnes de variance unité), celle de variance inter-groupe (la variable canonique doit optimiser cette quantité). Il est bon de voir dans l'analyse discriminante linéaire (**ADL**) l'analyse d'inertie d'un nuage de points et de définir une classe de métriques qu'on appelle par extensions les métriques de Mahalanobis, bien qu'il s'agisse de généralisation qui s'écarte complètement des conceptions du fondateur.

2 n individus p variables g groupes



La question des valeurs relatives du nombre d'individus, de variables et de classes permet d'introduire aux schémas théoriques de ce type d'analyse. Considérons

d'abord un exemple où il n'y a que deux classes [30, Tableau 8 p. 45]. Les variables sont : poids vif, poids de la carcasse, poids de la viande de première qualité, poids de la viande totale, poids du gras et poids des os en kg pour 23 bovins, 12 Charolais et 11 Zébus.

```
data(chazeb)
pairs(chazeb$tab,panel=function(x,y,...) s.class(cbind.data.frame(x,y),chazeb$cla,add.p=T))
```

Il s'agit d'une illustration pédagogique.

```
summary(aov(as.matrix(chazeb$tab)~chazeb$cla))
plot(discrimin(dudi.pca(chazeb$tab,scan=F),chazeb$cla,scan=F))
```

Qu'obtient-on avec la commande `aov` ?

```
summary(manova(as.matrix(chazeb$tab)~chazeb$cla))
```

```
           Df  Pillai approx F num Df den Df   Pr(>F)
chazeb$cla  1 0.84508   14.547      6     16 1.11e-05 ***
Residuals  21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

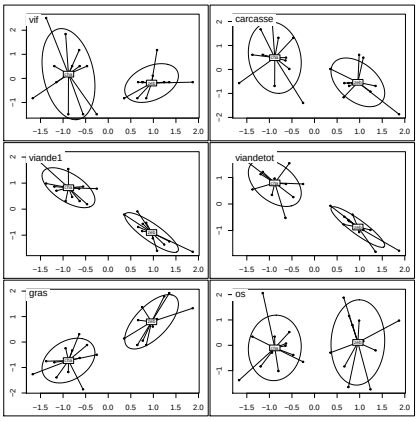
```
summary(manova(as.matrix(chazeb$tab)~chazeb$cla),"Wilks")
```

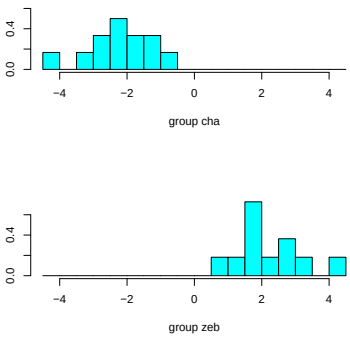
```
           Df   Wilks approx F num Df den Df   Pr(>F)
chazeb$cla  1 0.15492   14.547      6     16 1.11e-05 ***
Residuals  21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(manova(as.matrix(chazeb$tab)~chazeb$cla),"Hotelling-Lawley")
```

```
           Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
chazeb$cla  1           5.4551   14.547      6     16 1.11e-05 ***
Residuals  21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

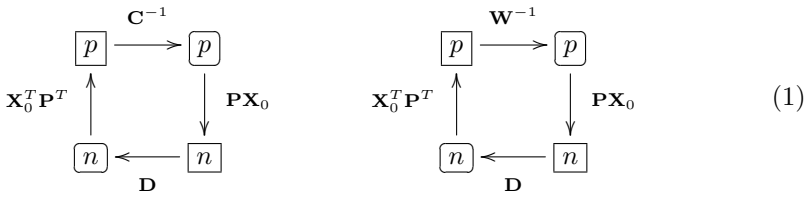
```
par(mar=c(5.1,4.1,4.1,2.1))
plot(lda(chazeb$tab,chazeb$cla))
```





Pourquoi obtient-on un résultat unique pour plusieurs tests ? Pourquoi le plot des analyses discriminantes a-t-il été modifié ? Tout simplement, la présence de deux groupes impose celle d'une variable canonique unique et d'une seule valeur propre non nulle. Il y a donc besoin de préciser ce qu'est l'analyse discriminante pour voir dans un même schéma le cas de nombreuses populations pour un faible nombre de variables, comme celui de peu de populations connues par un grand nombre de variables.

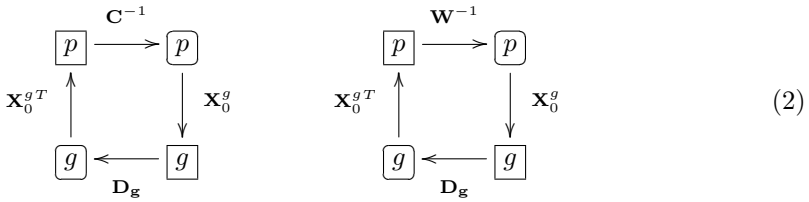
Rappelons que les schéma déjà rencontrés sont :



(1)

Nous garderons celui de gauche après avoir montré la relation étroite qu'il entretient celui de droite (voir `tdr63`).

On note $\mathbf{X}_0^g$ le tableau à g lignes et p colonnes constitué des moyennes par variables et par groupes à partir du tableau $\mathbf{X}_0$. En toute généralité, les schémas (1) se transforment par :



(2)

Dans (1) le tableau a n lignes qui sont identiques pour chaque groupe et p colonnes. L'inertie du nuage de points est identique dans chaque direction (dont les axes principaux) quand on remplace un paquet de points superposés par un seul auquel on affecte la somme des poids de ces points superposés. Dans (2) le tableau a g lignes et p colonnes. Plus formellement cela revient à l'écriture matricielle :

$$\mathbf{P}^T\mathbf{D}\mathbf{P} = \mathbf{D}\mathbf{P} = \mathbf{D}\mathbf{H}\mathbf{D}_g^{-1}\mathbf{H}^T\mathbf{D} = \mathbf{K}^T\mathbf{D}_g\mathbf{K}$$

avec :

$$\mathbf{K} = \mathbf{D}_g^{-1}\mathbf{H}^T\mathbf{D}$$

opérateur qui transforme un tableau en tableau des centres de gravité par classes définies par les indicatrices colonnes de $\mathbf{H}$.

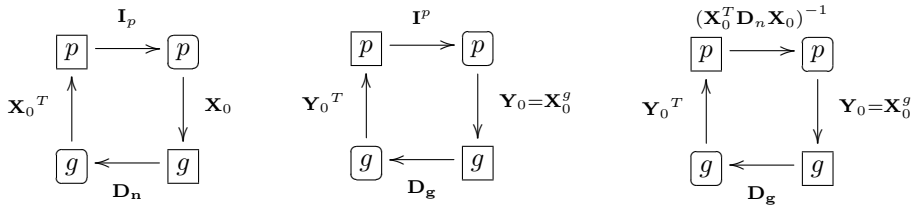
Si $g = 2$ le tableau $\mathbf{X}_0^g$ à deux lignes. Il est centré par colonne et de rang 1. Il n'y a qu'une seule valeur propre non nulle.

En général, le vecteur $\mathbf{1}_g$ est axe principal car :

$$\mathbf{X}_0^{gT}\mathbf{D}_\mathbf{g}\mathbf{1}_g = \mathbf{0}_p$$

Il y a au plus $g - 1$ valeurs propres non nulles. Les schémas de gauche dans (1) et (2) ont les mêmes axes et facteurs principaux et mêmes valeurs propres. Ils appartiennent à la même analyse dans deux interprétations différentes. C'est là une des difficulté de fond de la compréhension de ces techniques. Un même calcul, qui permet d'identifier une *méthode*, a plusieurs interprétations donc des *usages* multiples.

A l'inverse, des méthodes différentes ont des objectifs différents. Après une ACP, nous devons retenir les trois schémas :



Le schéma de gauche est celui de l'**ACP**, le second du centre est celui de l'**ACP inter-classe**, le troisième est celui de l'**analyse discriminante**. Elle utilise la métrique de Mahalanobis au sens large ([6, p. 487] , [18, p.258]). A partir de là des extensions sont possibles.

3 Extension au double centrage

On voudrait se débarrasser définitivement de l'effet taille dans la comparaison entre groupes. Une manière de faire est d'utiliser un double centrage sur les logarithmes des mesures (deux individus reliés par une constante multiplicatives deviennent identiques) :

```
omblog <- bicenter.wt(log(omb))
```

L'analyse discriminante :

```
lda(omblog,pop)
```

enverra un message d'alerte sans bloquer les calculs :

```
Warning message:
les variables sont collinéaires in: lda.default(x, grouping, ...)
```

tandisque la MANOVA :

```
summary(manova(omblog~pop))
```

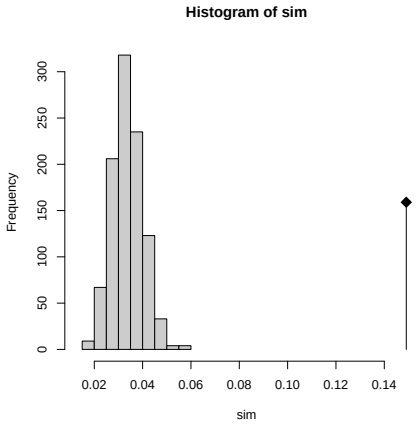
renvoie un message d'erreur :

```
Erreur dans summary.manova(manova(omblog ~ pop)) :
        residuals have rank 12 < 13
```

indiquant clairement que le tableau n'est plus de rang plein.

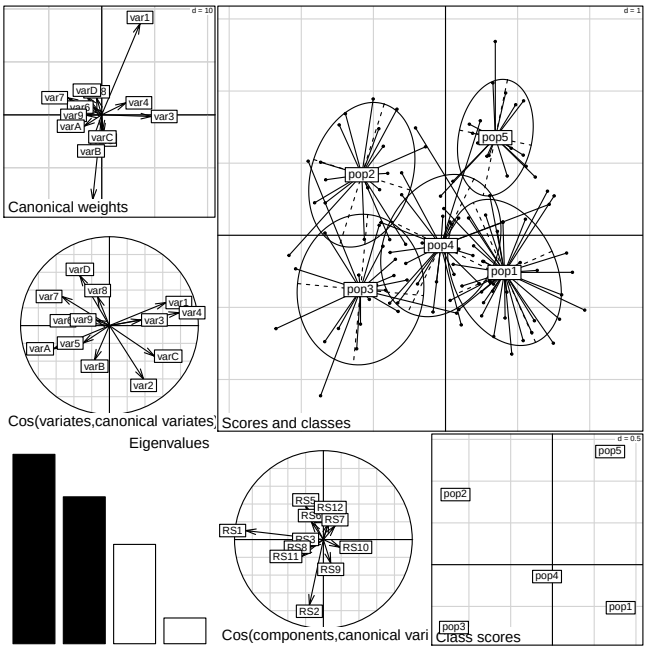
On peut cependant refaire une ACP (non centrée et non normalisée) puis une analyse discriminante et un test de signification :

```
ombl.pca <- dudi.pca(omblog,center=F,scannf=F,scal=F)
ombl.dis <- discrimin(ombl.pca,pop,scannf=F)
plot(randtest(ombl.dis))
```



On notera que les corrélations entre composantes principales (scores de variances maximales) et scores canoniques (scores discriminants) de même rang sont élevées pour les deux premières, que poids et corrélations des variables sont à nouveaux cohérents. La discrimination est légitime. Elle se fait sans référence avec la taille des individus, ce qui est l'objectif visé.

```
plot(ombl.dis)
```

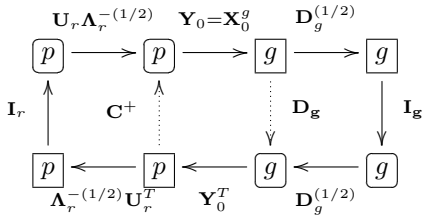


Ceci est possible car fonction `discrimin` utilise en fait des inverses généralisés (bases dans [2]), introduites en statistiques par C.R. Rao [25], au centre des questions d'élimination de la taille [5] [31]), sous la forme :

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^T \Rightarrow \mathbf{C}^+ = \mathbf{U}_r\mathbf{\Lambda}_r^{-1}\mathbf{U}_r^T$$

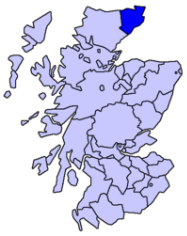
r est le rang et $\mathbf{U}_r$ la matrice des vecteurs propres réduites aux r premiers, les r valeurs propres non nulles étant dans $\mathbf{\Lambda}_r$. On a ainsi l'inverse généralisé de $\mathbf{C}$ de rang minimal [12].

Le calcul se fait alors sur le schéma :



qui permet de diagonaliser dans la plus petite des deux dimensions p (le nombre de variables) ou g (le nombre de groupes) une matrice symétrique. Ce qui fait que l'analyse discriminante est construite sur l'ACP du nuages des moyennes par classes des coordonnées normalisées de l'ACP sous-jacente. A partir de cette observation on peut généraliser à toute analyse de niveau 1. Observons d'abord que l'AFC est une analyse discriminante.

4 AFC et discrimination sur une variable qualitative



Les gens de Caithness [2] sont célèbres (chez les statisticiens) pour la couleur de leurs yeux et celle de leurs cheveux. C'est Sir R.A. Fisher qui a inventé, entre autre, l'analyse des correspondances. [11].

```
data(caith)
caith
```

```
       fair red medium dark black
blue    326  38    241  110     3
light   688 116    584  188     4
medium  343  84    909  412    26
dark     98  48    403  681    85
```

Il y a 5387 personnes dont on connaît la couleur des yeux (en lignes) et des cheveux (en colonnes). On peut faire un tableau avec autant de lignes que de personnes :

```
fac.col <- factor(rep(col(as.matrix(caith)),as.matrix(caith)))
levels(fac.col) <- names(caith)
fac.lig <- factor(rep(row(as.matrix(caith)),as.matrix(caith)))
levels(fac.lig) <- row.names(caith)
```

Vérifier l'identité des deux types d'information :

```
table(fac.lig,fac.col)
```

```
        fac.col
fac.lig  fair red medium dark black
  blue    326  38    241  110     3
  light   688 116    584  188     4
  medium  343  84    909  412    26
  dark     98  48    403  681    85
```

2. Carte : http ://fr.wikipedia.org/wiki/Caithness

Faire l'analyse des correspondances :

```
coa1 <- dudi.coa(caith,scan=F)
w1 <- model.matrix(~-1+fac.col)
w1[sample(1000,5),]
```

```
    fac.colfair fac.colred fac.colmedium fac.coldark fac.colblack
997           1          0             0           0            0
705           1          0             0           0            0
473           1          0             0           0            0
722           1          0             0           0            0
340           1          0             0           0            0
```

Caractériser l'objet w1.

```
lda5 <- lda(w1,fac.lig)
lda5
```

```
Call:
lda(w1, grouping = fac.lig)
Prior probabilities of groups:
     blue     light    medium      dark
0.1332838 0.2932987 0.3293113 0.2441062

Group means:
       fac.colfair fac.colred fac.colmedium fac.coldark fac.colblack
blue    0.45403900 0.05292479     0.3356546   0.1532033  0.004178273
light   0.43544304 0.07341772     0.3696203   0.1189873  0.002531646
medium  0.19334837 0.04735062     0.5124014   0.2322435  0.014656144
dark    0.07452471 0.03650190     0.3064639   0.5178707  0.064638783

Coefficients of linear discriminants:
                    LD1        LD2        LD3
fac.colfair    1.3838266 -0.9484982 -0.5464516
fac.colred     0.6061757 -0.2137223  3.9077918
fac.colmedium  0.1275821  1.2877337 -0.2297008
fac.coldark   -1.4509037 -0.5394969 -0.4643861
fac.colblack  -2.7164292 -1.6073577  1.4544913

Proportion of trace:
   LD1    LD2    LD3
0.8864 0.1105 0.0031
```

Commenter ce qui sort de ces commandes.

```
coa1$li
```

```
             Axis1       Axis2
blue   -0.40029985  0.16541100
light  -0.44070764  0.08846303
medium  0.03361434 -0.24500190
dark    0.70273880  0.13391383
```

```
coa1$co
```

```
             Comp1       Comp2
fair   -0.54399533  0.17384449
red    -0.23326097  0.04827895
medium -0.04202412 -0.20830421
dark    0.58870853  0.10395044
black   1.09438828  0.28643670
```

```
coa1$eig
```

```
[1] 0.1992447520 0.0300867741 0.0008594814
```

```
discri1 <- discrimin(dudi.pca(data.frame(w1),scal=F,scan=F),fac.lig,scan=F)
discri1$eig
```

```
[1] 0.1992447520 0.0300867741 0.0008594814
```

```
unique(discri1$li[,1])
```

```
[1]  1.21871379  0.52257500  0.09414671 -1.31888486 -2.45176017
```

```
unique(discri1$li[,2])
```

```
[1]  1.0022432  0.2783364 -1.2009094  0.5992920  1.6513565
```

```
coa1$c1
```

```
               CS1        CS2
fair   -1.21871379  1.0022432
red    -0.52257500  0.2783364
medium -0.09414671 -1.2009094
dark    1.31888486  0.5992920
black   2.45176017  1.6513565
```

Expliquer la relation entre les objets coa1$c1 et discri1$li.

```
tapply(discri1$li[,1],fac.lig,mean)
      blue      light     medium       dark
0.40029985 0.44070764 -0.03361434 -0.70273880
tapply(discri1$li[,2],fac.lig,mean)
      blue      light     medium       dark
0.16541100 0.08846303 -0.24500190 0.13391383
coa1$li
            Axis1       Axis2
blue   -0.40029985  0.16541100
light  -0.44070764  0.08846303
medium  0.03361434 -0.24500190
dark    0.70273880  0.13391383
```

L'AFC est bien une analyse discriminante. C'est surtout une *double analyse discriminante*. Faire l'autre.

Du point de vue mathématique, c'est plus difficile. Considérons une table de contingence $\mathbf{T}$ avec I lignes et J colonnes. Elle s'écrit

$$\mathbf{T} = \mathbf{X}^T\mathbf{Y}$$

Le tableau $\mathbf{X}$ est celui des indicatrices des lignes et $\mathbf{Y}$ est celui des indicatrices des colonnes. Soit n le nombre total d'individus. $\mathbf{X}$ a n lignes et I colonnes. $\mathbf{Y}$ a n lignes et J colonnes. Faisons l'analyse discriminante de $\mathbf{X}$ par les classes définies par $\mathbf{Y}$.

Il faut calculer les moyennes. Les individus portent un poids uniforme et les moyennes sont dans le vecteur colonne $\mathbf{X}^T\mathbf{D}_{1/n}\mathbf{1}_n$, c'est-à-dire le vecteur des fréquences marginales des lignes de $\mathbf{T}$. Si, comme il est habituel, on note $\mathbf{D}_I$ la diagonale des poids des lignes, ces moyennes forment le vecteur $\mathbf{f}_I = \mathbf{D}_I\mathbf{1}_I$.

Le tableau centré est donc $\mathbf{X} - \mathbf{1}_n\mathbf{f}_I^T$.

La matrice de variances-covariances est donc :

$$\begin{aligned}
\mathbf{C} &= (\mathbf{X}^T - \mathbf{f}_I\mathbf{1}_n^T)\mathbf{D}_{1/n}(\mathbf{X} - \mathbf{1}_n\mathbf{f}_I^T) \\
&= \mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T - \mathbf{f}_I\mathbf{f}_I^T + \mathbf{f}_I\mathbf{f}_I^T \\
&= \mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T
\end{aligned}$$

laquelle n'est évidemment pas inversible comme l'a indiqué la fonction `lda` (qui utilise des décompositions en valeurs singulières qui élimine le problème).

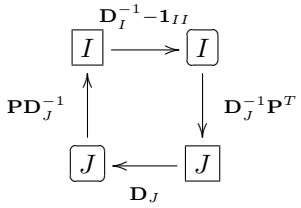
Mais elle a un inverse généralisé $\mathbf{C}^\sharp$ qui s'écrit ($\mathbf{1}_{II}$ est la matrice à I lignes et I colonnes dont toutes les valeurs égalent l'unité) :

$$\mathbf{C}^\sharp = \mathbf{D}_I^{-1} - \mathbf{1}_{II}$$

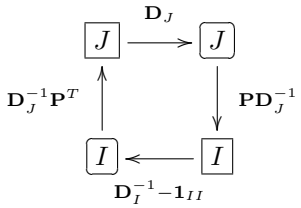
Il suffit de vérifier la définition :

$$\begin{aligned}
\mathbf{C}\mathbf{C}^\sharp\mathbf{C} &= (\mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T)(\mathbf{D}_I^{-1} - \mathbf{1}_{II})(\mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T) \\
&= (\mathbf{I}_I - \mathbf{f}_I\mathbf{1}_I^T - \mathbf{f}_I\mathbf{1}_I^T + \mathbf{f}_I\mathbf{1}_I^T)(\mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T) \\
&= (\mathbf{I}_I - \mathbf{f}_I\mathbf{1}_I^T)(\mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T) \\
&= \mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T = \mathbf{C} \\
\mathbf{C}^\sharp\mathbf{C}\mathbf{C}^\sharp &= (\mathbf{D}_I^{-1} - \mathbf{1}_{II})(\mathbf{D}_I - \mathbf{f}_I\mathbf{f}_I^T)(\mathbf{D}_I^{-1} - \mathbf{1}_{II}) \\
&= (\mathbf{D}_I^{-1} - \mathbf{1}_{II}) = \mathbf{C}^\sharp
\end{aligned}$$

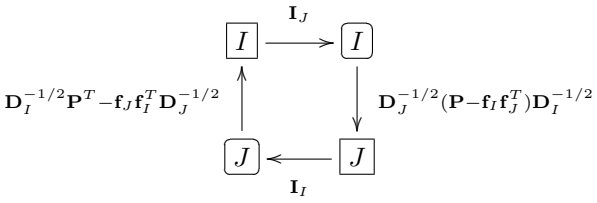
Il ne s'agit pas, d'ailleurs, de l'inverse généralisé de Moore-Penrose. Pour achever l'analyse discriminante, il faut alors calculer le poids des classes qui se retrouve dans $\mathbf{D}_J$ et le tableau des centres de gravité par classe. Il suffit d'additionner les profils (disjonctifs) des porteurs de la modalité j donc de les compter par modalité i puis diviser par leur effectif, ce qui redonne le profil conditionnel de la colonne j. Le tableau des centres de gravité est donc $\mathbf{D}_J^{-1}\mathbf{P}^T$ qui donne le schéma :



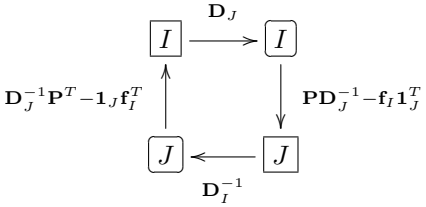
On le réécrit sans le modifier par transposition (ce que fait la fonction `t.dudi` dans `ade4`) :

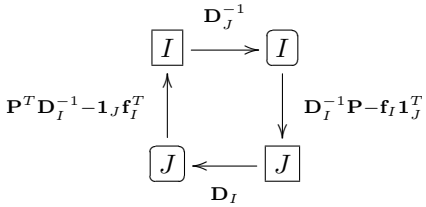


où on reconnaît l'analyse des correspondances, ou du moins une nouvelle forme d'un schéma dont on connaissait cinq versions équivalentes [7] à commencer par celle du programme (dans sa version la plus commune) :

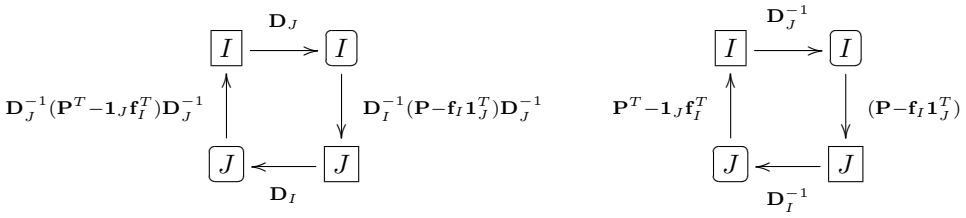


puis deux ACP sur profils avec la métrique du χ^2 :

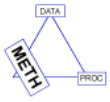




et deux ACP symétriques :



L'AFC est une analyse discriminante. M. O. Hill a complété son introduction célèbre en écologie[13] par un article[14] dont le titre au moins a beaucoup irrité et une reconnaissance claire de ce fait[15] (discrimination des sites par les espèces). Il faut insister sur le fait que c'est une ***double*** analyse discriminante : quand on ne se sert que de l'une des deux, l'autre est présente et devient une contrainte (discrimination des espèces par les sites).

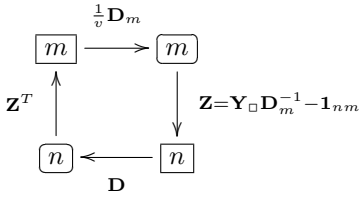


5 Dscriminations sur variables qualitatives

Notons $\mathbf{Y}$ un tableau portant sur n individus (lignes) et v facteurs (colonnes). L'équivalent de l'ACP de ce type de tableau est l'ACM (analyse des correspondances multiples) qui demande d'identifier les effectifs de modalités par variable et le nombre total de modalités :

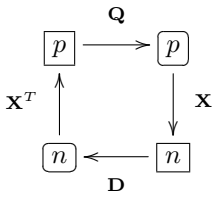
$$m = m_1 + \cdots + m_v$$

le tableau disjonctif complet $\mathbf{Y}_\square$ avec n lignes et m colonnes (voir `tdr52`), la diagonale des poids des individus $\mathbf{D}$ (en général, $(1/n)\mathbf{I}_n$), la diagonale des poids des modalités par variable $\mathbf{D}_m$ (diagonale des sommes des poids des individus porteurs par modalité) et se définit comme l'analyse du schéma :



Il est important de noter que le nombre de modalités de l'ACM remplace le nombre de variables de l'ACP et que le tableau $\mathbf{Z}$ a un rang inférieur ou égal à $m - v$. Une difficulté supplémentaire intervient par la présence d'une diagonale des poids des modalités qui remplace la matrice identité. Pour étendre l'analyse discriminante à tout type de tableau, il faut utiliser une extension de l'inverse généralisée de Moore-Penrose (IGMP), sous la forme des IMGP pondérés définis dans [1, Annexe p. 30-32].

Pour simplifier, en partant du schéma :



(3)

on note r le rang du schéma, $\mathbf{U}_r$ la base des axes principaux $\mathbf{Q}$-orthonormés et $\mathbf{\Lambda}_r$ la diagonale des valeurs propres non nulles du schéma. Les deux matrices :

$$\mathbf{C} = \mathbf{X}^T\mathbf{D}\mathbf{X} \qquad \mathbf{W} = \mathbf{Q}\mathbf{U}_r\mathbf{\Lambda}_r^{-1}\mathbf{U}_r^T\mathbf{Q}$$

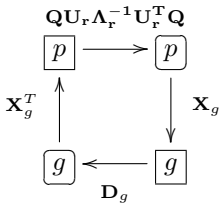
vérifient les relations :

$$\begin{aligned} \mathbf{WC} &= \mathbf{Q}\mathbf{U}_r\mathbf{U}_r^T \\ \mathbf{CW} &= \mathbf{U}_r\mathbf{U}_r^T\mathbf{Q} \\ \mathbf{WCW} &= \mathbf{W} \\ \mathbf{CWC} &= \mathbf{C} \end{aligned}$$

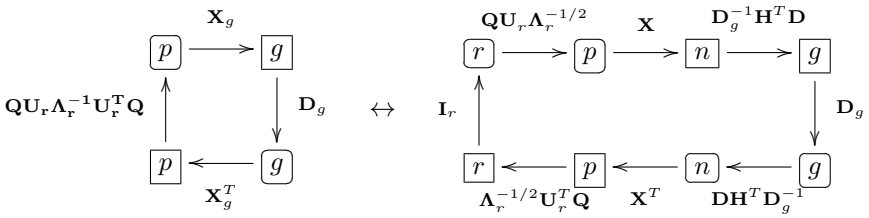
$\mathbf{W}$ est le $\mathbf{Q}$-inverse généralisé de $\mathbf{C}$. On le notera :

$$\mathbf{W} = \mathbf{C}_{\mathbf{Q}}^{+}$$

Pour une partition des n lignes et g groupes exprimée par la matrices des indicatrices de classes $\mathbf{H}$ laquelle définit la pondération des classes $\mathbf{D}_g = \mathbf{H}^T\mathbf{D}\mathbf{H}$, le projecteur $\mathbf{P_H} = \mathbf{H}\mathbf{D}_g^{-1}\mathbf{H}\mathbf{D}$, l'opérateur d'averaging (qui fait les moyennes par classes et par variables) $\mathbf{K_H} = \mathbf{D}_g^{-1}\mathbf{H}^T\mathbf{D}$, et le tableau des moyennes par classes $\mathbf{X}_g = \mathbf{K_H}\mathbf{X}$, l'analyse discriminante du schéma 3 est celle du schéma :



En développant :



on voit simplement que l'analyse discriminante du schéma 3 est l'analyse interclasses du tableau des coordonnées réduites (composantes principales) issu du

schéma 3. L'usage des composantes principales de l'ACP dans l'analyse discriminante linéaire (qui suit l'usage des composantes principales de l'ACP dans la régression multiple ou **PCR** *Principal Components Regression*) est une idée ancienne et utile [16]. On en a ici une forme canonique qui permet d'étendre l'ADL à tout type de variables. On retrouve alors l'analyse discriminante classique, sur variables qualitatives, des correspondances, des mélanges comme cas particulier.

6 Truites ancestrales et modernes

On utilise comme illustration les travaux de J.M. Lascaux et le jeux de données décrit dans :

`pbil.univ-lyon1.fr/R/pps/pps022.pdf`

Il y a 306 individus. La variable qui définit les groupes est typiquement un facteur à modalités ordonnées. Elle indique une appartenance génétique à 7 modalités notées de 0 à 6. 0 indique un poisson assimilé à un homozygote atlantique (domestique ou moderne) et 6 indique un poisson assimilé à un homozygote méditerranéen (ancestral ou sauvage).

```
data(lascaux)
gen <- lascaux$gen
summary(gen)
```

```
 0  1  2  3  4  5  6
73 41 21 24 32 32 83
```

Les variables concernent la morphométrie :

```
A <- lascaux$morpho[unlist(lapply(lascaux$morpho,function(x) !any(is.na(x))))]
Alog <-  bicenter.wt(log(A))
A.pca <- dudi.pca(Alog, scale=F, scan=F,center=F,nf=4)
A.dis <-discrimin(A.pca,gen,scan=F)
```

la coloration de la robe :

```
B<- lascaux$colo
B.pca <- dudi.pca(B,scan=F,nf=4)
B.dis <- discrimin(B.pca,gen,scan=F)
```

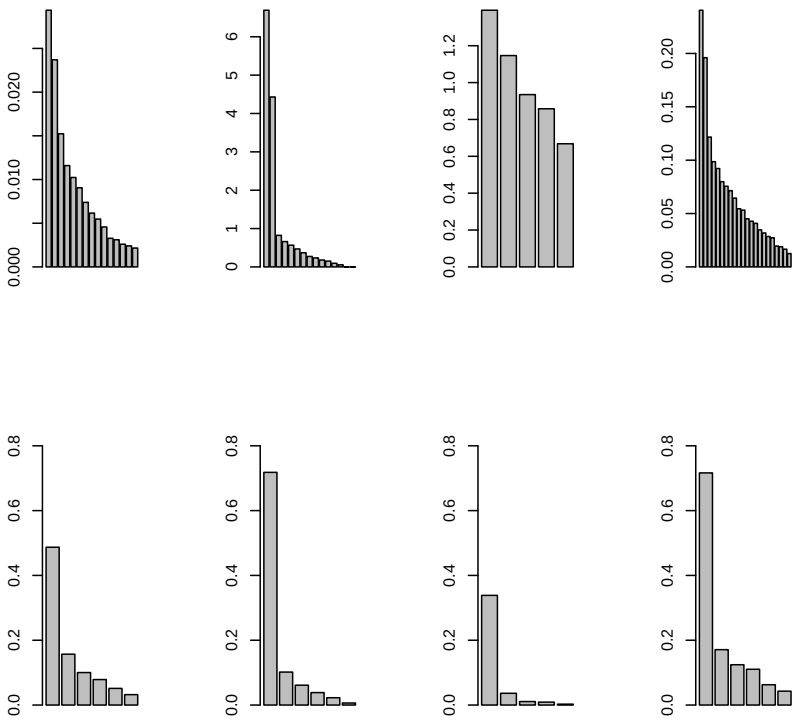
les variables méristiques, indépendantes de la croissance :

```
C <- lascaux$meris
C.pca <- dudi.pca(C,scan=F,nf=4)
C.dis <- discrimin(C.pca,gen,scan=F)
```

et l'ornementation :

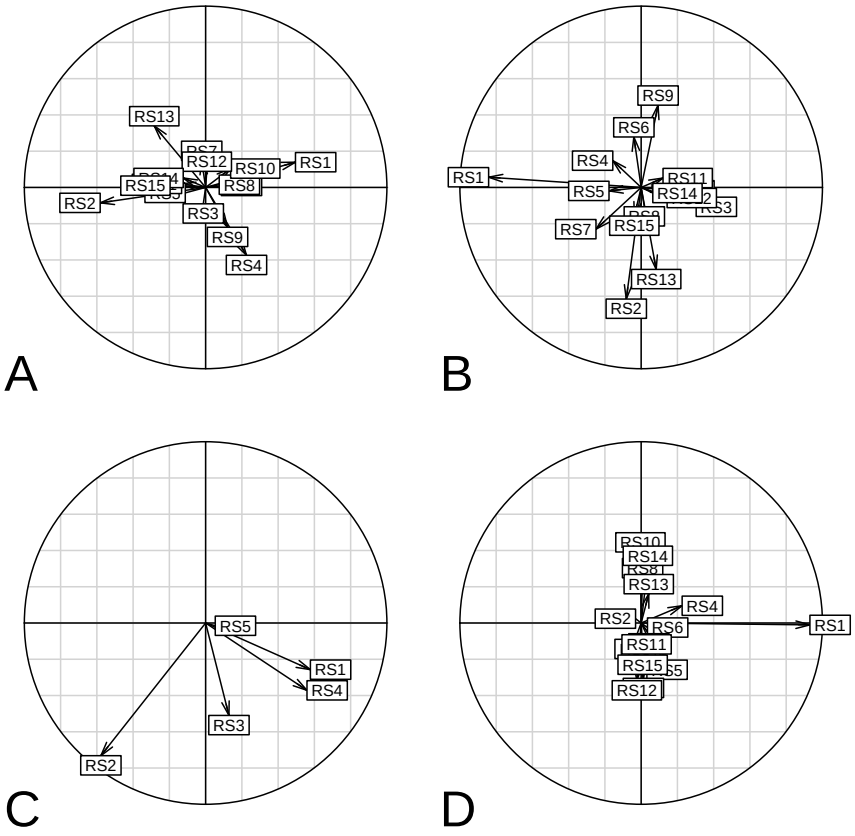
```
D <- lascaux$ornem
D.acm <- dudi.acm(D,scan=F,nf=4)
D.dis <- discrimin(D.acm,gen, scan=F)
```

On trace, en haut, dans l'ordre A, B, C et D les valeurs propres des analyses simples, en bas les valeurs propres des analyses discriminantes.



Quelle que soit la structure du tableau, il n'existe qu'une fonction discriminante pertinente. Le pouvoir discriminant de la seconde et la quatrième l'emporte largement sur celle des deux autres. Cette différence de qualité se retrouve dans la définition des variables canoniques. C'est pratiquement une composante principale pour la seconde et la quatrième, l'interprétation étant beaucoup plus difficile pour les deux autres.

```
par(mfrow=c(2,2))
s.corcircle (A.dis$cp[1:15,],sub="A",csub=3)
s.corcircle (B.dis$cp[1:15,],sub="B",csub=3)
s.corcircle (C.dis$cp,sub="C",csub=3)
s.corcircle (D.dis$cp[1:15,],sub="D",csub=3)
```

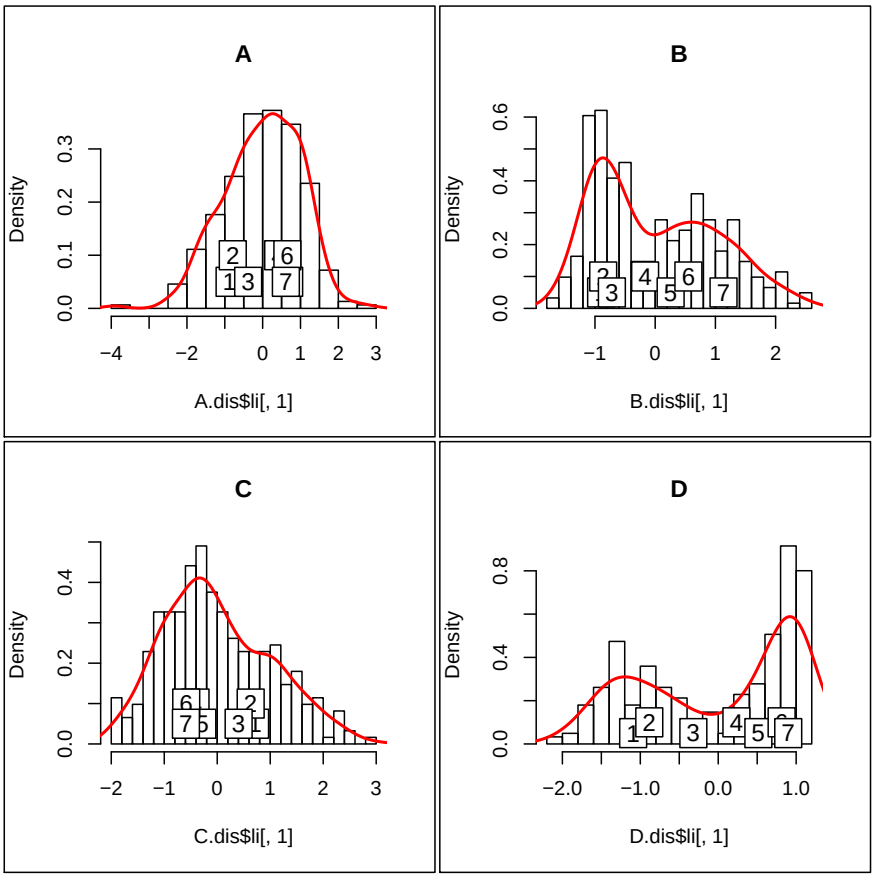


De même les variables canoniques les plus liées sont la seconde et la quatrième :

```
cor(cbind.data.frame(A.dis$li[,1],B.dis$li[,1],C.dis$li[,1],D.dis$li[,1]))
```

```
              A.dis$li[, 1] B.dis$li[, 1] C.dis$li[, 1] D.dis$li[, 1]
A.dis$li[, 1]     1.0000000     0.5757384    -0.5950184     0.6540988
B.dis$li[, 1]     0.5757384     1.0000000    -0.5120205     0.7585973
C.dis$li[, 1]    -0.5950184    -0.5120205     1.0000000    -0.5367691
D.dis$li[, 1]     0.6540988     0.7585973    -0.5367691     1.0000000
```

Mais chaque groupe de variables permet d'approcher la même structure :



On pourra continuer l'exploration de ce remarquable ensemble de données, en particulier mélanger les types de variables en passant par `dudi.mix`. On retiendra que la discrimination linéaire s'étend aux variables qualitatives [27]. Elle est cependant moins utilisée (et moins accessible) que la discrimination barycentrique [20] [21] qui est l'inter-classe correspondante et qui n'est rien d'autre que l'AFC du tableau croisé juxtaposant les table de contingence :

```
D.bet <- bca(D.acm,gen,scan=F)
D.bet$eig
```

```
[1] 0.155873510 0.008430259 0.007513842 0.006100640 0.003471782 0.002453564
```

```
E <- acm.burt(D,as.data.frame(gen))
dudi.coa(E,scan=F)$eig
```

```
[1] 0.155873510 0.008430259 0.007513842 0.006100640 0.003471782 0.002453564
```

```
cor(D.bet$ls[,1],D.dis$li[,1])
```

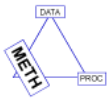
```
[1] 0.962197
```

On pourra éditer le tableau E, en comprendre la structure, montrer que son AFC est une AFC intra-classe (modalités par variables) implicite, et/ou observer qu'ici l'inter-classe et la discriminante sur variables qualitatives sont voisines. Mais il peut en être autrement.

7 Analyse Discriminante des Correspondances

Toute analyse de base induit une analyse inter-classe et une analyse discriminante associée. Cette extension à l'AFC était déjà disponible dans le logiciel ADE-4 (`http://pbil.univ-lyon1.fr/R/thema36.pdf`) et a été reprise dans [22] et [23]. Elle est en œuvre dans `discrimin.coa`. Nous l'appellerons ADC.

Il est utile de comprendre que l'AFC présente plusieurs facettes et des schémas cohérents mais multiples : la variante qui permet la généralisation de l'ADL est un de ceux-ci.



7.1 Base de l'ADC

Considérons un tableau $\mathbf{X}$ de nombres positifs ou nuls supportant une analyse des correspondances. L'introduction d'une partition de l'ensemble des lignes en groupes détruit, comme tout apport d'information sur une seule des deux marges, la symétrie naturellement en jeu dans le schéma classique de l'AFC. Il y a sous-jacent à une analyse des correspondances 4 triplets statistiques signifiants[7]. On utilise les notations classiques [18, pp. 67-107]. $\mathbf{X}$ a n lignes et p colonnes.

$$x_{\bullet\bullet} = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij} \quad f_{ij} = \frac{x_{ij}}{x_{\bullet\bullet}} \quad \mathbf{F} = [f_{ij}]$$

$$f_{i.} = \sum_{j=1}^{p} f_{ij} \quad \mathbf{D}_n = \mathrm{diag}\,(f_{i.}) \quad f_{.j} = \sum_{i=1}^{n} f_{ij} \quad \mathbf{D}_p = \mathrm{diag}\,(f_{.j})$$

Les diagonales des poids dérivent des données. Le triplet habituel est [9] :

$$\left(\mathbf{D}_n^{-1}\mathbf{F}\mathbf{D}_p^{-1} - \mathbf{U}_{np}, \mathbf{D}_p, \mathbf{D}_n\right)$$

C'est celui d'une double analyse d'inertie. On retrouve la même matrice à diagonaliser dans trois autres triplets. Le premier est :

$$\left(\mathbf{F}\mathbf{D}_p^{-1} - \mathbf{D}_n\mathbf{U}_{np}, \mathbf{D}_p, \mathbf{D}_n^{-1}\right)$$

Les colonnes du tableau sont les écarts entre les distributions conditionnelles par colonne et la distribution marginale des lignes. Les poids des colonnes sont les poids ordinaires et la métrique dans l'espace des colonnes est celles du Khi2 entre colonnes. On retrouve l'analyse de l'ensemble des profils colonnes. Le second est :

$$\left(\mathbf{D}_n^{-1}\mathbf{F} - \mathbf{U}_{np}\mathbf{D}_p, \mathbf{D}_p^{-1}, \mathbf{D}_n\right)$$

Les lignes du tableau sont les écarts entre les distributions conditionnelles par ligne et la distribution marginale des colonnes. Les poids des lignes sont les poids ordinaires et la métrique dans l'espace des lignes est celles du Khi2 entre lignes. On retrouve l'analyse de l'ensemble des profils lignes. Le dernier est :

$$\left(\mathbf{F} - \mathbf{D}_n\mathbf{U}_{np}\mathbf{D}_p, \mathbf{D}_p^{-1}, \mathbf{D}_n^{-1}\right)$$

Les valeurs du tableaux sont $f_{ij} - f_{i.}f_{.j}$, les écarts au modèle d'indépendance et les deux métriques sont celles de Mahalanobis associées aux indicatrices des

classes (modèle de l'AFC comme analyse canonique explicité dans [19] et utilisé dans [29]).

C'est la présence implicite de ces quatre triplets qui donne à l'AFC une place centrale en analyse des données. La section 1.3 de [18] fait le tour de cette question définitivement. Ces éléments sont également détaillés dans [26].

A cause de la dissymétrie lignes-colonnes introduite par la partition des lignes on retiendra pour la suite la présentation à partir du nuage des distributions par lignes soit :

$$\left(\mathbf{D}_n^{-1}\mathbf{F} - \mathbf{U}_{np}\mathbf{D}_p, \mathbf{D}_p^{-1}, \mathbf{D}_n\right).$$

On peut également considérer le triplet $\left(\mathbf{D}_n^{-1}\mathbf{F}, \mathbf{D}_p^{-1}, \mathbf{D}_n\right)$. Le non centrage introduit un vecteur propre artificiel associé à la valeur propre 1 classique dans [4] qui est ensuite éliminée explicitement. Le nuage est dans le sous-espace affine de $\mathbb{R}^p$ défini par l'équation $\sum_{j=1}^{p} x_j = 1$. Ce point de vue a l'intérêt de simplifier la présentation théorique mais est maintenant abandonné dans les programmes.

Soit alors $\mathbf{P}_{\mathbf{A}_0} = \mathbf{P}_{\mathbf{A}} - \mathbf{P}_{\mathbf{1}_n} = \mathbf{A}\left(\mathbf{A}^T\mathbf{D}_n\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{D}_n - \mathbf{U}_{nn}\mathbf{D}_n$ le projecteur sur le sous-espace de $\mathbb{R}^n$ associé à la partition des lignes définie par son tableau d'indicatrices $\mathbf{A}$.

Pour que la projection ait un sens statistique il faut que la dimension de ce sous-espace soit largement inférieure à n. Dans ces conditions, la matrice $\mathbf{F}^T\mathbf{D}_n^{-1}\mathbf{F}$ est de rang plein et son inverse symétrique et positive est un produit scalaire de $\mathbb{R}^p$. On peut alors définir deux nouvelles analyses. La première est associée au triplet

$$\left(\mathbf{P}_{\mathbf{A}_0}\mathbf{D}_n^{-1}\mathbf{F}, \mathbf{D}_p^{-1}, \mathbf{D}_n\right)$$

C'est tout simplement celui de l'AFC du tableau des sommes par classes des lignes du tableau $\mathbf{F}$, qui est encore celui de l'AFC inter-classes. Il s'associe avec l'analyse du triplet

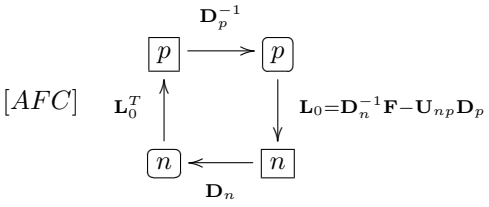
$$\left(\mathbf{D}_n^{-1}\mathbf{F} - \mathbf{P}_{A_0}\mathbf{D}_n^{-1}\mathbf{F}, \mathbf{D}_p^{-1}, \mathbf{D}_n\right)$$

qui est celui de l'analyse intra-classe[3]. Le couple est largement utilisé en hydrobiologie [17].

La seconde est associée au triplet $\left(\mathbf{P}_{\mathbf{A}_0}\mathbf{D}_n^{-1}\mathbf{F}, \left(\mathbf{F}^T\mathbf{D}_n^{-1}\mathbf{F}\right)^{-1}, \mathbf{D}_n\right)$ qui est celui de l'analyse discriminante associée à l'analyse des correspondances. On l'appellera analyse discriminante des correspondances.

7.2 Propriété principale de l'ADC

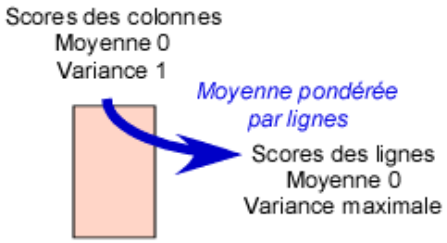
On utilise ici les propriétés générales des schémas de dualité [10] [6]. Nous avons retenu pour l'AFC le schéma :



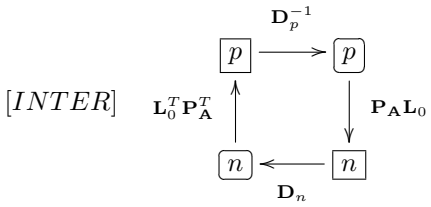
Un axe principal est un vecteur $\mathbf{a}$ de $\mathbb{R}^p$ qui est $\mathbf{D}_p^{-1}$-normé. Son image est un vecteur $\left(\mathbf{D}_p^{-1}\right)^{-1} = \mathbf{D}_p$-normé. C'est un score des colonnes centré et réduit pour la distribution marginale des colonnes qui maximise :

$$\left\|(\mathbf{D}_n^{-1}\mathbf{F} - \mathbf{U}_{np}\mathbf{D}_p)\mathbf{a}\right\|^2_{\mathbf{D}_n}$$

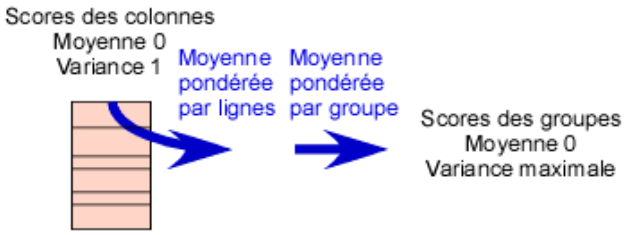
c'est-à-dire la variance des moyennes conditionnelles par lignes. Ce qu'on peut résumer par le petit schéma de principe :



L'AFC inter-classe est celle du schéma :



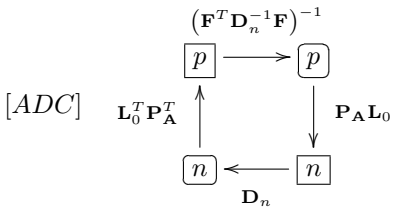
Le raisonnement est le même. Le projecteur rajoute l'opération qui consiste à calculer les moyennes par groupes des scores obtenus comme moyennes par lignes. On passe ainsi de la variance à la variance inter-classe, ce qui correspond au schéma de principe :



Noter que la moyenne pondérée dans un groupe des moyennes pondérées par distributions lignes du groupe est exactement la moyenne pondérée de la distribution globale du groupe. Si la moyenne l'a emporté sur la médiane c'est parce qu'elle rend possible cette unité.

Ceci permet de replacer la propriété fondamentale de l'ADC dont le schéma

est :



Comme

$$\mathbf{U}_{nn}\mathbf{F} = \mathbf{U}_{np}\mathbf{D}_p$$

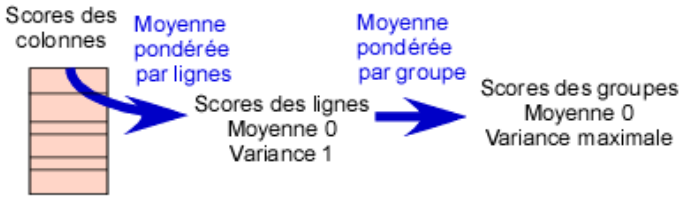
le centrage du nuage des lignes est restauré par la projection dans l'espace des colonnes. Un axe principal est un vecteur de $\mathbb{R}^p$ $\mathbf{a}$ qui est $\left(\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}\right)^{-1}$-normé. Son image $\mathbf{b} = \left(\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}\right)^{-1}\mathbf{a}$ est une forme linéaire exprimée dans la base duale de la base canonique. Elle est $\left(\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}\right)$-normée. C'est donc un score des colonnes qui a pour propriété :

$$\|\mathbf{b}\|^2_{\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}} = \mathbf{b}^t\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}\mathbf{b} = \mathbf{b}^t\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{D}_n\mathbf{D}_n^{-1}\mathbf{F}\mathbf{b} = \left\|\mathbf{D}_n^{-1}\mathbf{F}\mathbf{b}\right\|^2_{\mathbf{D}_n} = 1$$

Le vecteur $\mathbf{1}_p$ est dans le noyau de $\mathbf{P}_{A_0}\mathbf{D}_n^{-1}\mathbf{F}$. C'est donc un facteur principal associé à une valeur propre nulle et le vecteur $\mathbf{b}$ lui est $\left(\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}\right)$-orthogonal. Soit :

$$\mathbf{b}^t\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{F}\mathbf{1}_p = 0 = \mathbf{b}^t\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{D}_n\mathbf{D}_n^{-1}\mathbf{F}\mathbf{1}_p = \mathbf{b}^t\mathbf{F}^t\mathbf{D}_n^{-1}\mathbf{D}_n\mathbf{1}_n = 0$$

Ceci indique que le score moyen obtenu par averaging sur chaque ligne du score $\mathbf{b}$ est de moyenne nulle et de variance 1. Les variables canoniques de cette analyse sont donc obtenus par averaging et ils sont de variance inter-classe maximale. On a donc le schéma de principe :

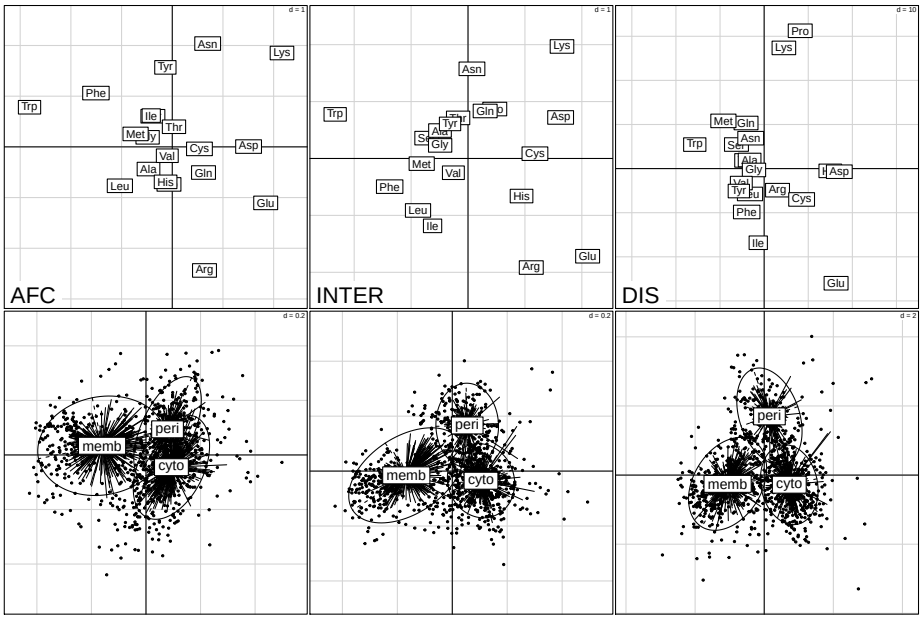


Ces schémas aident l'utilisateur potentiel. On a déplacé une contrainte sur la variance pour passer de l'optimisation de la variance inter-classe à l'optimisation du pourcentage de variance inter-classe. Il faut vérifier que la discrimination entre classes, qui a nécessairement augmenté, se fait encore avec des opérations qui ont un sens expérimental.

7.3 Exemples

L'unité des logiques sous-jacentes est bien perceptible dans cet exemple (904 protéines en 3 classes et 20 acides aminés) :

```
data(perthi02)
p.coa <- dudi.coa(perthi02$tab,scannf=F)
p.bet <- bca(p.coa,perthi02$cla,scannf=F)
p.dis <- discrimin.coa(perthi02$tab,perthi02$cla,scannf=F)
par(mfrow=c(2,3))
s.label(p.coa$c1,sub="AFC",csub=3,clab=1.5)
s.label(p.bet$c1,sub="INTER",csub=3,clab=1.5)
s.label(p.dis$fa,sub="DIS",csub=3,clab=1.5)
s.class(p.coa$li,perthi02$cla,csta=0.5,clab=2)
s.class(p.bet$ls,perthi02$cla,csta=0.5,clab=2)
s.class(p.dis$li,perthi02$cla,csta=0.5,clab=2)
```



Les trois analyses sont ici voisines, mais ce n'est pas une obligation. L'objectif de discriminer oublie toute autre contrainte. Dans le prochain exemple (302 relevés répartis en 12 classes pour 51 espèces d'oiseaux), la structure (AFC) est clairement la structure INTER-classe (le plan d'observation gouverne la variabilité faunistique) mais la discriminante (DIS) trouvera 5 variables canoniques indépendantes présentant plus de 80% de variabilité inter-classe.

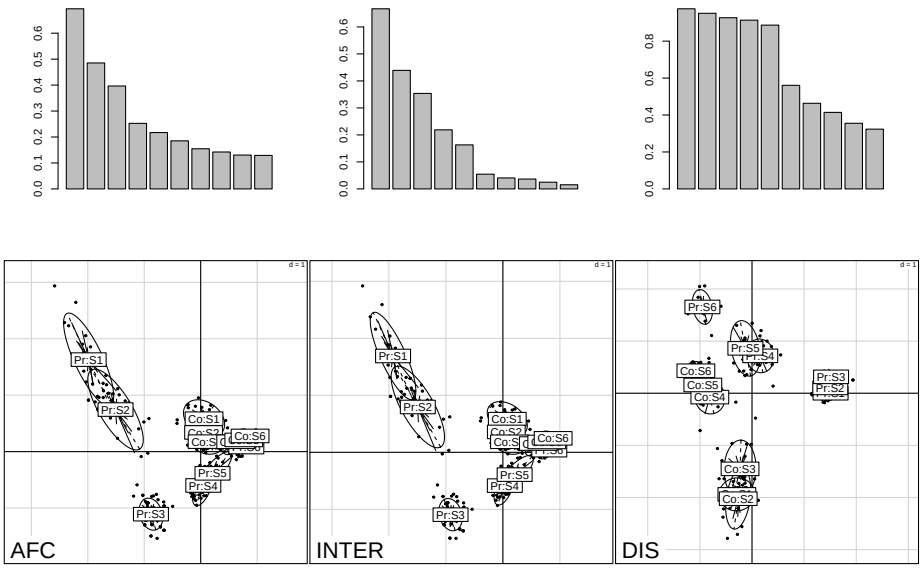
```
data(avimedi)
names(avimedi)
```

```
[1] "fau"    "plan"   "nomesp"
```

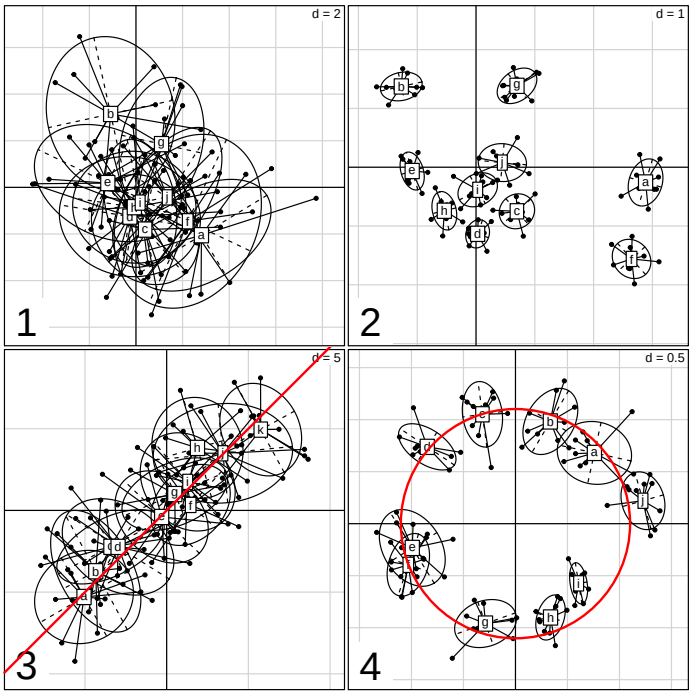
```
cla <- avimedi$plan$reg:avimedi$plan$str
summary(cla)
```

```
Pr:S1 Pr:S2 Pr:S3 Pr:S4 Pr:S5 Pr:S6 Co:S1 Co:S2 Co:S3 Co:S4 Co:S5 Co:S6
   24    24    70    28    16    16    16    16    22    18    24    28
```

```
a.coa <- dudi.coa(avimedi$fau,scannf=F)
a.bet <- bca(a.coa,cla,scannf=F)
a.dis <- discrimin.coa(avimedi$fau,cla,scannf=F)
par(mfcol=c(2,3))
barplot(a.coa$eig[1:10])
s.class(a.coa$li,cla,csta=0.5,sub="AFC",csub=3,clab=1.5)
barplot(a.bet$eig[1:10])
s.class(a.bet$ls,cla,csta=0.5,sub="INTER",csub=3,clab=1.5)
barplot(a.dis$eig[1:10])
s.class(a.dis$li,cla,csta=0.5,sub="DIS",csub=3,clab=1.5)
```



En haut, observer le même nuage de centre de gravité sans structure. En bas les centres de gravité forment une figure simple. A gauche, on ne peut pas discriminer, c'est-à-dire déterminer le groupe avec les valeurs. A droite, c'est relativement aisé.



Prédire le groupe conduit à des méthodes spécialisées et performantes (discrimination PLS, régression logistique, discrimination non paramétrique, réseau de neurones, ...) : l'analyse discrimnante linéaire est un point d'entrée. Etudier

la structure des groupes est plus typiquement le rôle de l'analyse multivariée. L'analyse discriminante est alors une option, les analyses inter-classes sont plutôt un standard. Noter que l'analyse des correspondances inter-classes s'appelle discrimination barycentrique, ce qui ne simplifie pas la question.

Références

[1] R. Abdesselam and Y. Schektman. Une analyse factorielle de l'association dissymétrique entre deux variables qualitatives. *Revue de Statistique Appliquée*, 44(2) :5–34, 1996.

[2] A. Ben-Israel and T.N.E. Greville. *Generalized Inverses : Theory and Applications*. Wiley-Interscience, 1974.

[3] J.P. Benzécri. Analyse de l'inertie intra-classe par l'analyse d'un tableau de correspondances. *Les Cahiers de l'Analyse des données*, 8 :351–358, 1983.

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

[5] T.P. Burnaby. Growth-invariant discriminant functions and generalized distances. *Biometrics*, 22 :96–110, 1966.

[6] F. Cailliez and J.P. Pagès. *Introduction à l'analyse des données*. SMASH, 9 rue Duban, 75016 Paris, 1976.

[7] P. Cazes, D. Chessel, and S. Dolédec. L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Revue de Statistique Appliquée*, 36 :39–54, 1988.

[8] D. Chessel and S. Dolédec. *ADE Version 3.6 : HyperCard Stacks and Programme library for the Analysis of Environmental Data. Manuel d'utilisation. 8 fascicules, 750 pp.* URA CNRS 1451, Université Lyon 1, 69622 Villeurbanne cedex, 1993.

[9] Y. Escoufier. L'analyse des tableaux de contingence simples et multiples. *Metron*, 40 :53–77, 1982.

[10] Y. Escoufier. The duality diagramm : a means of better practical applications. In P. Legendre and L. Legendre, editors, *Development in numerical ecology*, pages 139–156. NATO advanced Institute , Serie G .Springer Verlag, Berlin, 1987.

[11] R.A. Fisher. The precision of discriminant functions. *Annals of Eugenics*, 10 :422–438., 1940.

[12] J.C. Gower. Growth-free canonical variates and generalized inverses. *Bulletin of the Geological Institutions of the University of Uppsala, N.S*, 7 :1–10, 1976.

[13] M.O. Hill. Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology*, 61 :237–249, 1973.

[14] M.O. Hill. Correspondence analysis : A neglected multivariate method. *Applied Statistics*, 23 :340–354, 1974.

[15] 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.

[16] Ian T. Jolliffe, Byron J.t. Morgan, and Philip J. Young. A simulation study of the use of principal components in linear discriminant analysis. *Journal of statistical computation and simulation*, 55(4) :353–366, 1996.

[17] N. Lair and D. Sargos. A 10 years study at four sites of the middle course of the river loire. i - patterns of change in hydrological, physical and chemical variables in relation to algal biomass. *Hydroécologie Appliquée*, 5 :1–27, 1993.

[18] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.

[19] L. Lebart, L. Morineau, and K.M. Warwick. *Multivariate descriptive analysis : correspondence and related techniques for large matrices*. John Wiley and Sons, New York, 1984.

[20] A. Leclerc. L'analyse des correspondances sur juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, XXIII :5–16, 1975.

[21] A. Leclerc. Une étude de la relation entre une variable qualitative et un groupe de variables qualitatives. *International Statistical Review*, 44 :241–248, 1976.

[22] G. Perrière, J.R. Lobry, and J. Thioulouse. Correspondence discriminant analysis : a multivariate method for comparing classes of protein and nucleic acid sequences. *CABIOS*, 12 :519–524, 1996.

[23] G. Perrière and J. Thioulouse. Use of correspondence discriminant analysis to predict the subcellular location of bacterial proteins. *Computer Methods and Programs in Biomedicine*, 70 :99–105, 2003.

[24] H. Persat. *De la biologie des populations de l'Ombre commun (Thymallus thymallus (L. 1758)) à la dynamique des communautés dans un hydrosystème fluvial aménagé, le Haut-Rhône français. Eléments pour un changement d'échelles*. Thèse d'état, Université Lyon 1, 1988.

[25] C.R. Rao. A note on a generalized inverse of a matrix with applications to problems in mathematical statistics. *Journal of the Royal Statistical Society, B*, 24 :152–158, 1962.

[26] H. Rouanet and B. Le Roux. *Analyse des données multidimensionnelles*. Dunod, paris, 1993.

[27] G. Saporta. *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. PhD thesis, Thèse de 3° cycle, Université Pierre et Marie Curie, 1975.

[28] C. Surre, H. Persat, and J.M. Gaillard. A biometric study of three populations of the european grayling, thymallus thymallus (l.), from the french jura mountains. *Canadian Journal of Zoology*, 64 :2430–2438, 1986.

[29] J. Thioulouse and D. Chessel. A method for reciprocal scaling of species tolerance and sample diversity. *Ecology*, 73 :670–680, 1992.

[30] R. Tomassone, M. Danzard, J.J. Daudin, and J.P. Masson. *Discrimination et classement*. Masson, Paris, 1988.

[31] N. G. Yoccoz. *Le rôle du modèle euclidien d'analyse des données en biologie évolutive*. PhD thesis, Thèse de doctorat, Université Lyon 1, 1988.