Fiche de Biostatistique - Stage 5

Couplages de tableaux

D. Chessel, A.B. Dufour & J. Thioulouse

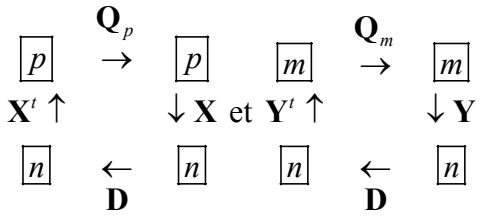
Résumé

La fiche introduit aux principales méthodes de couplage de deux tableaux.

Plan

1. Introduction

En toute généralité, une analyse à un tableau se décrit par un schéma de dualité. Pour deux tableaux on a deux schémas. Ils seront considérés comme cohérents s'ils partagent un espace euclidien sous-jacent. On supposera qu'ils sont appariés par les lignes (il suffit de transposer si ce n'est pas le cas). On a alors deux schémas appariés :

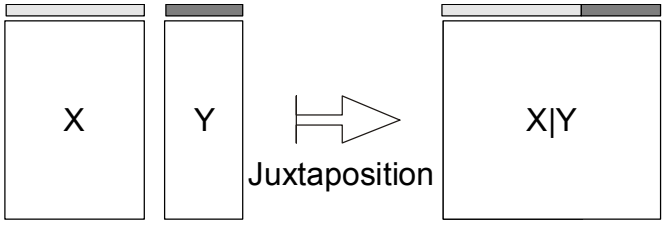


Il y a trois stratégies principales d'association de deux triplets. Elles recouvrent, à cause des nombreux jeux de paramètres, un vaste ensemble de pratiques. Il suffit de saisir ces trois principes pour faire un choix, voire pour construire des associations originales dont on peut avoir besoin.

Le couplage de deux tableaux de données est une opération fondamentale en écologie statistique. On dispose d'une énorme littérature sur le sujet. Parmi les bases historiques on doit rappeler quelques opérations fondamentales.

1.1. Juxtaposition

Deux tableaux ayant les mêmes lignes sont simplement accolés pour former un nouveau tableau qui appelle une analyse simple. La voie a été ouverte par P. Dagnélie (1965).



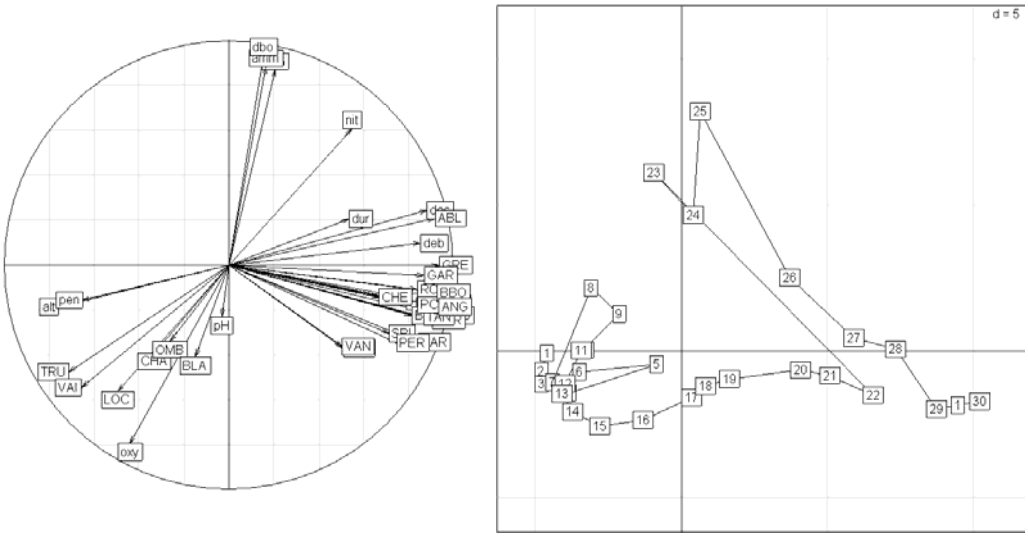
On associe un tableau 30 lignes-stations et 11 colonnes-variables (tableau mésologique) et un tableau 30 lignes-stations et 27 colonnes-espèces (tableau faunistique) extraits d'une thèse célèbre en hydrobiologie (Verneaux 1973).

```
library(ade4)
data(doubs)
summary(doubs)
    Length Class      Mode
mil 11     data.frame list
poi 27     data.frame list
xy   2     data.frame list
```

On accole les deux tableaux et on soumet le résultat à une ACP normée :

```
w = cbind.data.frame(doubs$mil, doubs$poi)
pcaw = dudi.pca(w,scan=F)
```

```
s.corcircle(pcaw$co)
s.traject(pcaw$li)
s.label(pcaw$li,add.plot=T)
```



A gauche l'amont, à droite l'aval, en haut la pollution

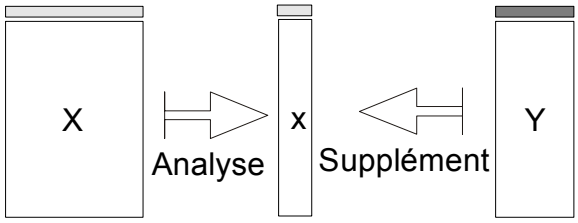
$$\boxed{p}\ \mathbf{X}_*^t \uparrow \quad \overset{\mathbf{I}_p}{\rightarrow} \quad \boxed{p}\ \downarrow \mathbf{X}_* \qquad \boxed{m}\ \mathbf{Y}_*^t \uparrow \quad \overset{\mathbf{I}_m}{\rightarrow} \quad \boxed{m}\ \downarrow \mathbf{Y}_* \qquad \mapsto \qquad \begin{matrix} \boxed{p+m} \\ \begin{bmatrix} \mathbf{X}_*^t \\ \mathbf{Y}_*^t \end{bmatrix} \uparrow \\ \boxed{n} \end{matrix} \quad \overset{\mathbf{I}_{p+m}}{\rightarrow} \quad \begin{matrix} \boxed{p+m} \\ \downarrow [\mathbf{X}_* \quad \mathbf{Y}_*] \\ \boxed{n} \end{matrix}$$

$$\boxed{n} \quad \overset{\frac{1}{n}\mathbf{I}_n}{\leftarrow} \quad \boxed{n} \qquad\qquad \boxed{n} \quad \overset{\frac{1}{n}\mathbf{I}_n}{\leftarrow} \quad \boxed{n} \qquad\qquad \boxed{n} \quad \overset{\frac{1}{n}\mathbf{I}_n}{\leftarrow} \quad \boxed{n}$$

$\mathbf{X}_*$ et $\mathbf{Y}_*$ désignent les tableaux de variables normalisées. C'est le premier pas vers l'analyse factorielle multiple (AFM voir la fiche *K*-tableaux). Cette approche ne fonctionne que si les inerties des deux tableaux sont comparables. Si l'un des deux l'emporte largement, il imposera son point de vue. L'AFM propose des modifications pour pallier au défaut et s'applique alors à un nombre quelconque de tableaux.

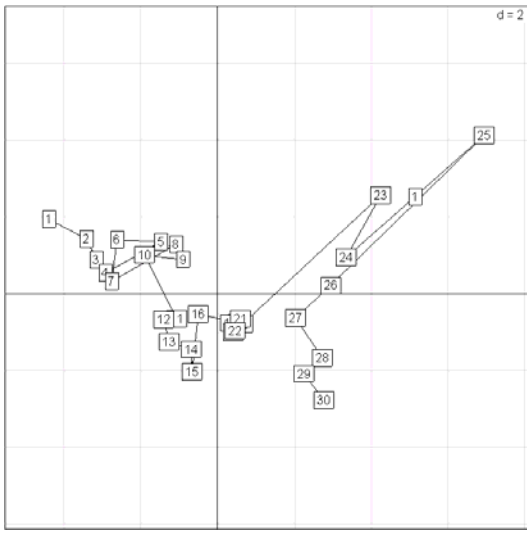
1.2. Illustration

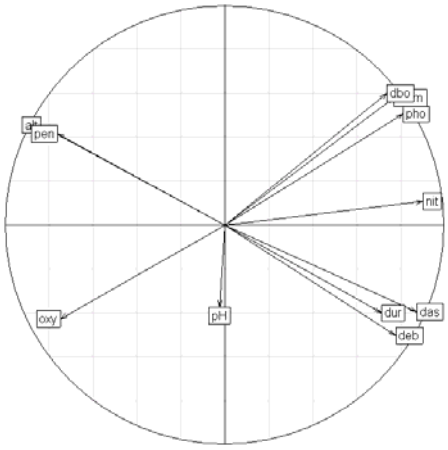
On pratique l'analyse d'un des deux tableaux et on introduit dans l'interprétation des éléments issus de l'autre. R. H. Whittaker, théoricien fondateur de l'analyse des données écologiques distingue ainsi l' ordination directe et l'ordination indirecte (Whittaker 1967).



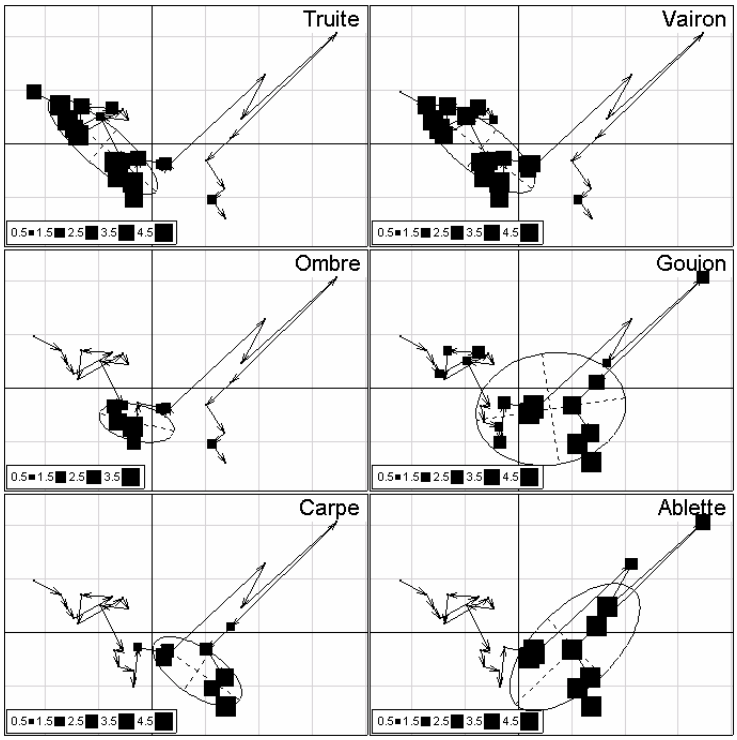
Dans la première, on analyse le tableau de milieu et on étudie la répartition des espèces.

```
pcamil = dudi.pca(doubs$mil, scann=F)
s.corcircle(pcamil$co)
s.traject(pcamil$li)
s.label(pcamil$li,add.plot=T)
```





A gauche et légèrement en haut l'amont , à droite et légèrement en bas l'aval, en haut un peu en amont et beaucoup en aval la pollution. La charge minérale et organique augmente quand on descend une rivière mais les deux phénomènes sont partiellement indépendants. La "niche" de quelques espèces :

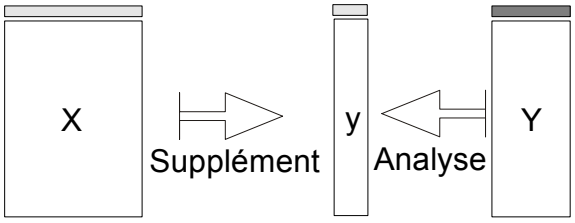



```
par(mfrow=c(3,2))
wnames = c("Truite", "Vairon", "Ombre", "Goujon", "Carpe", "Ablette")
wnum = c(2,3,5,13,19,26)
for (i in 1:6) {
    z = doubs$poi[,wnum[i]]
    nomesp = wnames[i]
    s.traject(pcamil$li,clab=0, cpoi=0, sub = nomesp,
        possub="topright", csub=3)
    s.value (pcamil$li,z, add.plot=T, cleg=1.5)
```

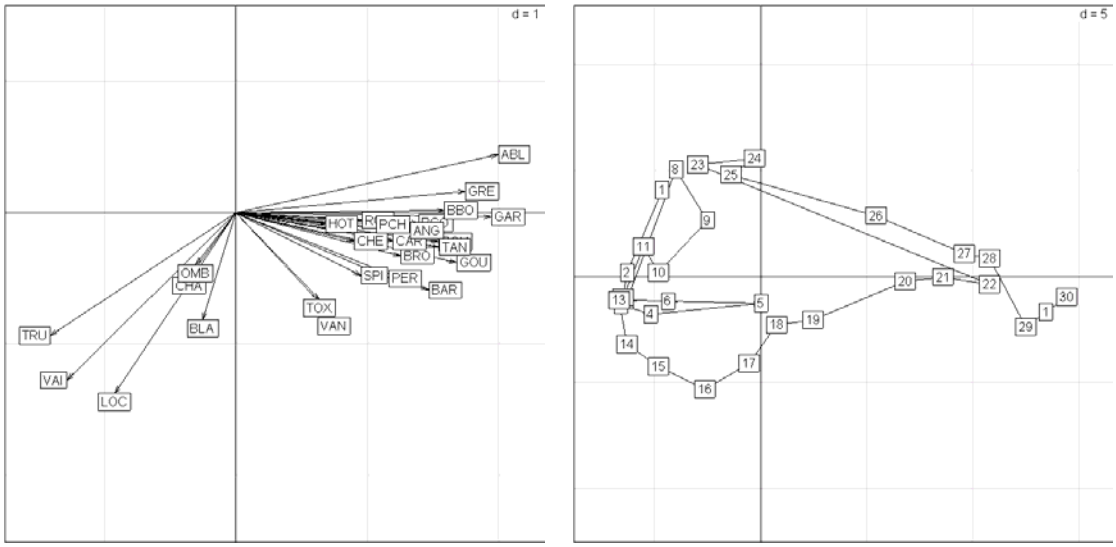
```
    s.distri(pcamil$li, z, add.plot=T, cstar=0)
}
```

Ceci est une représentation d'informations supplémentaires.

Dans la seconde (McIntosh 1958), on analyse le tableau relevés-espèces et on étudie la répartition des variables.



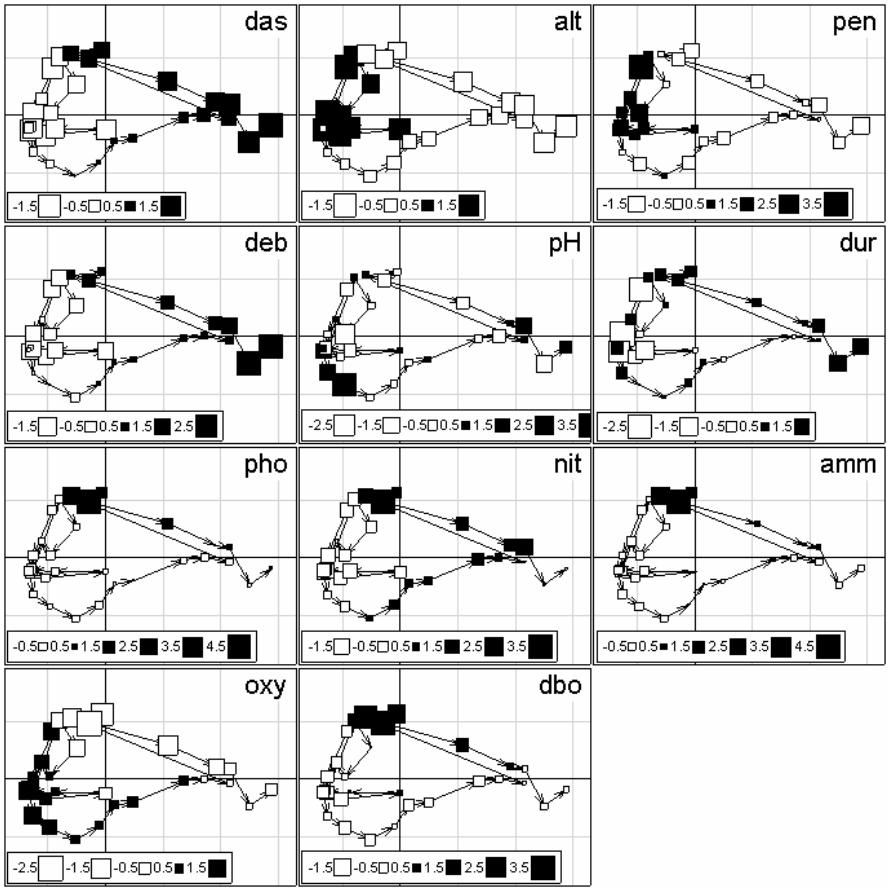

```
pcafau = dudi.pca(doubs$poi, scale = F, scann=F)
s.arrow(pcafau$co)
s.traject(pcafau$li)
s.label(pcafau$li,add.plot=T)
```



A gauche et légèrement en bas l'amont , à droite et légèrement en haut l'aval. En haut une forme repliée liée à l'absence de toutes les espèces en amont comme en aval (facteur taille négatif de pollution) avec une confusion pour la tête de rivière (richesse et abondance faible sans pollution).

L'évolution des variables dans le cadre de la structure faunistique :

```
par(mfrow=c(4,3))
for (i in 1:11) {
    s.traject(pcafau$li,clab=0,    cpoi=0,    sub    =    names(doubs$mil)[i],
possub="topright", csub=3)
    s.value (pcafau$li,pcamil$tab[,i], add.plot=T, cleg=1.5, csi=1.5)
}
```

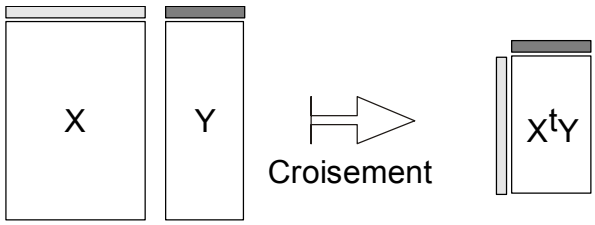


Ceci est une représentation d'informations supplémentaires.

1.3. Croisement

Le tableau croisé utilise simplement un produit de matrices.

Analyse d'un tableau croisé



C'est la base des analyses de co-inertie que nous verrons plus loin. Il suffit que le produit de matrice ainsi défini ait un sens expérimental. Passer le tableau faunistique en pourcentage par espèce et transposer :

```
w = t(apply(doubs$poi,2,function(x) x/sum(x)))
tabcroi = as.data.frame(w%*%as.matrix(pcamil$tab))
```

On obtient la position moyenne de chaque espèce sur chaque variable de milieu normalisée.

```
round(tabcroi,dig=1)
```

	das	alt	pen	deb	pH	dur	pho	nit	amm	oxy	dbo
CHA	-0.2	-0.2	-0.2	-0.1	0.8	0.3	-0.4	-0.3	-0.5	1.0	-0.6
TRU	**-0.7**	**0.6**	**0.5**	**-0.5**	**0.3**	**-0.5**	**-0.5**	**-0.7**	**-0.5**	**0.8**	**-0.6**
VAI	-0.5	0.4	0.3	-0.4	0.1	-0.3	-0.4	-0.5	-0.4	0.6	-0.5
LOC	-0.4	0.3	0.1	-0.4	0.1	-0.3	-0.4	-0.4	-0.4	0.5	-0.4
OMB	-0.1	-0.3	-0.2	0.1	0.6	0.5	-0.4	-0.4	-0.5	1.1	-0.5
BLA	0.0	-0.4	-0.3	0.1	0.8	0.3	-0.3	-0.1	-0.4	0.8	-0.5
HOT	0.9	-0.9	-0.7	0.8	-0.2	0.4	-0.1	0.5	0.0	-0.2	-0.1
TOX	0.6	-0.7	-0.6	0.6	-0.1	0.3	-0.2	0.4	-0.2	0.1	-0.3
VAN	0.3	-0.3	-0.3	0.3	0.1	0.3	-0.1	0.2	-0.1	-0.1	-0.2
CHE	0.4	-0.3	-0.4	0.4	0.0	0.3	0.0	0.2	0.0	-0.2	0.0
BAR	0.8	-0.8	-0.6	0.8	0.0	0.4	-0.1	0.4	-0.1	-0.1	-0.2
SPI	0.9	-0.8	-0.8	1.0	-0.1	0.5	-0.1	0.4	-0.1	0.0	-0.3
GOU	0.7	-0.7	-0.6	0.7	0.0	0.4	0.1	0.4	0.0	-0.2	0.0
BRO	0.6	-0.5	-0.4	0.7	0.0	0.3	0.0	0.3	0.0	-0.3	0.0
PER	0.6	-0.4	-0.5	0.6	-0.1	0.4	-0.2	0.1	-0.2	-0.1	-0.2
BOU	1.1	-0.9	-0.8	1.2	0.0	0.6	0.0	0.5	-0.1	-0.3	-0.2
PSO	1.1	-0.9	-0.7	1.1	-0.1	0.6	0.0	0.6	0.0	-0.4	-0.1
ROT	1.0	-0.8	-0.8	1.1	0.1	0.6	0.2	0.5	0.1	-0.5	0.0
CAR	**1.2**	**-0.9**	**-0.8**	**1.2**	**0.2**	**0.7**	**0.0**	**0.5**	**-0.1**	**-0.3**	**-0.2**
TAN	0.7	-0.6	-0.5	0.7	0.1	0.3	-0.1	0.3	-0.1	-0.3	-0.1
BCO	1.3	-1.0	-0.8	1.3	0.0	0.7	0.0	0.5	-0.1	-0.4	-0.1
PCH	1.5	-1.1	-0.9	1.7	0.1	0.9	0.0	0.4	-0.1	-0.5	-0.1
GRE	1.2	-1.0	-0.8	1.2	0.0	0.6	0.2	0.7	0.1	-0.6	0.2
GAR	0.7	-0.6	-0.6	0.6	-0.1	0.4	0.1	0.4	0.1	-0.4	0.1
BBO	1.3	-1.0	-0.8	1.2	0.1	0.6	0.0	0.6	0.0	-0.5	0.0
ABL	1.0	-0.9	-0.8	0.9	-0.1	0.5	0.4	0.8	0.4	-0.6	0.4
ANG	1.2	-1.0	-0.8	1.3	0.1	0.7	0.0	0.6	-0.1	-0.4	-0.1

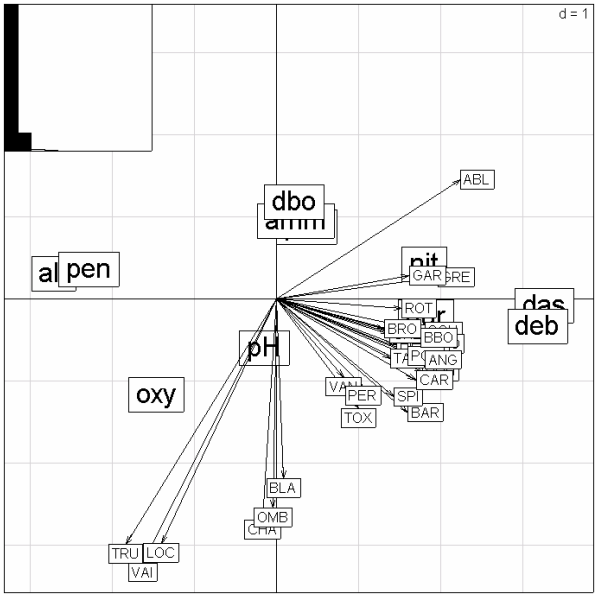
On peut faire l'ACP non centrée de ce tableau ou encore :

```
round(100*cor(doubs$poi,doubs$mil),dig=0)
    das alt pen deb  pH dur pho nit amm oxy dbo
CHA -11 -12  -9  -3  43  18 -20 -15 -26  55 -31
TRU -66  58  48 -51  24 -47 -44 -63 -44  75 -53
VAI -63  48  32 -50  16 -34 -48 -56 -47  74 -57
LOC -53  41  19 -47  16 -37 -46 -46 -46  61 -56
OMB  -6 -13  -9   7  32  26 -19 -18 -24  53 -26
BLA   0 -18 -16   3  37  17 -15  -3 -20  38 -27
HOT  62 -61 -47  56 -12  28  -4  34  -3 -16  -5
TOX  40 -49 -38  37  -9  23 -14  29 -10   8 -23
VAN  32 -33 -30  32   6  26  -9  17 -10  -5 -15
CHE  53 -47 -51  49   3  44  -2  28  -4 -28  -3
BAR  69 -67 -51  66   4  36 -10  37 -11  -6 -18
SPI  57 -54 -52  64  -8  36  -7  28  -7  -3 -17
GOU  75 -67 -56  72   1  38   8  44   3 -24  -3
BRO  57 -40 -39  61  -3  30  -1  24  -3 -26  -4
PER  45 -35 -38  51  -6  28 -13  10 -14 -11 -16
BOU  77 -64 -53  80  -2  41  -2  36  -6 -19 -10
PSO  79 -66 -51  79  -5  43   0  39  -3 -26  -4
ROT  63 -47 -47  68   6  38  10  29   6 -29   2
CAR  74 -60 -50  77  10  41  -3  32  -7 -19 -10
TAN  61 -49 -46  58   5  27  -7  29 -10 -25  -9
BCO  74 -57 -44  76   1  38  -1  27  -7 -25  -4
PCH  71 -50 -43  80   6  42   1  21  -6 -24  -3
GRE  84 -67 -53  81  -2  43  14  45   8 -41  11
GAR  64 -54 -56  58  -6  37   7  40   6 -41   7
BBO  77 -61 -48  76   6  39   1  36  -3 -30  -1
ABL  89 -79 -65  79 -10  46  33  65  31 -53  31
ANG  78 -61 -53  82   3  43   1  36  -4 -24  -7
```

Fondamentalement, dans l'analyse du tableau croisé, les lignes sont les variables d'un tableau et les colonnes sont les variables de l'autre.

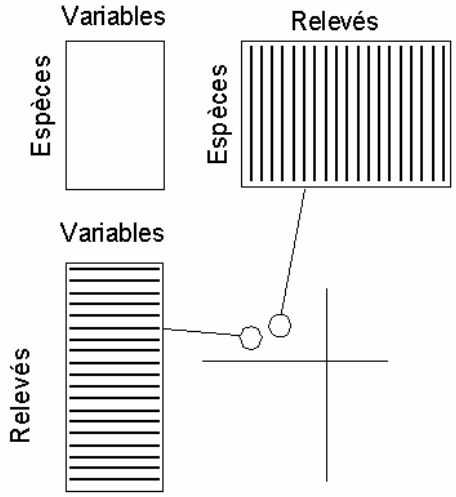
```
scatter(dudi.pca(cor(doubs$mil,doubs$poi), scal=F, cent=F, scan=F),clab.r=2)
```

On obtient un biplot qui donne deux typologies de variables et leur lien :



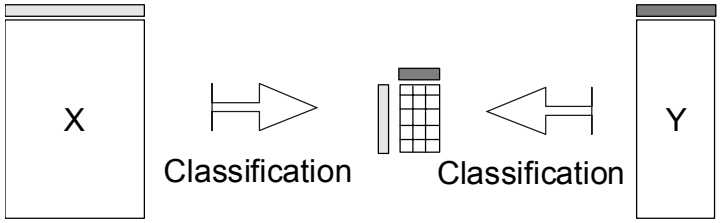
A gauche l'amont, à droite l'aval, en haut la pollution, en amont sans pollution Truite-Loche-Vairon, plus bas sans pollution Ombre-Chabot-Blageon, en bas larichesse augmente et la tolérance grandit (Cyprinidés).

On a alors trois tableaux et on comprend que les relevés sont supplémentaires de deux manières dans l'analyse du tableau croisé :

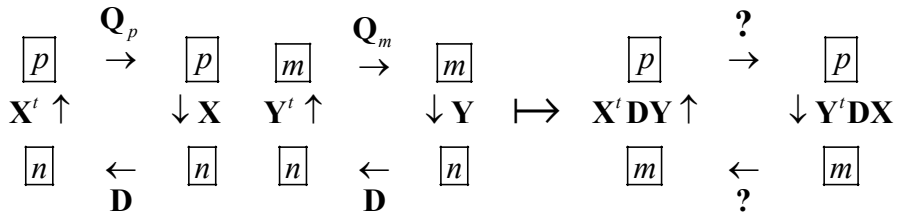


Le cas le plus célèbre d'un croisement de tableau signifiant est celui des profils écologiques (Godron et al. 1968, Gounot 1969). Les variables de milieu sont toutes qualitatives et les variables floristiques sont toutes binaires (0 absence, 1 présence). Le tableau **X** est formé des indicatrices des classes de chaque variable de milieu. Le tableau **Y** est formé des indicatrices de présence de chaque espèce. Le tableau croisé $\mathbf{Y}'\mathbf{X}$ a pour lignes les espèces et pour colonnes les modalités de milieu. Les cases contiennent le nombre de stations de chaque modalité de milieu contenant l'espèce. Sur une ligne on trouve une juxtaposition de profils écologiques bruts. P. Romane (1972) a eu l'idée d'envoyer un tel tableau dans l'analyse des correspondances, idée qui sera retrouvée plus tard (Montaña and Greig-Smith 1990). On obtient un cas particulier de l'analyse de co-inertie (Mercier et al. 1992).

Parmi nombre de pratiques basées sur les tableaux croisés, on trouve le croisement de partitions :



Le tableau croisé joue un rôle central dans les méthodes de couplage. Il faut, par nécessité théorique, utiliser le produit scalaire commun :



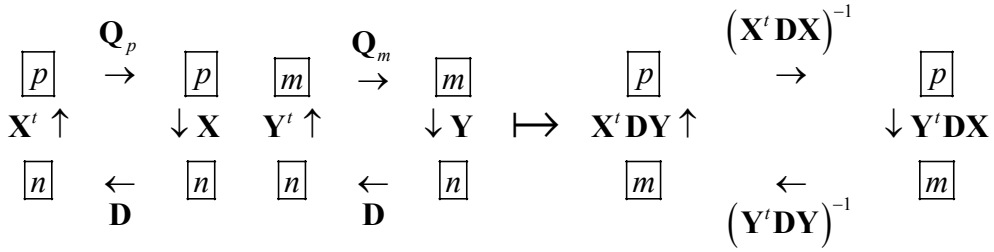
Le choix des métriques associées au croisement va définir les principales familles de méthodes.

2. Analyses canoniques

2.1. Analyse canonique des corrélations

L'analyse canonique des corrélations (Hotelling 1936) est la plus ancienne et la plus connue des méthodes de couplage introduite en détail pour l'écologie par Gittins(1985). Le fondement considère deux ACP normées. On appellera simplement **X** et **Y** les deux tableaux normalisés (moyennes nulles et variances unitaires par colonnes).

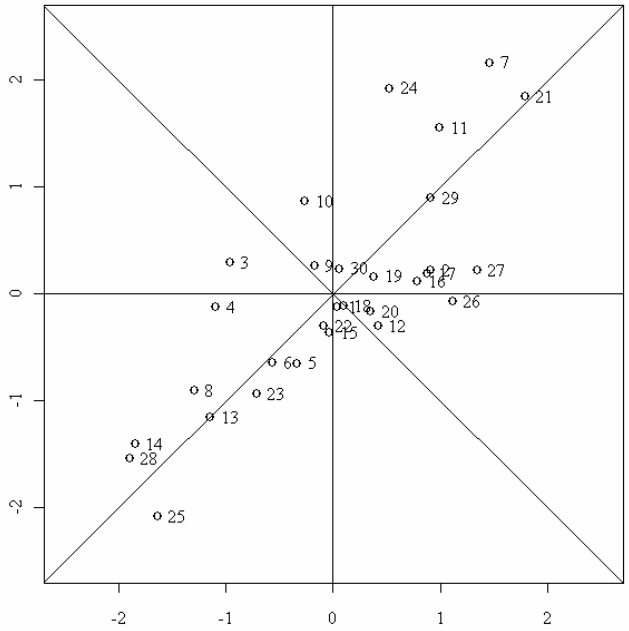
On suppose que les deux paquets de variables (colonnes de **X** et **Y**) sont sans redondances (la régression d'une variables de **Y** sur **X** ou d'une variable de **X** sur **Y** est définie sans problème) donc que les matrices de corrélations des deux paquets sont inversibles. L'analyse canonique est celle du schéma :



Les matrices $\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}$ et $\left(\mathbf{Y}^t\mathbf{D}\mathbf{Y}\right)^{-1}$ sont symétriques, positives et inversibles donc des matrices de produits scalaires. La géométrie qu'elles définissent est très particulière.

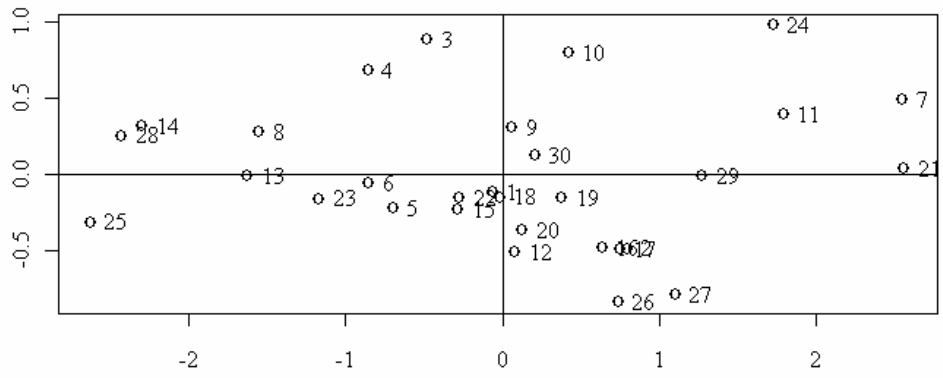
```
> cor <- matrix(c(1,0.8,0.8,1),nrow=2)
```

```
> cor
     [,1] [,2]
[1,]  1.0  0.8
[2,]  0.8  1.0
> X <- mvrnorm(30,c(0,0),cor)
> X <- scale(X)
> cor(X)
       [,1]   [,2]
[1,] 1.0000 0.7809
[2,] 0.7809 1.0000

f1 <- function () {
   par(mar=rep(4,4))
   plot(X,xlim=c(-2.5,2.5),ylim=c(-2.5,2.5))
   abline(h=0)
   abline(v=0)
   abline(0,1)
   abline(0,-1)
   text(X,as.character(1:30),pos=4)
}
```



Avec le produit scalaire canonique, le carré de la distance entre les points 7 et 10 vaut :

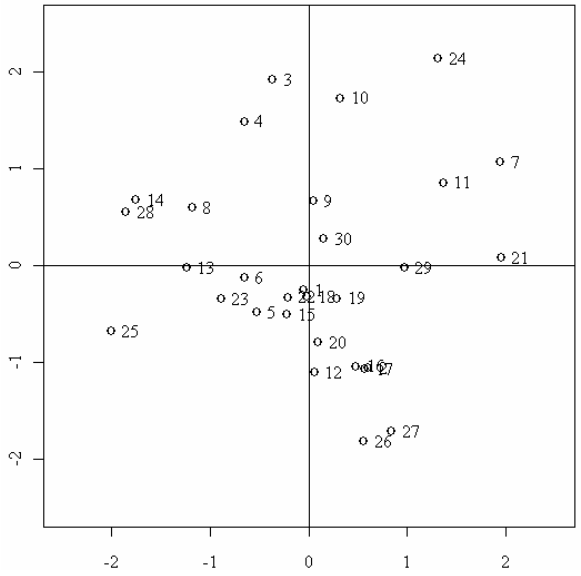
```
> sum((X[7,]-X[10,])^2)
[1] 4.643
> sum((X[7,]-X[21,])^2)
[1] 0.2102

> pr0 <- princomp(X)
> plot(-pr0$scores)
> text(-pr0$scores,as.character(1:30),pos=4)
> abline(h=0)
> abline(v=0)
```



En se mettant dans la base des axes principaux sans changer de métrique, on conserve les distances :

```
> Y <- -pr0$scores
> sum((Y[7,]-Y[10,])^2)
[1] 4.643
> sum((Y[7,]-Y[21,])^2)
[1] 0.2102
> Z <- Y
> Z[,1] <- Z[,1]/pr0$sdev[1]
> Z[,2] <- Z[,2]/pr0$sdev[2]
> plot(Z,xlim=c(-2.5,2.5),ylim=c(-2.5,2.5))
> text(Z,as.character(1:30),pos=4)
> abline(h=0)
> abline(v=0)
```

La variabilité sur chacun des axes principaux a été ramenée volontairement à 1 en divisant par la racine de la valeur propre (valeur singulière) qui est l'écart-type de la coordonnée :



Les distances sont changées profondément :

```
> sum((Z[7,]-Z[10,])^2)
[1] 3.076
> sum((Z[7,]-Z[21,])^2)
[1] 0.9924
```

On ne tient plus compte de la corrélation dans le calcul de la distance. L'opération consiste à prendre le produit scalaire de matrice Λ^{-1} dans la base des axes principaux donc $\mathbf{A}\Lambda^{-1}\mathbf{A}^t$ dans la base canonique. Or :

$$\mathbf{X}^t\mathbf{D}\mathbf{X} = \mathbf{A}\Lambda\mathbf{A}^t \Rightarrow \left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1} = \mathbf{A}\Lambda^{-1}\mathbf{A}^t$$

Les facteurs principaux du schéma :

$$\begin{array}{ccc} & \left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1} & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{X}^t\mathbf{D}\mathbf{Y} \uparrow & & \downarrow \mathbf{Y}^t\mathbf{D}\mathbf{X} \\ \boxed{m} & \leftarrow & \boxed{m} \\ & \left(\mathbf{Y}^t\mathbf{D}\mathbf{Y}\right)^{-1} & \end{array}$$

sont $\left(\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\right)^{-1} = \mathbf{X}^t\mathbf{DX}$ normés donc vérifient $\mathbf{a}^{*t}\,\mathbf{X}^t\mathbf{DXa}^* = \left\|\mathbf{Xa}^*\right\|_{\mathbf{D}}^2 = 1$. Ce sont des coefficients de combinaisons linéaires de variables de variance unité.

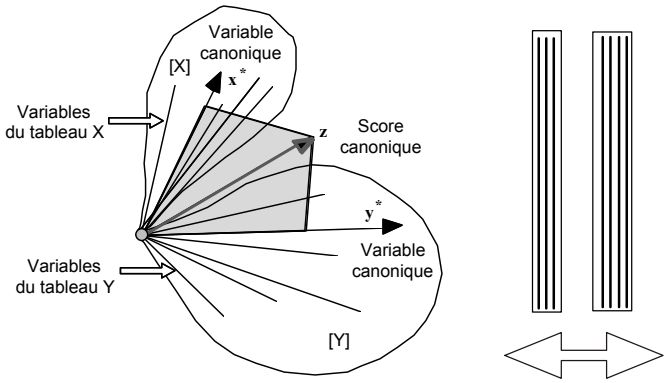
Les cofacteurs de ce schéma sont $\left(\left(\mathbf{Y}^t\mathbf{DY}\right)^{-1}\right)^{-1} = \mathbf{Y}^t\mathbf{DY}$ normés donc vérifient $\mathbf{b}^{*t}\,\mathbf{Y}^t\mathbf{DYb}^* = \left\|\mathbf{Yb}^*\right\|_{\mathbf{D}}^2 = 1$. Ce sont des coefficients de combinaisons linéaires de variables de variance unité.

Les théorèmes généraux indiquent que le premier facteur et le premier cofacteur optimise la quantité :

$$\left\langle \left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\mathbf{X}^t\mathbf{DYb}^* \middle| \mathbf{a}^* \right\rangle_{\left(\mathbf{X}^t\mathbf{DX}\right)} = \mathbf{a}^{*t}\left(\mathbf{X}^t\mathbf{DX}\right)\left(\mathbf{X}^t\mathbf{DX}\right)^{-1}\mathbf{X}^t\mathbf{DYb}^*$$
$$= \mathbf{a}^{*t}\,\mathbf{X}^t\mathbf{DYb}^* = \left\langle \mathbf{Xa}^* \middle| \mathbf{Yb}^* \right\rangle_{\mathbf{D}}$$

On obtient donc une combinaison de variables du tableau **X** de variance 1 (le vecteur $\mathbf{Xa}^*$) et une combinaison de variables du tableau **Y** de variance 1 (le vecteur $\mathbf{Yb}^*$) de corrélation maximale (la quantité $\left\langle \mathbf{Xa}^* \middle| \mathbf{Yb}^* \right\rangle_{\mathbf{D}}$). $\mathbf{Xa}^*$ et $\mathbf{Yb}^*$ sont appelées variables canoniques. L'optimum de la corrélation qu'on peut faire avec une combinaison linéaire de chaque tableau est la corrélation canonique (racine carrée de la première valeur propre). La première valeur propre prend donc le nom de carré de corrélation canonique entre les deux tableaux.

L'analyse canonique prend sa signification dans l'espace des variables. On peut la résumer par la figure :



La variable canonique du tableau **X** est la combinaison des variables de **X** la mieux prédite par une régression multiple sur les variables de **Y** tout comme la variable canonique du tableau **Y** est la combinaison des variables de **Y** la mieux prédite par une régression multiple sur les variables de **X**. On appelle score canonique la bissectrice des deux variables canoniques (somme normalisée). Il est capital de voir dans l'analyse canonique toutes les contraintes associées à deux régressions multiples. Le nombre de prédicteurs doit être forcément limité par rapport au nombre de variables.

C'est pourquoi, sur les couplages de tableaux écologiques, elle a mauvaise presse. On ne peut l'utiliser que si le nombre de variables est faible par rapport au nombre d'individus.

```
> cancor(pcamil$tab,pcafau$tab)$cor
 [1] 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 0.9052
[11] 0.7643
```

Ce n'est pas la peine de continuer. Il existe une multitude de combinaisons linéaires des variables faunistiques parfaitement corrélées à des combinaisons linéaires de variables de milieu, mais cela ne peut rien nous apprendre. Par contre, on l'utilise avec un tableau et les coordonnées de l'analyse de l'autre ou avec deux ensembles de coordonnées (Barkham and Norris 1970) :

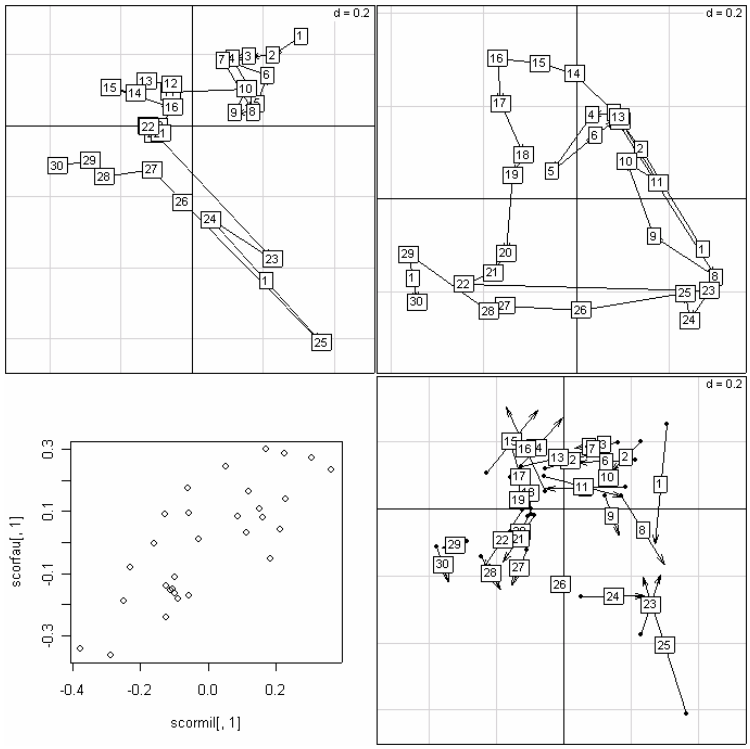
```
cancor(pcamil$l1,pcafau$l1)$cor
[1] 0.7928 0.6577
```

L'analyse canonique ignore la notion de variance :

```
cancor(pcamil$li,pcafau$li)$cor
[1] 0.7928 0.6577

cancor(50*pcamil$li,0.5*pcafau$li)$cor
[1] 0.7928 0.6577
```

On garde ici deux variables canoniques pour deux variables (coordonnées) dans chaque tableau.

```
par(mfrow=c(2,2))
s.traject(scormil); s.label(scormil,add.p=T)
s.traject(scorfau); s.label(scorfau,add.p=T)
plot(scormil[,1],scorfau[,1])
s.match(scormil,scorfau)
```




```
var(scormil)
              [,1]          [,2]
[1,]  3.448e-02 -2.393e-18
[2,] -2.393e-18  3.448e-02 # utile pour comprendre 1/29 = 0.03448

var(scorfau)
              [,1]          [,2]
[1,]  3.448e-02 -3.350e-18
[2,] -3.350e-18  3.448e-02
```

On peut évidemment utiliser ces codes comme fond de carte pour toute expression directe des données.

2.2. Analyse canonique de deux sous-espaces

On comprend que la géométrie associée aux deux ensembles de variables centrées et réduites s'étend à deux ensembles quelconques de vecteurs. L'opération se réduit à chercher des combinaisons de variables faisant des angles les plus petits possibles (sous contrainte d'orthogonalité successive dans chaque paquet).

La documentation de la fonction de R résume parfaitement cette approche :

```
cancor                package:mva            R Documentation

Canonical Correlations

Description:

     Compute the canonical correlations between two data matrices.

Usage:

     cancor(x, y, xcenter = TRUE, ycenter = TRUE)

Arguments:

       x: numeric matrix (n * p1), containing the x coordinates.

       y: numeric matrix (n * p2), containing the y coordinates.

 xcenter: logical or numeric vector of length p1, describing any
          centering to be done on the x values before the analysis.  If
          `TRUE' (default), subtract the column means. If `FALSE', do
          not adjust the columns.  Otherwise, a vector of values to be
          subtracted from the columns.

 ycenter: analogous to `xcenter', but for the y values.

Details:

     The canonical correlation analysis seeks linear combinations of
     the `y' variables which are well explained by linear combinations
     of the `x' variables. The relationship is symmetric as `well
     explained' is measured by correlations.

Value:

     A list containing the following components:

     cor: correlations.

   xcoef: estimated coefficients for the `x' variables.

   ycoef: estimated coefficients for the `y' variables.

 xcenter: the values used to adjust the `x' variables.

 ycenter: the values used to adjust the `x' variables.

References:

     Hotelling H. (1936). Relations between two sets of variables.
     Biometrika, 28, 321-327.

     Seber, G. A. F. (1984). Multivariate Analysis. New York: Wiley, p.
     506f.
```

Remarque : quand on cherche des références bibliographiques sur une méthode le plus sérieux est de prendre celles de la documentation de R. Elles sont toujours incontournables.

2.3. L'AFC est une analyse canonique

La procédure de R permet d'illustrer le fait que l'AFC est une analyse canonique.

```
> fauv
  V1 V2 V3 V4
1  2  2  1  0
2  2  2  1  0
3  3  2  2  0
4  2  2  1  0
5  2  2  2  0
6  2  3  3  3
7  1  3  2  3

> as.vector(unlist(fauv))
 [1] 2 2 3 2 2 2 1 2 2 2 2 2 3 3 1 1 2 1 2 3 2 0 0 0 0 0 3 3
> fau.vec<-as.vector(unlist(fauv))
> as.vector(row(as.matrix(fauv)))
 [1] 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7
> nlig.vec<-as.vector(row(as.matrix(fauv)))
> as.vector(col(as.matrix(fauv)))
 [1] 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 3 4 4 4 4 4 4 4
> ncol.vec<-as.vector(col(as.matrix(fauv)))

X.fau<-matrix(0, nrow=48, ncol=7)
Y.fau<-matrix(0, nrow=48, ncol=4)

j<-0
for (i in 1:28) {
   if (fau.vec[i]>0) {
      for (k in 1:fau.vec[i]) {
         j<-j+1
         X.fau[j,nlig.vec[i]]<-1
         Y.fau[j,ncol.vec[i]]<-1
      }
   }
}
> X.fau                       > Y.fau
      [,1] [,2] [,3] [,4]         [,1] [,2] [,3] [,4] [,5] [,6] [,7]
 [1,]    1    0    0    0         1    0    0    0    0    0    0
 [2,]    1    0    0    0         1    0    0    0    0    0    0
 [3,]    1    0    0    0         0    1    0    0    0    0    0
 [4,]    1    0    0    0         0    1    0    0    0    0    0
 [5,]    1    0    0    0         0    0    1    0    0    0    0
 [6,]    1    0    0    0         0    0    1    0    0    0    0
 [7,]    1    0    0    0         0    0    1    0    0    0    0
 [8,]    1    0    0    0         0    0    0    1    0    0    0
 [9,]    1    0    0    0         0    0    0    1    0    0    0
[10,]    1    0    0    0         0    0    0    0    1    0    0
[11,]    1    0    0    0         0    0    0    0    1    0    0
[12,]    1    0    0    0         0    0    0    0    0    1    0
[13,]    1    0    0    0         0    0    0    0    0    1    0
[14,]    1    0    0    0         0    0    0    0    0    0    1
[15,]    0    1    0    0         1    0    0    0    0    0    0
[16,]    0    1    0    0         1    0    0    0    0    0    0
...
[41,]    0    0    1    0         0    0    0    0    0    0    1
[42,]    0    0    1    0         0    0    0    0    0    0    1
[43,]    0    0    0    1         0    0    0    0    0    1    0
[44,]    0    0    0    1         0    0    0    0    0    1    0
[45,]    0    0    0    1         0    0    0    0    0    1    0
[46,]    0    0    0    1         0    0    0    0    0    0    1
[47,]    0    0    0    1         0    0    0    0    0    0    1
[48,]    0    0    0    1         0    0    0    0    0    0    1
```

```
> cano.fau<-cancor(X.fau,Y.fau,xcenter=F,ycenter=F)

> cano.fau
$cor:
[1] 1.00000 0.48086 0.11367 0.05559

> cano.fau$cor^2
[1] 1.000000 0.231225 0.012921 0.003091

-----------------------
Total inertia: 0.247236
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.3122E-01 +0.9352 +0.9352   |02   +1.2921E-02 +0.0523 +0.9875 |
03   +3.0906E-03 +0.0125 +1.0000   |04   +0.0000E+00 +0.0000 +1.0000 |

> sqrt(48)*cano.fau$xcoef
     [,1]    [,2]    [,3]     [,4]         [,5]         [,6]         [,7]
[1,]    1  0.8629  1.1259  0.03902  2.120e+000  9.646e-001  1.078e+000
[2,]    1  0.8629  1.1259  0.03902  3.157e-001 -1.741e+000 -1.859e+000
[3,]    1  0.8783 -1.1166  1.47126 -3.632e-001  1.017e+000 -7.133e-001
[4,]    1  0.8629  1.1259  0.03902 -2.072e+000 -2.408e-001  1.495e+000
[5,]    1  0.7214 -1.3666 -2.14752 -1.649e-015 -3.925e-016 -7.850e-017
[6,]    1 -0.9812 -0.6509  0.55022  3.632e-001 -1.017e+000  7.133e-001
[7,]    1 -1.4030  0.6985 -0.45015 -3.632e-001  1.017e+000 -7.133e-001
-------------------------------------------------
Binary input file: D:\…\fauv.fcl1 - 7 rows, 4 cols.
   1 |  0.8629 -1.1259  0.0390     Inf
   2 |  0.8629 -1.1259  0.0390     Inf
   3 |  0.8783  1.1166  1.4713     Inf
   4 |  0.8629 -1.1259  0.0390     Inf
   5 |  0.7214  1.3666 -2.1475     Inf
   6 | -0.9812  0.6509  0.5502    -Inf
   7 | -1.4030 -0.6985 -0.4502     Inf

> sqrt(48)*cano.fau$ycoef
     [,1]      [,2]     [,3]    [,4]
[1,]    1  0.874910  0.04364  1.2889
[2,]    1  0.159135  1.06232 -0.9199
[3,]    1  0.006676 -1.57200 -0.7272
[4,]    1 -2.479171  0.20931  0.8999
-------------------------------------------------
Binary input file: D:\…\fauv.fcc1 - 4 rows, 4 cols.
   1 |  0.8749 -0.0436  1.2889    -Inf
   2 |  0.1591 -1.0623 -0.9199    -Inf
   3 |  0.0067  1.5720 -0.7272    -Inf
   4 | -2.4792 -0.2093  0.8999    -Inf
```

Pour comprendre le lien, il suffit de voir que si **X** est le paquet des indicatrices des colonnes et si **Y** est le paquet des indicatrices des lignes les deux schémas suivants sont strictement identiques :

$$\begin{array}{ccc} \boxed{p} & \xrightarrow{\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}} & \boxed{p} \\ \mathbf{X}^t\mathbf{D}\mathbf{Y}\uparrow & & \downarrow\mathbf{Y}^t\mathbf{D}\mathbf{X} \\ \boxed{m} & \xleftarrow{\left(\mathbf{Y}^t\mathbf{D}\mathbf{Y}\right)^{-1}} & \boxed{m} \end{array} \qquad \begin{array}{ccc} \boxed{J} & \xrightarrow{\mathbf{D}_J^{-1}} & \boxed{J} \\ \mathbf{P}^t\uparrow & & \downarrow\mathbf{P} \\ \boxed{I} & \xleftarrow{\mathbf{D}_I^{-1}} & \boxed{I} \end{array}$$

Le vecteur $\mathbf{1}_n$ est dans chacun des sous-espaces engendrés, d'où la corrélation canonique de 1 dans celle de R et la valeur propre de 0 dans celle de ADE-4. On trouve dans Estève (1978) l'affirmation de l'importance fondamentale de l'AFC comme analyse canonique.

L'opération est basée sur une vision du tableau de données très parlante au plan expérimental :

	Chabot	Truite	Vairon	Loche	Ombre	Blageon	Hotu	Toxostome	Vandoise	Chevaine	Barbeau	Spirlin	Goujon	Brochet	Perche	Bouvière	Perche_soleil	Rotengle	Carpe	Tanche	Brème_comm.	Poisson_chat	Grémille	Gardon	Brème_bord.	Ablette	Anguille
1		3																									
2		5	4	3																							
3		5	5	5										1													
4		4	5	5						1			1	2	2					1							
5		2	3	2					5	2			2	4	4			2		3				5			
6		3	4	5					1	2			1	1	1					2				1			
7		5	4	5					1	1																	
8																											
9			1	3						5										1				4			
10		1	4	4					2	2			1														
11	1	3	4	1	1					1																	
12	2	5	4	4	2					1																	
13	2	5	5	2	3	2																					
14	3	5	5	4	4	3				1	1		1	1													
15	3	4	4	5	2	4			3	3	2		2							1							
16	2	3	3	5		5		4	5	2	2	1	2	1	1		1		1	1				1			
17	1	2	4	4	1	2	1	4	3	2	3	4	1	1	2	1	1		1	1				2		2	1
18	1	1	3	3	1	1	1	3	2	3	3	3	2	1	3	2	1		1	1			1	2		2	1
19			3	5		1	2	3	2	1	2	2	4	1	1	2	1	1	1	2	1		1	5	1	3	1
20			1	2			2	2	2	3	4	3	4	2	2	3	2	2	1	4	1		2	5	2	5	2
21			1	1			2	2	2	2	4	2	5	3	3	3	2	2	2	4	3	1	3	5	3	5	2
22				1			3	2	3	4	5	1	5	3	4	3	3	2	3	4	4	2	4	5	4	5	2
23										1														1		2	
24							1			2			1				1						2	2	1	5	
25									1	1			2	1				1					1	1		3	
26				1			1		1	2	2	1	3	2	1	2	2	1	1	3	2	1	4	4	2	5	2
27				1			1	1	2	3	4	1	4	4	1	3	3	1	2	5	3	2	5	5	4	5	3
28				1			1	1	2	4	3	1	4	3	2	4	4	2	4	4	3	3	5	5	5	5	4
29		1	1	1	1	1	2	2	3	4	5	3	5	5	4	5	5	2	3	3	4	4	5	5	4	5	4
30							1	2	3	3	3	5	5	4	5	5	3	5	5	5	5	5	5	5	5	5	5

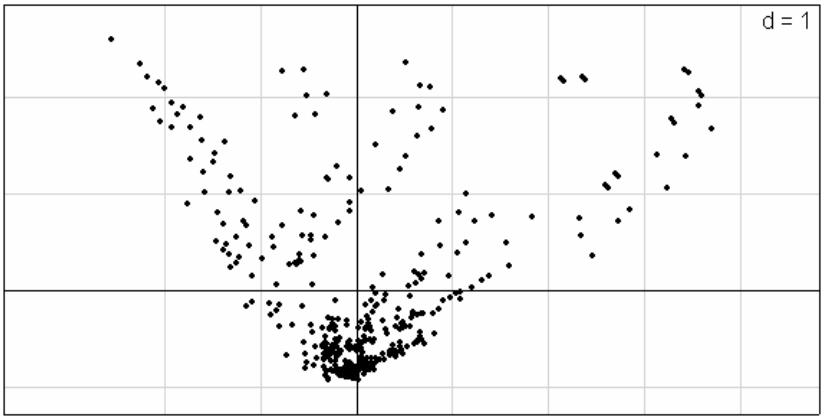
Jusqu'à présent, on a vu dans chaque espèce une variable qui prenait la valeur 0 pour une absence. On peut considérer que les absences n'ont aucune signification, soit que l'échantillonnage soit insuffisant pour détecter certaines des espèces, soit que des espèces pourraient se trouver dans certaines stations mais n'y sont pas pour des contingences historiques. Le raisonnement est dans Green (1971). Ne considérons donc que les présences. Il y a dans ce tableau 375 cases non vides, c'est-à-dire 375 occurrences d'une espèce dans un relevé. Par exemple, on a une occurrence Vairon dans la station 6. Cette occurrence est caractérisée par le nom de l'espèce *Vairon*, le nom du relevé *6* et son poids *4* ou plutôt 4/1004 (1004 est la somme de toutes les abondances). Les occurrences sont les nouveaux individus statistiques. On peut définir deux paquets de variables. Le premier compte 30 variables qui prennent la valeur 1 si l'occurrence est dans le relevé et 0 sinon. Le second compte 27 variables qui prennent la valeur 1 si l'occurrence est de l'espèce et 0 sinon. On a deux tableaux à 375 lignes et respectivement 30 et 27 variables. Faire l'analyse canonique de ces deux tableaux, c'est faire l'analyse des correspondances du tableau faunistique. Source historique dans Fisher (1940) .

```
> fau.coa = dudi.coa(doubs$poi,scan=F)
> fau.reci = reciprocal.coa(fau.coa)
> dim(fau.reci)
[1] 375   5
> names(fau.reci)
[1] "score1" "score2" "row"    "col"    "wei"
> fau.reci[c(1:3,20:22,225:227),]
       score1  score2 row col      wei
10CHA  2.1658  2.1678  10 CHA 0.000996
11CHA  2.1812  2.3857  11 CHA 0.001992
12CHA  2.2530  3.4588  12 CHA 0.001992
13TRU  2.1307  0.6583  13 TRU 0.004980
14TRU  1.8522  0.3835  14 TRU 0.003984
15TRU  1.5102  0.2000  15 TRU 0.002988
21PER -0.5797 -0.2006  21 PER 0.003984
```

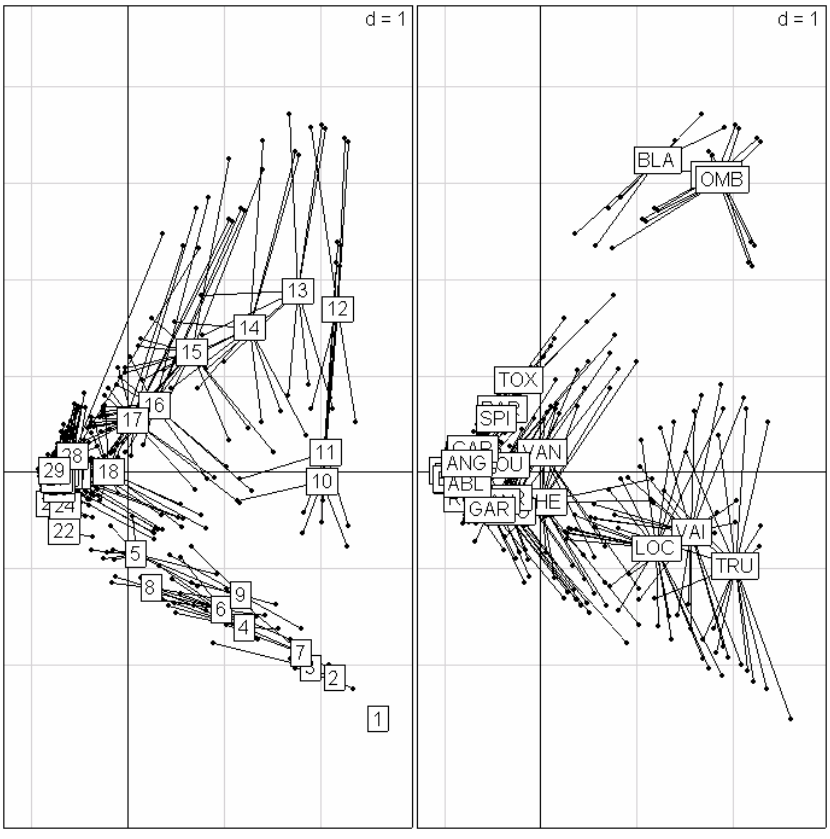
```
25PER -0.6003 -0.3183  25 PER 0.000996
26PER -0.5962 -0.2525  26 PER 0.000996
```

Chaque présence d'une espèce (une correspondance) a un score, un poids, un indice ligne et un indice colonne. On a une carte des correspondances (scores de l'analyse canonique) :

```
> s.label(fau.reci,2,1,clab=0)
```




```
par(mfrow=c(1,2))
s.class(fau.reci,fau.reci$row,fau.reci$wei,cell=0)
s.class(fau.reci,fau.reci$col,fau.reci$wei,cell=0)
```



Utilisée dans cette optique, l'analyse des correspondances a des propriétés uniques de définition réciproque de la diversité des relevés (à gauche) et de l'amplitude d'habitat des espèces (à droite) (Thioulouse and Chessel 1992).

2.4. Analyse discriminante

L'analyse discriminante étudie le lien entre un tableau et une partition des individus. C'est un problème de couplage de tableaux exactement comme précédemment.

Numéro	Qualitative		Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
1	1		1	0	0	0	0
2	3		0	0	1	0	0
3	2		0	1	0	0	0
4	2		0	1	0	0	0
5	2		0	1	0	0	0
6	3		0	0	1	0	0
7	3		0	0	1	0	0
8	3		0	0	1	0	0
9	1		1	0	0	0	0
10	1		1	0	0	0	0
11	4		0	0	0	1	0
12	4		0	0	0	1	0
13	4		0	0	0	1	0
14	5		0	0	0	0	1
15	5		0	0	0	0	1
16	1		1	0	0	0	0

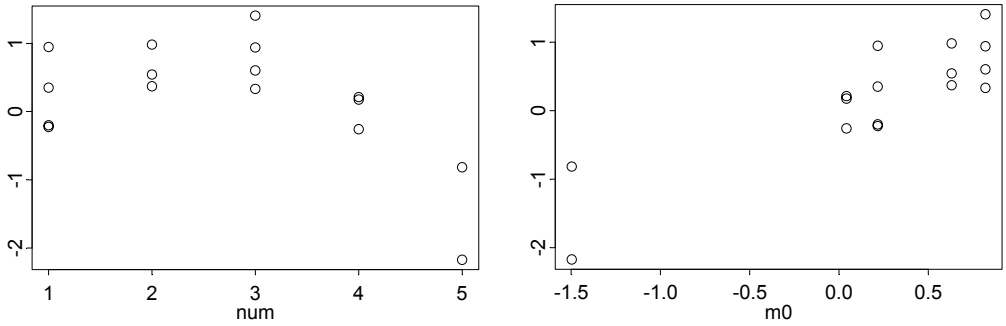
Une variable qualitative est toujours un ensemble d'indicatrices de classe. Une variable qualitative à 5 modalités est un ensemble de 5 variables quantitatives binaires (en fait 4, car la cinquième est entièrement définie par les autres).

```
> num
 [1] 1 3 2 2 2 3 3 3 1 1 4 4 4 5 5 1
> tabnum
   X1 X2 X3 X4 X5
 1  1  0  0  0  0
 2  0  0  1  0  0
 3  0  1  0  0  0
 4  0  1  0  0  0
 5  0  1  0  0  0
 6  0  0  1  0  0
 7  0  0  1  0  0
 8  0  0  1  0  0
 9  1  0  0  0  0
10  1  0  0  0  0
11  0  0  0  1  0
12  0  0  0  1  0
13  0  0  0  1  0
14  0  0  0  0  1
15  0  0  0  0  1
16  1  0  0  0  0
```

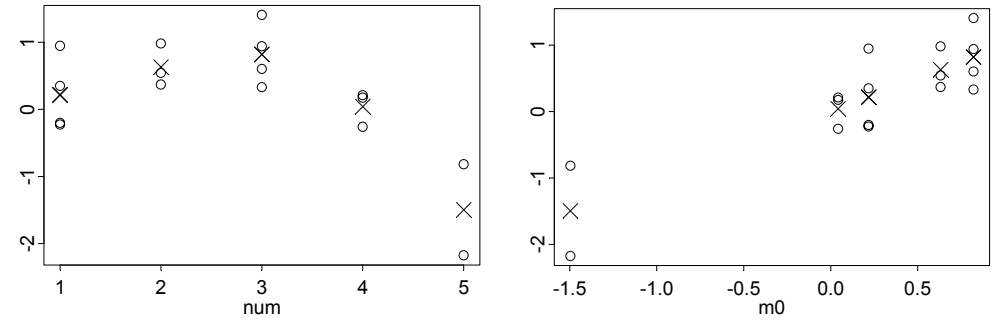
Le minimum nécessaire à la manipulation du lien entre une variable qualitative et une variable quantitative tient dans le rapport de corrélation.

```
> y
 [1]  0.3504  0.9401  0.3695  0.5425  0.9813  1.4071  0.3313  0.6016  0.9463
[10] -0.2008 -0.2587  0.2105  0.1749 -0.8171 -2.1751 -0.2260
> m0 <- predict(lm(y~as.factor(num)))
```

Comment mesurer le lien entre la variable qualitative (num) et la variable quantitative **y** ?



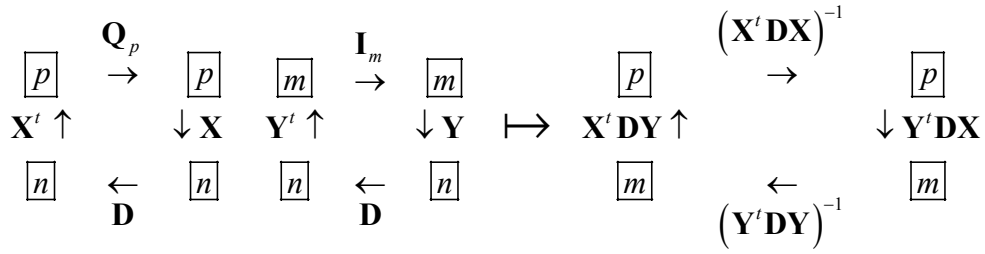
A gauche, les valeurs de **y** en fonction du numéro de classe, à droite les valeurs de **y** en fonction de la moyenne de la classe correspondante. On rajoute les moyennes par classe :



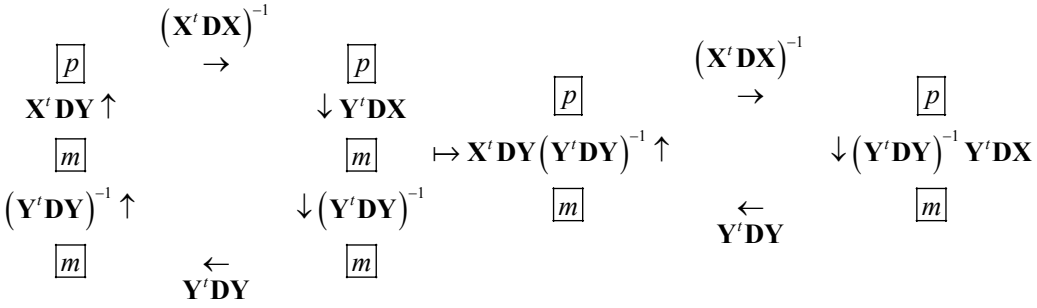

```
> tapply(y,num,mean)
      1      2    3       4      5
 0.2175 0.6311 0.82 0.04226 -1.496   Les moyennes par classe
> tapply(y,num,var,un=F)
      1      2      3      4     5
 0.2301 0.0663 0.1614 0.0455 0.461   Les variances par classe
> var(y,un=F)
[1] 0.6717                          La variance totale
> var(m0,un=F)
[1] 0.4953                 La variance des moyennes par classe (variance Inter)
> tapply(y,num,var,un=F)
      1      2      3      4     5
 0.2301 0.0663 0.1614 0.0455 0.461
> table(num)                 Le nombre d'individus par classe
 1 2 3 4 5
 4 3 4 3 2
> sum(tapply(y,num,var,un=F)*table(num))/16
[1] 0.1765                 La moyenne des variances par classe (variance Intra)
> sum(tapply(y,num,var,un=F)*table(num))/16+var(m0,un=F)
[1] 0.6717
> var(y,un=F)
[1] 0.6717                 La variance totale = Inter + Intra

> var(m0,un=F)/var(y,un=F)
[1] 0.7373         Le pourcentage de variance expliquée = rapport de
corrélation
> cor(m0,y)^2
[1] 0.7373         Le carré de corrélation avec la moyenne par classe
```

Le pourcentage de variance expliquée est un carré de corrélation. Donc chercher une combinaison de variables qui optimise le pourcentage de variance expliquée (rapport de corrélation), c'est chercher une combinaison de variables la plus corrélée avec une combinaison du paquet d'indicatrices de classe. C'est une analyse canonique. L'analyse des correspondances est ainsi une double analyse canonique et introduite comme telle dans Hill (1973).

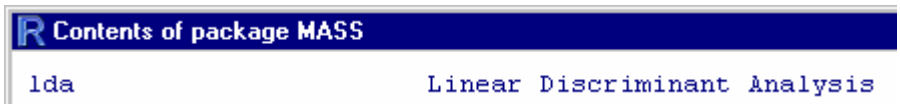
Les schémas de l'analyse discriminante sont donc :



Mais **Y** est un paquet d'indicatrices (*m* classes). Pour faire les calculs on utilise le schéma de gauche, mais pour interpréter, on le modifie sans changer sa nature :



Dans **D,** on trouve les poids des points. $\mathbf{Y}^t\mathbf{DY}$ est une matrice diagonale (deux indicatrices n'ont aucune valeur non nulle à la même place) qui contient les poids des classes (somme des poids des points de la classe). $\mathbf{Y}^t\mathbf{DX}$ fait les sommes pondérées des valeurs de **X** par classe et la multiplication par $\left(\mathbf{Y}^t\mathbf{DY}\right)^{-1}$ divise par le poids de la classe. On trouve donc avec les facteurs des coefficients de combinaisons linéaires des variables de **X** qui maximisent la variance des moyennes par classe, donc la variance inter-classes. Quand on part d'une ACP normée, on obtient l'analyse discriminante linéaire (ADL) fondamentale en morphométrie où elle est née.




```
> library(MASS)
> ?lda
> data(iris)
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> dim(iris)
[1] 150   5
> iris[1:5,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa

> lda(iris[,1:4],iris$Species)
Call:
lda.data.frame(iris[, 1:4], iris$Species)

Prior probabilities of groups:
    setosa versicolor  virginica
    0.3333     0.3333     0.3333
```

```
Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1     LD2
Sepal.Length -0.8294  0.0241
Sepal.Width  -1.5345  2.1645
Petal.Length  2.2012 -0.9319
Petal.Width   2.8105  2.8392

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

`> iris.lda = lda(iris[,1:4],iris$Species)` **dans MASS**

```
> iris.pca = dudi.pca(iris[,1:4])
Select the number of axes: 2
```

`> iris.dis = discrimin (iris.pca, iris$Species)` **dans ade4**

```
Select the number of axes: 2

> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length       1.0000     -0.1176       0.8718      0.8179
Sepal.Width       -0.1176      1.0000      -0.4284     -0.3661
Petal.Length       0.8718     -0.4284       1.0000      0.9629
Petal.Width        0.8179     -0.3661       0.9629      1.0000

> iris.pca$eig
[1] 2.91850 0.91403 0.14676 0.02071
```

Les deux programmes sont assez différents. On ne retrouvera pas facilement les concordances. Si **X** est le tableau normalisé :

```
> names(iris.lda)
[1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"
[8] "call"
> names(iris.dis)
[1] "eig"  "nf"   "fa"   "li"   "va"   "cp"   "gc"   "call"

> iris.lda$scaling
                    LD1     LD2
Sepal.Length -0.8294  0.0241
Sepal.Width  -1.5345  2.1645
Petal.Length  2.2012 -0.9319
Petal.Width   2.8105  2.8392
```

On trouve des poids des variables qui permettent de calculer des combinaisons linéaires des variables de départ.

```
> iris.dis$fa
                   DS1      DS2
Sepal.Length -0.1200  0.01772
Sepal.Width  -0.1169  0.83778
Petal.Length  0.6790 -1.46088
Petal.Width   0.3744  1.92177
```

La parenté est lointaine et ce ne sont pas les mêmes.

```
 scaling: a matrix which transforms observations to discriminant
          functions, normalized so that within groups covariance matrix
          is spherical.
```

Il y a une grosse difficulté théorique. Si **X** est le tableau normalisé, les variables (colonnes) sont des vecteurs de $\mathbb{R}^n$. Les indicatrices des classes forment un sous-espace

vectoriel sur lequel sont projetées les variables. Ce projecteur s'écrit $\mathbf{P} = \mathbf{Y}\left(\mathbf{Y}^t\mathbf{D}\mathbf{Y}\right)^{-1}\mathbf{Y}^t\mathbf{D}$, $\mathbf{PX}$ est le tableau projeté où les données sont remplacées par le centre de gravité (moyenne) de la classe correspondante. La fonction `lda` donne ces valeurs brutes :

```
> iris.lda$means
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026
```

Le tableau $\mathbf{X} - \mathbf{PX}$ donne la différence, soit le tableau des écarts aux moyennes par classe. On peut calculer sa matrice de covariances :

$$\left(\mathbf{X} - \mathbf{PX}\right)^t \mathbf{D}\left(\mathbf{X} - \mathbf{PX}\right) = \mathbf{X}^t\mathbf{D}\mathbf{X} - \mathbf{X}^t\mathbf{P}^t\mathbf{D}\mathbf{X} - \mathbf{X}^t\mathbf{D}\mathbf{P}\mathbf{X} + \mathbf{X}^t\mathbf{P}^t\mathbf{D}\mathbf{P}\mathbf{X}$$

Un projecteur a pour propriété fondamentale d'assurer que $\mathbf{X}^t\mathbf{P}^t\mathbf{D}\left(\mathbf{X} - \mathbf{PX}\right) = 0$ car les variables projetées sont dans le sous-espace et les écarts aux variables projetées sont orthogonales à ce sous-espace. Donc $\left(\mathbf{X} - \mathbf{PX}\right)^t \mathbf{D}\left(\mathbf{X} - \mathbf{PX}\right) = \mathbf{X}^t\mathbf{D}\mathbf{X} - \mathbf{X}^t\mathbf{P}^t\mathbf{D}\mathbf{X}$, ou encore :

$$\mathbf{X}^t\mathbf{D}\mathbf{X} = \mathbf{X}^t\mathbf{P}^t\mathbf{D}\mathbf{P}\mathbf{X} + \left(\mathbf{X} - \mathbf{PX}\right)^t \mathbf{D}\left(\mathbf{X} - \mathbf{PX}\right)$$

La matrice de covariances de **X** se décompose en matrice de covariances du tableau des moyennes par classe (covariances inter-classes) et matrice de covariances du tableau des écarts aux moyennes par classe (covariances intra-classes). L'équation de l'analyse de variance (variance = variance inter + variance intra) s'étend à l'ensemble de la matrice de covariances :

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

On peut donc réécrire le schéma de l'analyse discriminante en appelant **T** la matrice des covariances totales ($\mathbf{T} = \mathbf{X}^t\mathbf{D}\mathbf{X}$), **G** le tableau des moyennes par classe ($\mathbf{G} = \left(\mathbf{Y}^t\mathbf{D}\mathbf{Y}\right)^{-1}\mathbf{Y}^t\mathbf{D}\mathbf{X}$) et $\mathbf{D}_m$ est la diagonale des poids des classes ($\mathbf{D}_m = \mathbf{Y}^t\mathbf{D}\mathbf{Y}$) :

$$\begin{array}{ccc} & \mathbf{T}^{-1} & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{G}^t \uparrow & & \downarrow \mathbf{G} \\ \boxed{m} & \leftarrow & \boxed{m} \\ & \mathbf{D}_m & \end{array}$$

On y trouve les combinaisons linéaires de variance unité et de covariance totale nulle deux à deux qui maximisent la variance inter-classe. La note "*a matrix which transforms observations to discriminant functions, normalized so that within groups covariance matrix is spherical*" indique qu'on utilise le schéma :

$$\begin{array}{ccc} & \mathbf{W}^{-1} & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{G}^t \uparrow & & \downarrow \mathbf{G} \\ \boxed{m} & \leftarrow & \boxed{m} \\ & \mathbf{D}_m & \end{array}$$

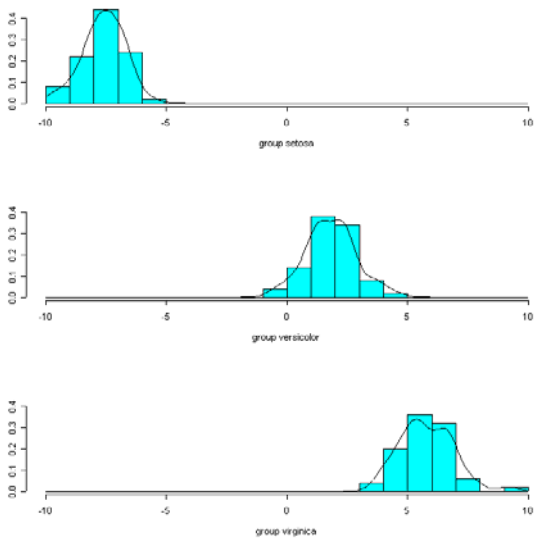
On y trouve les combinaisons linéaires de variance intra-classe unité et de covariance intra-classe nulle deux à deux qui maximisent la variance inter-classe. C'est un problème très voisin car $\mathbf{G}^t\mathbf{D}_m\mathbf{G} = \mathbf{X}^t\mathbf{P}^t\mathbf{D}\mathbf{P}\mathbf{X} = \mathbf{B}$. Si λ_k est valeur propre du premier et μ_k est valeur propre du second on a :

$$\mathbf{T}^{-1}\mathbf{B}\mathbf{u}_k = \lambda_k\mathbf{u}_k \Rightarrow \mathbf{B}\mathbf{u}_k = \lambda_k\mathbf{T}\mathbf{u}_k = \lambda_k(\mathbf{W}+\mathbf{B})\mathbf{u}_k \Rightarrow \mathbf{B}\mathbf{u}_k = \lambda_k\mathbf{W}\mathbf{u}_k + \lambda_k\mathbf{B}\mathbf{u}_k$$

$$\Rightarrow (1-\lambda_k)\mathbf{B}\mathbf{u}_k = \lambda_k\mathbf{W}\mathbf{u}_k \Rightarrow \mathbf{W}^{-1}\mathbf{B}\mathbf{u}_k = \frac{\lambda_k}{1-\lambda_k}\mathbf{u}_k \Rightarrow \mu_k = \frac{\lambda_k}{1-\lambda_k} \Rightarrow \lambda_k = \frac{\mu_k}{1+\mu_k}$$

La trace du premier $Trace(\mathbf{T}^{-1}\mathbf{B})$ s'appelle le critère de Pillai (Pillai 1955) et celle du second $Trace(\mathbf{W}^{-1}\mathbf{B})$ s'appelle le critère généralisé de Hotelling dû à Lawley (1938). Ces quantités donnent des tests d'hypothèses dans le modèle gaussien : toute l'information est dans Tomassone et al. (1988).

```
> plot(iris.lda,dimen=1, type="both")
```



La fonction lda donne la combinaison de variables de variance intra-classe unité (la variance par classe en moyenne vaut 1) qui maximise la variance inter-classe. La fonction discrimin donne la combinaison de variables de variance unité qui maximise la variance inter-classe.

La procédure `discrimin` utilise le schéma :

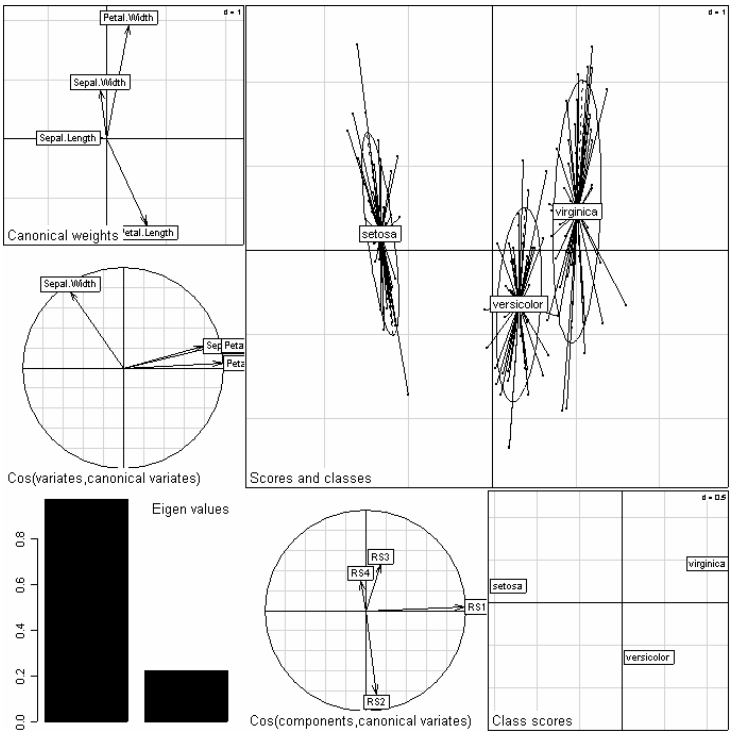
$$\begin{array}{ccc} & \mathbf{C}^{-1} & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{X}_0^t\mathbf{P}^t \uparrow & & \downarrow \mathbf{P}\mathbf{X}_0 \\ \boxed{n} & \leftarrow & \boxed{n} \\ & \mathbf{D} & \end{array} \qquad [1]$$

La procédure `lda` utilise le schéma :

$$\begin{array}{ccc} & \mathbf{W}^{-1} & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{X}_0^t\mathbf{P}^t \uparrow & & \downarrow \mathbf{P}\mathbf{X}_0 \\ \boxed{n} & \leftarrow & \boxed{n} \\ & \mathbf{D} & \end{array} \qquad [2]$$

Les deux procédures donnent à une constante près la même fonction discriminante et les deux procédures ont des valeurs propres liées par une fonction simple. Sur cette base commune, les deux procédures ont des applications différentes.

```
> plot(iris.dis)
```



Dans la première colonne, représentation des poids des variables dans la constitution des scores canoniques, puis (cercle) représentation des corrélations entre variables et scores canoniques. Si des incohérences entre ces deux approches apparaissent, c'est le signe d'une analyse numériquement instable. Consulter Ter Braak (1990). En dessous, valeurs propres de l'analyse, à côté corrélations entre scores d'ACP et scores de discrimination (ou projection des composantes principales sur le plan des scores canoniques, enfin représentation des groupes (ACP inter-classes avec métrqiue de Mahalanobis). En haut et à droite, représentation des scores et des groupes, donc des variances inter et intra-classes.

Le choix fait dans ade4 permet d'introduire dans une analyse discriminante un triplet de départ arbitraire, ce qui donne par exemple, une analyse discriminante des correspondances (Perrière et al. 1996, Perrière and Thioulouse 2002). Voir la fonction `discrimin.coa`.

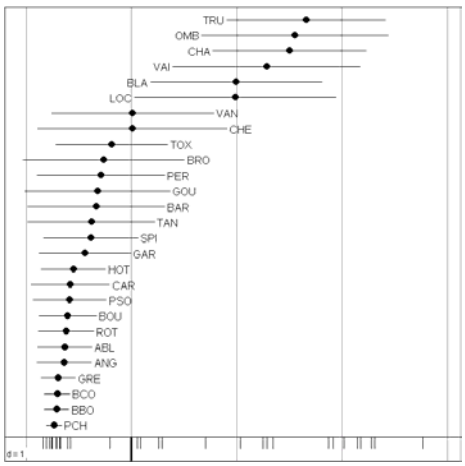
L'analyse discriminante est un cas particulier de l'analyse canonique. Notons que :

```
> iris.dis$eig
[1] 0.9699 0.2220

> x <- iris[,1:4]
> i1 <- as.numeric(iris$Species==levels(iris$Species)[1])
> i2 <- as.numeric(iris$Species==levels(iris$Species)[2])
> y <- cbind.data.frame(i1,i2)
> can0 <- cancor(x,y)
> can0
$cor
[1] 0.9848 0.4712

$xcoef
         [,1]      [,2]     [,3]     [,4]
[1,]  0.01187 -0.001753 -0.25105  0.07782
[2,]  0.02197 -0.157466  0.12503 -0.18250
[3,] -0.03151  0.067796  0.12746 -0.21329
[4,] -0.04023 -0.206546 -0.03786  0.37437

$ycoef
        [,1]    [,2]
[1,] 0.19465 0.04595
[2,] 0.05753 0.19155

$xcenter
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
       5.843        3.057        3.758        1.199

$ycenter
    i1     i2
0.3333 0.3333

> can0$cor^2
[1] 0.9699 0.2220

> w <- as.vector(scale(x,scale=F)%*%can0$xcoef[,1])*sqrt(150)
> w
[1]  1.41352  1.24991  1.31324  1.19460  1.42589  1.35043  1.26463  1.33348
…
[145] -1.20059 -0.98977 -0.90816 -0.87102 -1.03205 -0.82112

> iris.dis$li[c(1:5,141:150),1]
 [1] -1.4135 -1.2499 -1.3132 -1.1946 -1.4259  1.1665  0.8952  0.9657
 [9]  1.1916  1.2006  0.9898  0.9082  0.8710  1.0321  0.8211
```

L'analyse discriminante est donc bien une analyse canonique.

L'analyse des correspondances est donc une double analyse discriminante, comme on l'a vue, puisque chaque paquet d'indicatrices dans l'analyse canonique sert pour la discrimination de l'autre. Souvent on ne se sert que de l'une des deux (Hill 1977), par exemple de la discrimination des espèces par les relevés.

```
> doubs.coa = dudi.coa(doubs$poi,scan=F)
> sco.distri(doubs.coa$l1[,1],doubs$poi)
```



Une analyse canonique fort singulière a été baptisée LONGI car elle est utilisée pour l'étude de la croissance et les données longitudinales. Un tableau **X** de plusieurs mesures anthropométriques est confronté à deux indicatrices, la première **B** portant sur l'individu mesuré et la seconde **A** portant sur son âge (mesures longitudinales au cours de la croissance). On peut définir le sous-espace $A \cap B^{\perp}$ (Afriat 1957) ensemble des variables constantes par classe d'âge et de moyenne nulle par individu pour faire l'analyse canonique avec l'espace des variables définies par le tableau **X**. Ces approches sont développées dans Pontier et al. (1990).

A noter, dans les calculs, le présence de la propriété générale :

Soit **P** la matrice d'un projecteur ($\mathbf{P}^2 = \mathbf{P}$) dans une base quelconque de $\mathbb{R}^n$, **Q** la matrice d'un produit scalaire de $\mathbb{R}^n$ dans cette même base. **P** est un projecteur **Q**-orthogonal si et seulement si $\mathbf{QP} = \mathbf{P}^t\mathbf{Q} = \mathbf{P}^t\mathbf{QP}$.

La condition nécessaire est vérifiée par l'écriture de $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^t\mathbf{QX}\right)^{-1}\mathbf{X}^t\mathbf{Q}$ où **X** contient en colonnes une base de l'image du projecteur exprimée dans la base de référence. Réciproquement, si la propriété est vérifiée :

$$\left\langle \mathbf{Py} \middle| \left(\mathbf{x} - \mathbf{Px}\right) \right\rangle_{\mathbf{Q}} = \mathbf{y}^t\mathbf{P}^t\mathbf{Q}\left(\mathbf{x} - \mathbf{Px}\right) = 0$$

2.5. Analyse canonique des correspondances

L'analyse canonique des correspondances (CCA) étend la stratégie de l'analyse des correspondances. En effet en réécrivant le tableau faunistique comme deux tableaux d'indicatrices, la mention 00010000000...000 qui indique auxquels des relevés doit être rattachée l'occurrence peut être remplacée par l'enregistrement des variables de milieu de ce relevé. On fait alors l'analyse canonique entre les indicatrices des espèces et les variables de milieu couplées par le biais des occurrences. On appelle cette méthode l'analyse canonique des correspondances qui se retrouve être l'analyse discriminante des occurrences (milieu) par la variable qualitative nom d'espèce. Elle donne des combinaisons linéaires de variables de milieu de variance unité qui maximisent la variance des positions moyennes des espèces (Ter Braak 1986, Chessel et al. 1987, Ter Braak 1987, Lebreton et al. 1988a, Lebreton et al. 1988b). Cette méthode, « dominante sur le marché », suppose qu'on ne voit dans le tableau floro-faunistique que des

présences c'est-à-dire dans les relevés des assemblages d'espèces qui sont chacune dans un milieu pour des raisons qui leur sont propres (théorie de la niche).

On peut comprendre l'opération sur un exemple. Soient 4 sites, 3 espèces et 2 variables de milieu. Le couple de tableaux appariés par les sites devient un couple de tableaux appariés par les occurrences. Cette dualité est une source de difficulté. On peut *naïvement* écrire :

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1.5 & 10.0 \\ 1.8 & 8.7 \\ 3.1 & 8.3 \\ 3.2 & 8.4 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 & 10.0 \\ 1.5 & 10.0 \\ 1.8 & 8.7 \\ 1.8 & 8.7 \\ 3.1 & 8.3 \\ 1.8 & 8.7 \\ 3.1 & 8.3 \\ 3.2 & 8.4 \\ 3.2 & 8.4 \end{bmatrix}$$

Si les poids des occurrences dans le tableau sites-espèces ne sont pas entiers il faudrait réécrire une colonne de poids au lieu de décomposer les entiers. On place les poids à l'extérieur :

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1.5 & 10.0 \\ 1.8 & 8.7 \\ 3.1 & 8.3 \\ 3.2 & 8.4 \end{bmatrix} \begin{Bmatrix} 2 \\ 3 \\ 2 \\ 2 \end{Bmatrix} \mapsto \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1.5 & 10.0 \\ 1.8 & 8.7 \\ 1.8 & 8.7 \\ 1.8 & 8.7 \\ 3.1 & 8.3 \\ 3.1 & 8.3 \\ 3.2 & 8.4 \end{bmatrix} \begin{Bmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 2 \end{Bmatrix}$$

Mathématiquement, le premier tableau relève (entre autre) d'une analyse des correspondances qui donne un tableau de fréquences bivarié. Le second relève d'une analyse en composantes principales normée. Pour que le couplage fonctionne les poids doivent être cohérents. Ce sont nécessairement ceux de l'AFC. Centrer et normaliser le tableau de milieu avec les poids issus du tableau faunistique donne un résultat cohérent si on utilise les $f_{i.}$ sur le tableau sites-variables ou les f_{ij} sur le tableau occurrences-variables. La position moyenne de l'espèce 1 (profil 2 1 0 0) sur la première variable de milieu (valeurs 1.5 1.8 3.1 3.2) vaut $\left(2\times 1.5+1\times 1.8\right)/3$ à gauche et évidemment la même chose à droite. On note *I* le nombre de sites, *J* le nombre d'espèces et *p* le nombre de variables pour retrouver les schémas des analyses simples :

$$\begin{array}{ccccc} \boxed{J} & \overset{\mathbf{D}_J}{\rightarrow} & \boxed{J} & \boxed{p} & \overset{\mathbf{I}_p}{\rightarrow} & \boxed{p} \\ \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1}\uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} & \mathbf{X}^t\uparrow & & \downarrow \mathbf{X} \\ \boxed{I} & \underset{\mathbf{D}_I}{\leftarrow} & \boxed{I} & \boxed{I} & \underset{\mathbf{D}_I}{\leftarrow} & \boxed{I} \end{array}$$

En passant aux occurrences, on aura :

$$\begin{array}{ccccc} \boxed{J} & \overset{\mathbf{I}_J}{\rightarrow} & \boxed{J} & \boxed{p} & \overset{\mathbf{I}_p}{\rightarrow} & \boxed{p} \\ \mathbf{L}^t\uparrow & & \downarrow \mathbf{L} & \mathbf{Y}^t\uparrow & & \downarrow \mathbf{Y} \\ \boxed{O} & \underset{\mathbf{D}_O}{\leftarrow} & \boxed{O} & \boxed{O} & \underset{\mathbf{D}_O}{\leftarrow} & \boxed{O} \end{array}$$

Y est le tableau occurrences-milieu après normalisation, **X** est le tableau sites-milieu après normalisation. Les deux matrices de corrélations sont les mêmes :

$$\mathbf{Y}^t\mathbf{D}_O\mathbf{Y} = \mathbf{X}^t\mathbf{D}_I\mathbf{X}$$

L est le tableau des indicatrices occurrences-espèces. Son schéma est entièrement artificiel puisque $\mathbf{L}^t\mathbf{D}_O\mathbf{L} = \mathbf{D}_J$. Le tableau espèces-variables des positions moyennes des espèces sur les variables se calcule alors de deux manières en restant exactement lui-même :

$$\mathbf{M} = \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{X} = \mathbf{D}_J^{-1}\mathbf{L}^t\mathbf{D}_O\mathbf{Y}$$

Si on fait l'analyse canonique du couple vu par les occurrences, on a le schéma :

$$\begin{array}{ccccc} \boxed{p} & \overset{\left(\mathbf{Y}^t\mathbf{D}_O\mathbf{Y}\right)^{-1}}{\rightarrow} & \boxed{p} & \Leftrightarrow & \boxed{p} & \overset{\left(\mathbf{Y}^t\mathbf{D}_O\mathbf{Y}\right)^{-1}}{\rightarrow} & \boxed{p} \\ \mathbf{Y}^t\mathbf{D}_O\mathbf{L}\uparrow & & \downarrow \mathbf{L}^t\mathbf{D}_O\mathbf{Y} & & \mathbf{Y}^t\mathbf{D}_O\mathbf{L}\mathbf{D}_J^{-1}\uparrow & & \downarrow \mathbf{D}_J^{-1}\mathbf{L}^t\mathbf{D}_O\mathbf{Y} \\ \boxed{J} & \underset{\mathbf{D}_J^{-1}}{\leftarrow} & \boxed{J} & & \boxed{J} & \underset{\mathbf{D}_J}{\leftarrow} & \boxed{J} \end{array}$$

Si on fait l'ACPVI (voir ci-dessous) du couple vu par les sites, on a le schéma :

$$\begin{array}{ccc} & \overset{\left(\mathbf{X}^t\mathbf{D}_I\mathbf{X}\right)^{-1}}{\rightarrow} & \\ \boxed{p} & & \boxed{p} \\ \mathbf{X}^t\mathbf{D}_I\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}\uparrow & & \downarrow \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1}\mathbf{D}_I\mathbf{X} \\ \boxed{J} & \underset{\mathbf{D}_J}{\leftarrow} & \boxed{J} \end{array}$$

C'est-à-dire exactement la même chose qui s'écrit :

$$
\begin{array}{ccc}
 & \left(\mathbf{X}^t\mathbf{D}_O\mathbf{X}\right)^{-1} & \\
\boxed{p} & \rightarrow & \boxed{p} \\
\mathbf{M}^t \uparrow & & \downarrow \mathbf{M} \\
\boxed{J} & \leftarrow & \boxed{J} \\
 & \mathbf{D}_J &
\end{array}
$$

On obtient, comme cas particulier du modèle général, que les facteurs de cette analyse donnent des combinaisons de variables de milieu qui maximisent la variance des positions moyennes des espèces. Mais, très curieusement, pour obtenir ce schéma unique, on a pris un couplage d'analyse canonique en associant les tableaux par les occurrences et un couplage avec une seule inversion de métrique (donc une ACPVI, chapitre suivant) en associant les tableaux par les sites. Quand il existe plusieurs justificatifs théoriques distincts pour une même procédure, en général la procédure a du succès (bibliographie dans Birks and Austin (1992)) et on retiendra des justificatifs les plus adéquats à défendre un point de vue. Plusieurs autres modèles se cachent dans ce schéma complexe et on trouvera des compléments dans des synthèses récentes (Lebreton et al. 1991, Ter Braak and Verdonschot 1995).

L'important à retenir : l'ACC se pratique après une AFC du tableau faune et une ACP normée du tableau de milieu *utilisant la pondération des relevés issue de l'AFC* du tableau faune. De ce point de vue encore, le milieu enregistré dans un relevé est d'autant plus important qu'il y a plus de taxons présents (un relevé vide ne compte pas). C'est le milieu de l'occurrence qui est ainsi pris en compte. Les facteurs limitant sont minimisés, les facteurs de séparation de niches écologiques sont maximisés. C'est une contrainte forte.

```
> cca1 = cca(doubs$poi,doubs$mil)
Select the number of axes: 2
> cca1
Principal Component Analysis with Instrumental Variables
call: cca(sitspe = doubs$poi, sitenv = doubs$mil)
class: cca pcaiv dudi

$rank (rank)     : 11
$nf (axis saved) : 2

eigen values: 0.5345 0.1218 0.0687 0.04917 0.02709 ...

 vector length mode    content
 $eig   11     numeric eigen values
 $lw    30     numeric row weigths (from dudi)
 $cw    27     numeric col weigths (from dudi)

 data.frame nrow ncol content
 $Y         30   27   Dependant variables
 $X         30   11   Explanatory variables
 $tab       30   27   modified array (projected variables)

 data.frame nrow ncol content
 $c1        27   2    PPA Pseudo Principal Axes
 $as        8    2    Principal axis of dudi$tab on PAP
 $ls        30   2    projection of lines of dudi$tab on PPA
 $li        30   2    $ls predicted by X

 data.frame nrow ncol content
 $fa        12   2    Loadings (CPC as linear combinations of X
 $l1        30   2    CPC Constraint Principal Components
 $co        27   2    inner product CPC - Y
 $cor       12   2    correlation CPC - X

 iner  inercum inerC inercumC ratio R2    lambda
```

```
0.601 0.601   0.597 0.597    0.994 0.895 0.535
0.144 0.745   0.142 0.739    0.992 0.857 0.122
```

La procédure a été écrite sur la base du schéma :

$$\begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{Y}^t \uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} \\ \boxed{I} & & \boxed{I} \\ \mathbf{P}_{\mathbf{X}}^t \uparrow & & \downarrow \mathbf{P}_{\mathbf{X}} = \mathbf{X}\left(\mathbf{X}^t\mathbf{D}_I\mathbf{X}\right)^- \mathbf{X}^t\mathbf{D}_I \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array}$$

C'est une écriture possible du précédent en s'arrêtant sur *I* relevés et *J* espèces. Il y a encore deux lectures du schéma et deux modes d'interprétation de l'analyse. On peut garder le langage de l'écologie (*J* espèces, *I* sites et *p* variables de milieu). Le premier mode utilise :

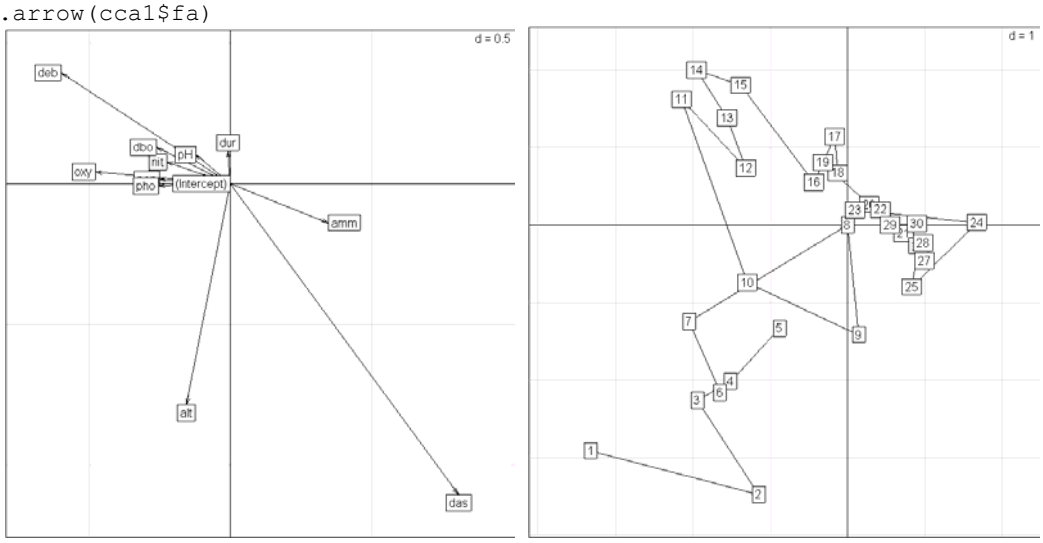
tab le tableau $\mathbf{P}_{\mathbf{X}}\mathbf{Y}$ d'usage purement technique

cw poids des colonnes (la pondération marginales des espéces dans le tableau faunistique).

lw poids des lignes (la pondération marginale des sites dans le tableau faunistique qui sert aux régressions pondérées)

fa poids des variables (coefficient des combinaisons linéaires de variables de milieu, du type loadings)

```
> s.arrow(cca1$fa)
```



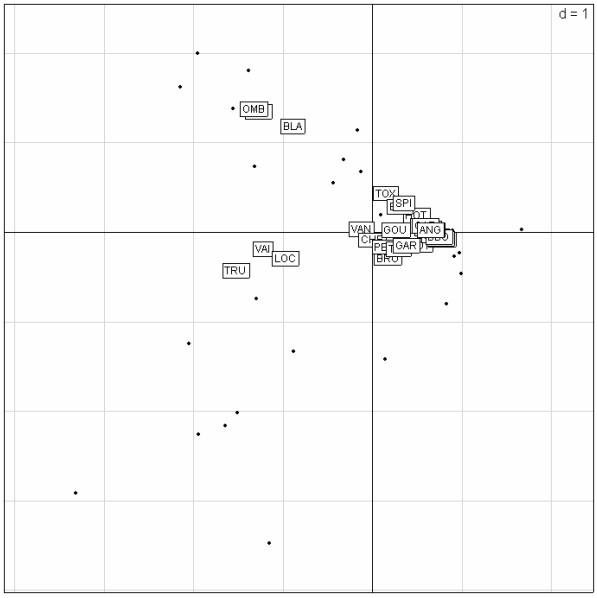
fa l1

l1 composantes principales sous contrainte ou CPC (combinaisons linéaires de variables de **x** maximisant le critère de l'analyse). Ce sont des scores des relevés de variance 1, combinaison linéaire des variables de milieu.

```
> s.traject(cca1$l1,clab=0)
> s.label(cca1$l1,add.p=T)
```

co Le calcul montre que ce sont les positions des espèces au barycentre des relevés pour leur distribution (espèces à la moyenne des relevés).

```
> s.label(cca1$l1,clab=0)
> s.label(cca1$co,add.p=T,clab=0.75)
```



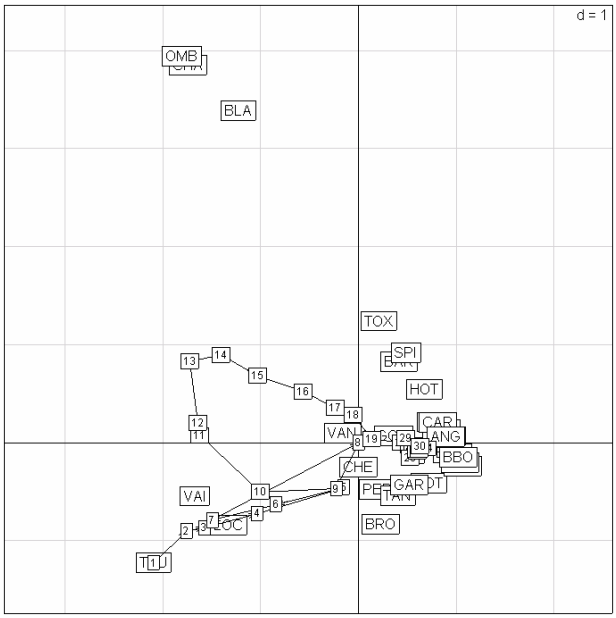
eig valeurs propres, optimum du critère somme des carrés des écarts entre moyenne par espèce et origine, c'est à dire variance des moyennes conditionelles (double centrage).

Quand on interprète l'analyse avec ce point de vue on fait une analyse canonique des correspondances CCA au sens de Ter Braak (1986).

Ce schéma est parfaitement adapté à la vision de la niche écologique et des gradients environnementaux sur lesquels se séparent les niches des espèces. Mais dans un schéma de dualité, il y a toujours deux points de vue. Le second est formé de :

c1 pseudo-axes principaux ou PAP, vecteurs normés de $\mathbb{R}^J$. Ce sont des scores des espèces centrés et réduits pour la pondération des espèces.

```
> s.label(cca1$c1)
> s.traject(cca1$ls,clab=0,add.p=T)
> s.label(cca1$ls,add.p=T,clab=0.75)
```

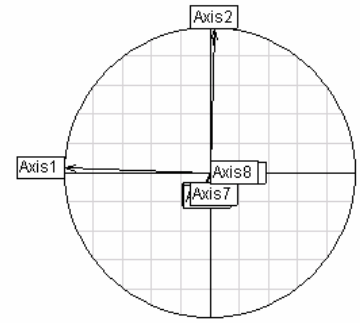


c1 / ls

ls coordonnées des projections des lignes de **Y** sur les PAP. Le calcul montre qu'on calcule ainsi les positions des sites par moyennes des espèces qu'ils contiennent (sites à la moyenne des espèces).

as coordonnées des projections des axes principaux (AP) de **Y** sur les PAP. Ceci permet de replacer les plans principaux de l'AFC d'origine sur les PAP et ainsi de comparer l'analyse simple et l'analyse sous contraintes.

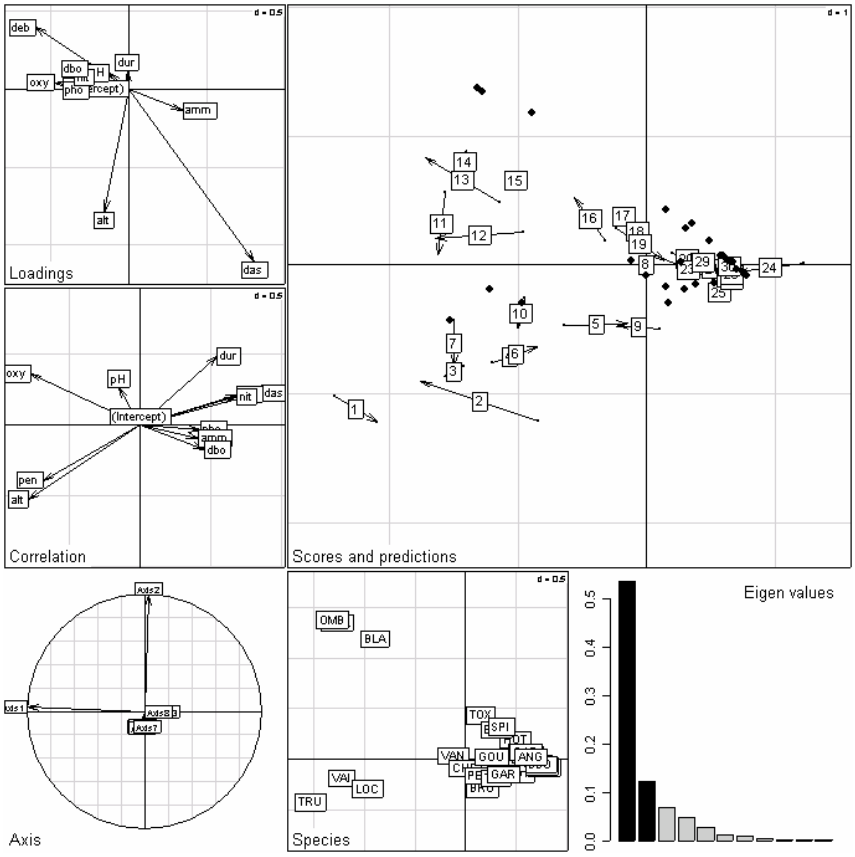
```
> s.corcircle(cca1$as)
```



li prédiction des coordonnées des projections des lignes de **Y** sur les PAP par régressions multiples sur X. Ces régressions définissent des carrés de corrélation multiple ou pourcentage de variance expliquée. Les PAP optimisent la variance expliquée, c'est à dire le produit de la variance de la projection (critère d'ACP) par le carré de corrélation multiple (critère d'analyse canonique). On peut superposer **ls** (projections sur les PAP) et **li** (prédiction des positions).

Quand on utilise ce modèle, on fait une AFCVI au sens de Lebreton et al. (1991).

```
> plot(cca1)
```

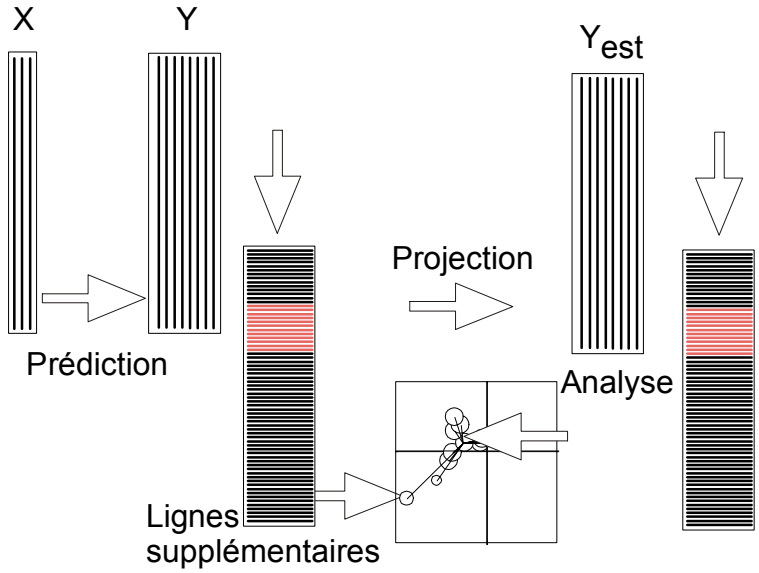


Les très fortes incohérences entre les coefficients qui fabriquent les scores à partir des variables et les corrélations entre ces mêmes scores et ces mêmes variables indiquent que les conditions d'emploi ne sont pas bonnes. C'est un problème très général en régression multiple et toutes les méthodes dérivées (Ter Braak 1990). A utiliser avec la plus grande prudence. L'ACC fait souvent guère plus que l'AFC sans qu'on s'en rende compte.

A retenir : on maximise une corrélation entre combinaisons de variables quantitatives en *analyse canonique des corrélations*. On maximise la corrélation entre scores des lignes et des colonnes dans une *analyse des correspondances*. On maximise un pourcentage de variance expliquée par une partition en *analyse discriminante*. On maximise la variance des moyennes par espèces avec des combinaisons de variables de milieu normalisées en *analyse canonique des correspondances.* Ces méthodes relèvent de la stratégie des analyses canoniques.

3. Stratégie des variables instrumentales

Ce sont des méthodes dissymétriques. Un tableau est formé d'explicatives (variables instrumentales) et un tableau est formé de variables à étudier. Dans ces méthodes, on étudie le second en utilisant le premier. On parle en général d'Analyses en Composantes Principales sur Variables Instrumentales ou ACPVI. Fondations dans Rao (1964). Le principe de base se trouve dans le schéma :



X est le tableau des variables instrumentales. **Y** est le tableau à analyser. Chacune des variables de **Y** est prédite par une régression multiple sur les variables de **X**. Les modèles sont rangés dans le tableau $\mathbf{Y}_{\mathbf{est}}$. Les variables de $\mathbf{Y}_{\mathbf{est}}$ sont obtenues par projection des variables de **Y** sur le sous-espace engendré par les variables de **X**.

Cette opération est linéaire.

On considère alors que $\mathbf{Y}_{\mathbf{est}}$ est un ensemble de lignes. Quand on passe de la ligne *i* du tableau $\mathbf{Y}$ à la ligne *i* du tableau $\mathbf{Y}_{\mathbf{est}}$ l'opération *n'est plus linéaire*. On analyse $\mathbf{Y}_{\mathbf{est}}$ et on projette en lignes supplémentaires celle de $\mathbf{Y}$.

On fait une analyse de $\mathbf{Y}$ sous contrainte de $\mathbf{X}$. Suivant $\mathbf{X}$ et $\mathbf{Y}$, on a un ensemble de méthodes dites d'ordination sous contraintes ou d'analyses sur variables instrumentales. Les schémas utiles sont :

$$\begin{array}{ccccccc}
\boxed{p} & \overset{\mathbf{Q}_p}{\rightarrow} & \boxed{p} & \boxed{m} & \overset{\mathbf{Q}_m}{\rightarrow} & \boxed{m} & \qquad \boxed{p} \quad \overset{\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}}{\rightarrow} \quad \boxed{p} \\
\mathbf{X}^t \uparrow & & \downarrow \mathbf{X} & \mathbf{Y}^t \uparrow & & \downarrow \mathbf{Y} & \mapsto \quad \mathbf{X}^t\mathbf{D}\mathbf{Y} \uparrow \qquad\qquad \downarrow \mathbf{Y}^t\mathbf{D}\mathbf{X} \\
\boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n} & \boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n} & \qquad \boxed{m} \quad \underset{\mathbf{Q}_m}{\leftarrow} \quad \boxed{m}
\end{array}$$

Le schéma de droite est celui d'une ACPVI. On le réécrit sous la forme :

$$\begin{array}{ccc}
\boxed{n} & \overset{\mathbf{D}}{\rightarrow} & \boxed{n} \\
\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D} \uparrow & & \downarrow \mathbf{D}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^t \\
 & \mathbf{D}\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D} & \\
\boxed{n} & \rightarrow & \boxed{n} \\
\mathbf{Y} \uparrow & & \downarrow \mathbf{Y}^t \\
\boxed{m} & \underset{\mathbf{Q}_m}{\leftarrow} & \boxed{m}
\end{array}$$

ou encore :

$$\begin{array}{ccccccc}
\boxed{n} & \overset{\mathbf{D}}{\rightarrow} & \boxed{n} & \boxed{m} & \overset{\mathbf{Q}_m}{\rightarrow} & \boxed{m} \\
\mathbf{P}_{\mathbf{X}} \uparrow & & \downarrow \mathbf{P}_{\mathbf{X}}^t & \mathbf{Y}^t \uparrow & & \downarrow \mathbf{Y} \\
\boxed{n} & & \boxed{n} & \boxed{n} & & \boxed{n} \\
\mathbf{Y} \uparrow & & \downarrow \mathbf{Y}^t & \mathbf{P}_{\mathbf{X}}^t \uparrow & & \downarrow \mathbf{P}_{\mathbf{X}} \\
\boxed{m} & \underset{\mathbf{Q}_m}{\leftarrow} & \boxed{m} & \boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n}
\end{array} \quad \Leftrightarrow$$

qui montre $\mathbf{Y}_{\mathbf{est}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$. On peut changer de projecteurs ce qui permet de distinguer les ACPVI directes et les ACPVI orthogonales qui décomposent une analyse simple avec deux projecteurs complémentaires :

$$\begin{array}{ccccccc}
\boxed{m} & \overset{\mathbf{Q}_m}{\rightarrow} & \boxed{m} & \boxed{m} & \overset{\mathbf{Q}_m}{\rightarrow} & \boxed{m} \\
\mathbf{Y}^t \uparrow & & \downarrow \mathbf{Y} & \mathbf{Y}^t \uparrow & & \downarrow \mathbf{Y} \\
\boxed{n} & & \boxed{n} & \boxed{n} & & \boxed{n} \\
\mathbf{P}_{\mathbf{X}}^t \uparrow & & \downarrow \mathbf{P}_{\mathbf{X}} & \mathbf{I}_n - \mathbf{P}_{\mathbf{X}}^t \uparrow & & \downarrow \mathbf{I}_n - \mathbf{P}_{\mathbf{X}} \\
\boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n} & \boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n}
\end{array}$$

3.1. Analyses inter-classes

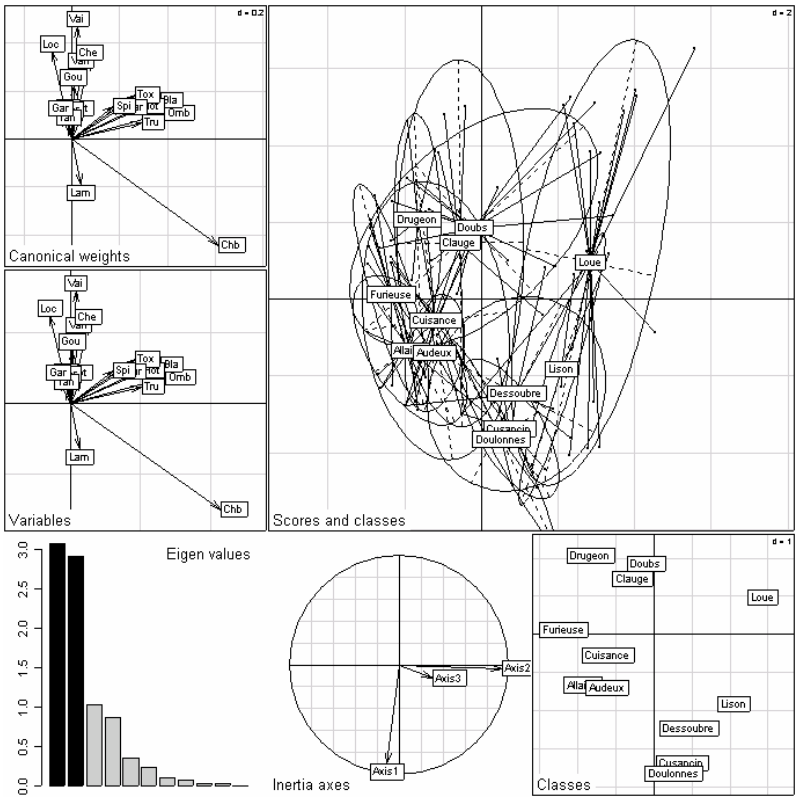
Le cas le plus simple est formé d'un tableau **Y** quelconque et d'un tableau **X** des indicatrices d'une variable qualitative. Prédire **Y** par **X**, c'est simplement remplacer une valeur par la moyenne des individus de la même classe pour la même variable. La transformation opérée par les variables instrumentales est figurée par le graphe en étoiles. L'analyse inter-classe se résume au changement de schéma :

$$\begin{array}{ccc} \boxed{p} & \xrightarrow{\mathbf{Q}_p} & \boxed{p} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \boxed{n} & \xleftarrow[\mathbf{D}]{} & \boxed{n} \end{array} \quad \longmapsto \quad \begin{array}{ccc} \boxed{p} & \xrightarrow{\mathbf{Q}_p} & \boxed{p} \\ \mathbf{X}_g^t \uparrow & & \downarrow \mathbf{X}_g \\ \boxed{g} & \xleftarrow[\mathbf{D}_g]{} & \boxed{g} \end{array}$$

A gauche, p variables mesurées sur n individus. A droite, p variables mesurées sur g groupes. Le second tableau est déduit du premier en faisant les moyennes par classe. Les poids des groupes sont les sommes par classe des poids des points. La métrique est inchangée. Toutes les analyses de base permettent de faire des inter-classes.

Utiliser la liste jv73 (Verneaux 1973). Faire une typologie de rivière à partir d'une typologie de stations.

```
data(jv73)
jv.pca.poi = dudi.pca(jv73$poi,scal=F)
jv.bet.poi = between(jv.pca.poi,jv73$fac.riv)
Select the number of axes: 2

> plot(jv.bet.poi)
```



L'analyse du nuage des centres de gravités par groupe est une analyse simple. On rajoute pour l'interprétation la projection des points d'origine sur les axes inter-classes et la projection des axes d'inertie du nuage initial sur les axes d'inertie inter-classes.

```
> jv.bet.poi
Between analysis
call: between(dudi = jv.pca.poi, fac = jv73$fac.riv)
class: between dudi

$nf (axis saved) : 2
$rank:  11
$ratio:  0.3135

# 31% de variabilité entre rivières
# 69 % de variabilité interne aux rivières

eigen values: 3.079 2.915 1.037 0.8727 0.3558 ...

  vector length mode    content
1 $eig   11     numeric eigen values
2 $lw    12     numeric group weigths
3 $cw    19     numeric col weigths

  data.frame nrow ncol content
1 $tab       12   19   array class-variables
2 $li        12   2    class coordinates
3 $l1        12   2    class normed scores
4 $co        19   2    column coordinates
5 $c1        19   2    column normed scores
6 $ls        92   2    row coordinates
7 $as        3    2    inertia axis onto between axis
```

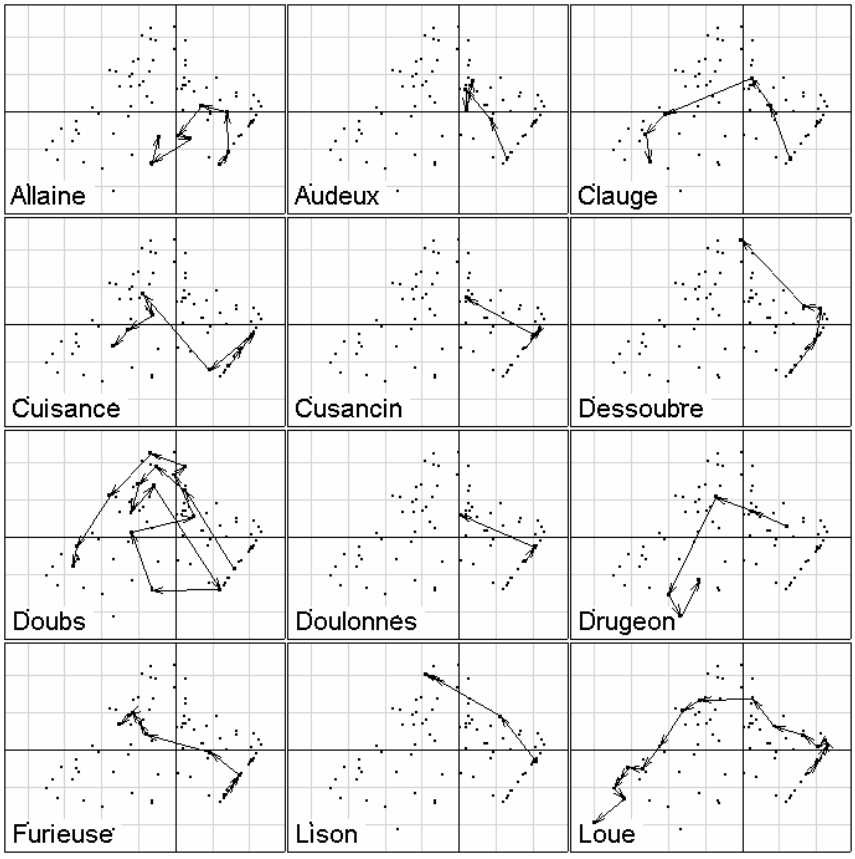
3.2. Analyses intra-classes

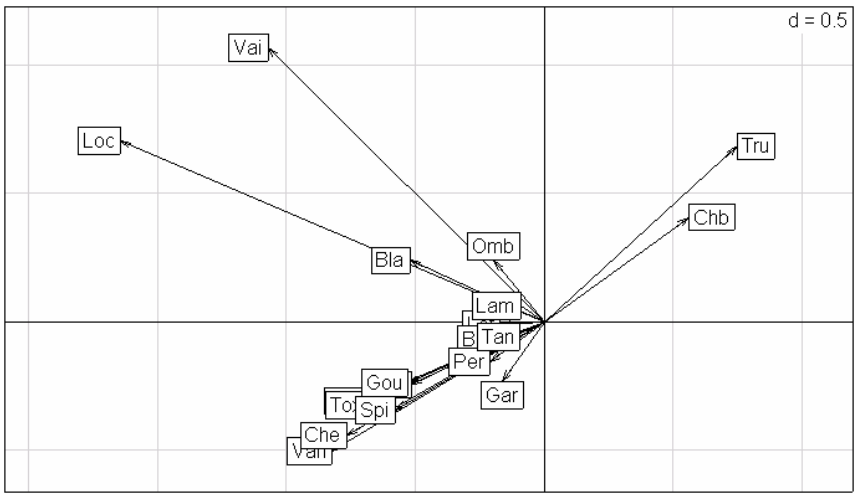
Si on utilise, en lieu de l'analyse du tableau estimé par les variables instrumentales, l'analyse du tableau des résidus des prédictions, on fait une ACPVI orthogonale. Pour le cas des indicatrices des classes, il s'agit de mettre les centres de gravité des sous-nuages à l'origine. Les deux analyses sont strictement complémentaires.

```
> jv.wit.poi = within(jv.pca.poi,jv73$fac.riv)
Select the number of axes: 2

> jv.wit.poi
Within analysis
call: within(dudi = jv.pca.poi, fac = jv73$fac.riv)
class: within dudi

$nf (axis saved) : 2
$rank:  19
$ratio:  0.6865
# 69 % de variabilité interne aux rivières
# 31% de variabilité entre rivières

eigen values: 8.289 3.347 1.974 1.096 1.009 ...

  vector length mode    content
1 $eig   19     numeric eigen values
2 $lw    92     numeric row weigths
3 $cw    19     numeric col weigths
4 $tabw  12     numeric table weigths
5 $fac   92     numeric factor for grouping

  data.frame nrow ncol content
1 $tab       92   19   array class-variables
2 $li        92   2    row coordinates
3 $l1        92   2    row normed scores
```

```
4 $co        19   2    column coordinates
5 $c1        19   2    column normed scores
6 $ls        92   2    supplementary row coordinates
7 $as        3    2    inertia axis onto within axis
```

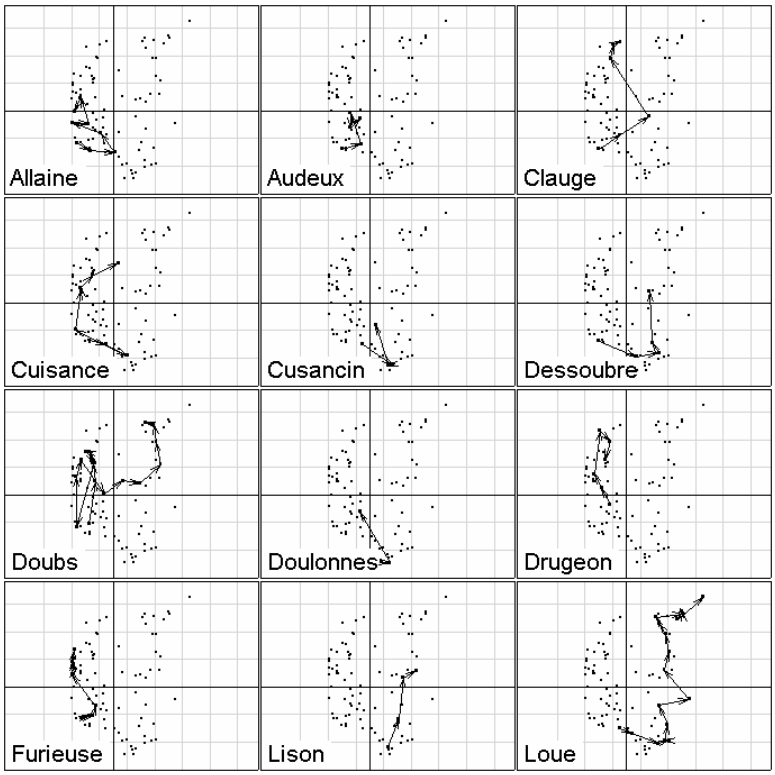
```
par(mfrow=c(4,3))
for (gr in levels(jv.wit.poi$fac)) {
   w = which (jv.wit.poi$fac==gr)
   s.label(jv.wit.poi$ls,clab=0,sub=gr, csub=3, cgrid=0)
   s.traject(jv.wit.poi$ls[w,],add.plot=T,clab=0)
}
par(mfrow=c(1,1))
s.arrow(jv.wit.poi$co)
```

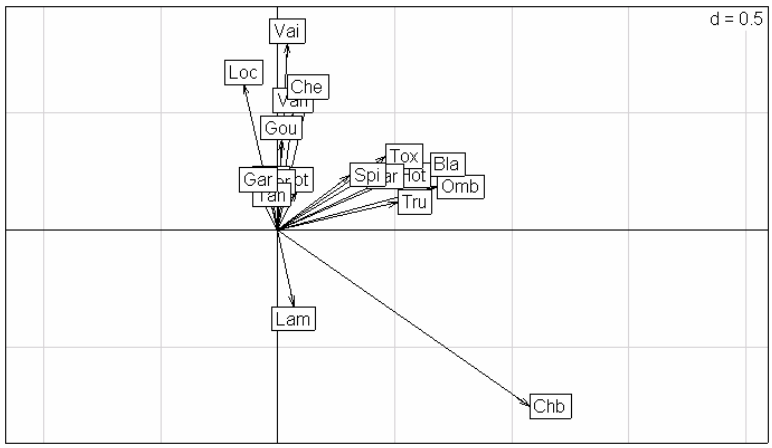





```
par(mfrow=c(4,3))
for (gr in levels(jv.wit.poi$fac)) {
   w = which (jv.wit.poi$fac==gr)
   s.label(jv.bet.poi$ls,clab=0,sub=gr, csub=3, cgrid=0)
   s.traject(jv.bet.poi$ls[w,],add.plot=T,clab=0)
```

```
}
par(mfrow=c(1,1))
s.arrow(jv.bet.poi$co)
```





Les contraintes qu'on peut imposer sont donc très fortes. Il faut identifier les objectifs. Discriminante et inter-classes ont les mêmes objectifs mais pas les mêmes contraintes. L'analyse des correspondances inter-classes est le cas particulier de l'analyse inter-classes après une AFC. On peut aussi la pratiquer en faisant l'AFC du tableau des sommes par classes avec projection en individus supplémentaires des lignes du tableau de départ. On parle alors de discrimination barycentrique (Bergougnan and Couraud 1982) . Retenir les schémas de principe :

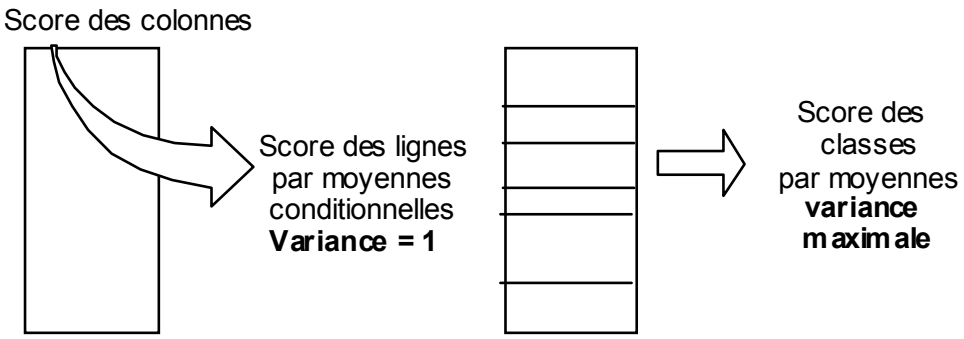


Schéma de principe de l'analyse discriminante des correspondances (Perrière et al. 1996, Perrière and Thioulouse 2002).

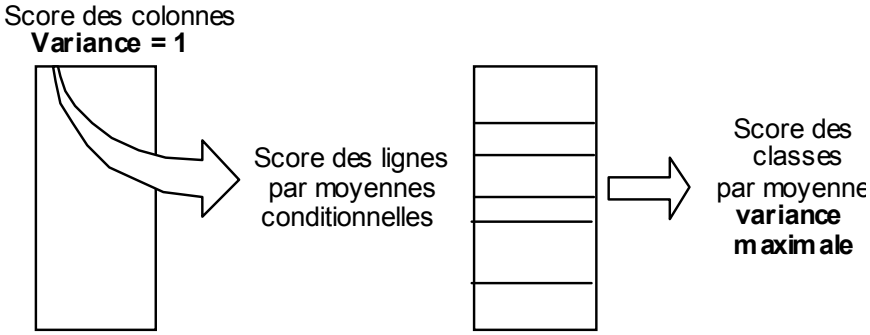


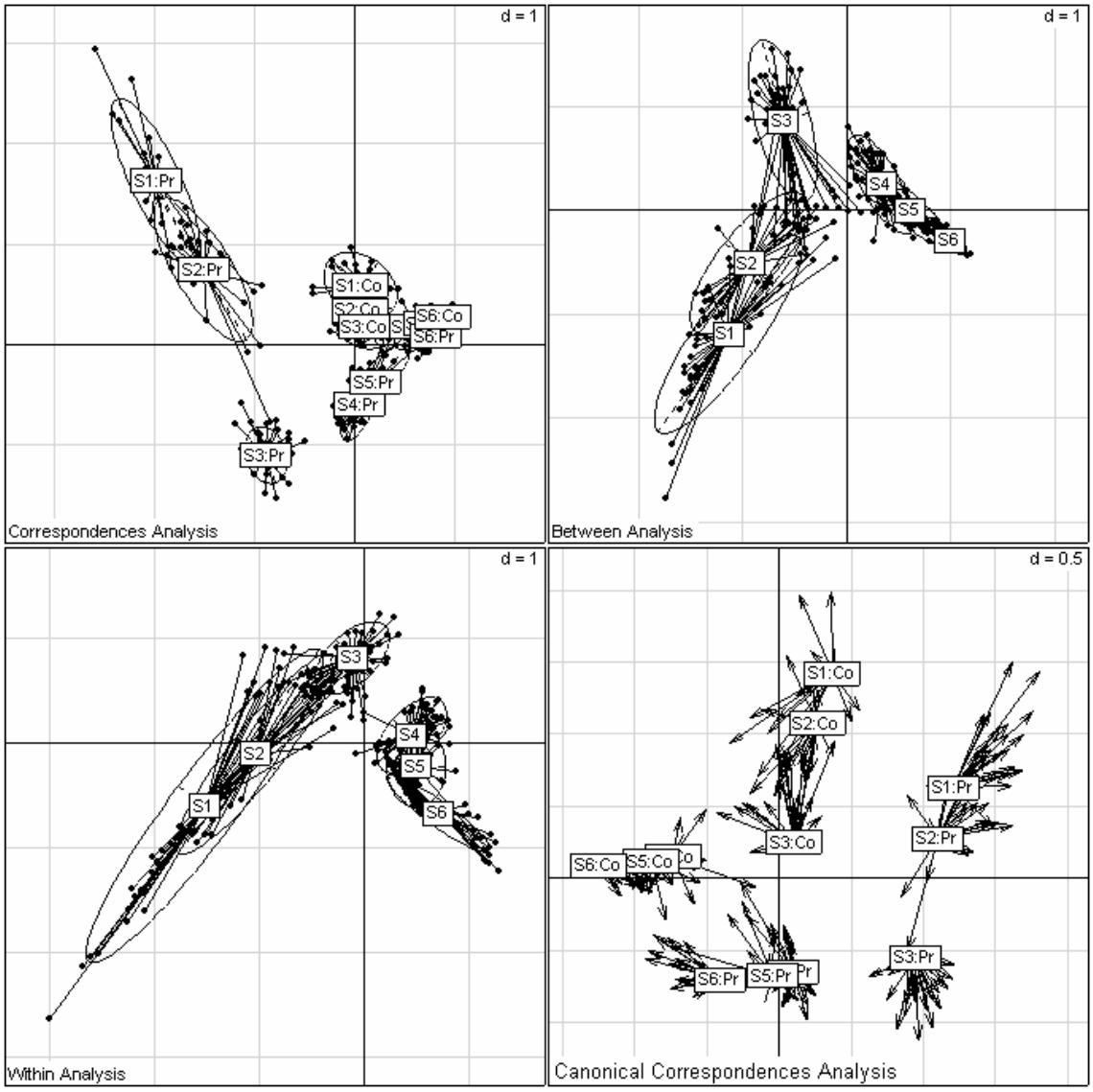
Schéma de principe de l'analyse des correspondances inter-classes.

L'intra-classes en AFC peut s'étendre aux blocs de colonnes et aux doubles contraintes (fonction `witwit.coa`) (Cazes et al. 1988).

3.3. Ordinations sous contraintes

Les contraintes en inter et intra-classes sont simples. On peut les rendre assez complexes en introduisant des sous-espaces de projection (module Projectors). Utiliser la liste avimedi (Blondel et al. 1988) pour faire une analyse sous contrainte **A**+**B** :

```
data(avimedi)
par(mfrow = c(2,2))
coa1 <- dudi.coa(avimedi$fau, scan = FALSE, nf = 3)
s.class(coa1$li,avimedi$plan$str:avimedi$plan$reg,
    sub = "Correspondences Analysis")
bet1 <- between(coa1, avimedi$plan$str, scan = FALSE)
s.class(bet1$ls, avimedi$plan$str,
    sub = "Between Analysis")
wit1 <- within(coa1, avimedi$plan$reg, scan=FALSE)
s.class(wit1$li, avimedi$plan$str,
    sub = "Within Analysis")
pcaiv1 <- pcaiv(coa1, avimedi$plan, scan = FALSE)
s.match(pcaiv1$li, pcaiv1$ls, clab = 0,
    sub = "Canonical Correspondences Analysis")
s.class(pcaiv1$li, avimedi$plan$str:avimedi$plan$reg,
    add.plot = TRUE)
par(mfrow = c(1,1))
```



Des relevés d'avifaune sont rangés par région (Pr Provence, Co Corse) et strates de végétation (S1 pelouses à S6 forêts élevées). On peut illustrer la typologie brute par les variables de contrôle (en haut, à gauche), imposer la structure inter-strates (en haut à gauche), éliminer la structure inter-régions (en bas, à droite) ou prendre en compte les deux facteurs (en bas, à gauche). La dernière signifie qu'on veut une analyse dont les coordonnées des lignes soient de la meilleure manière (aux moindres carrés pondérés) du type **A**+**B**.

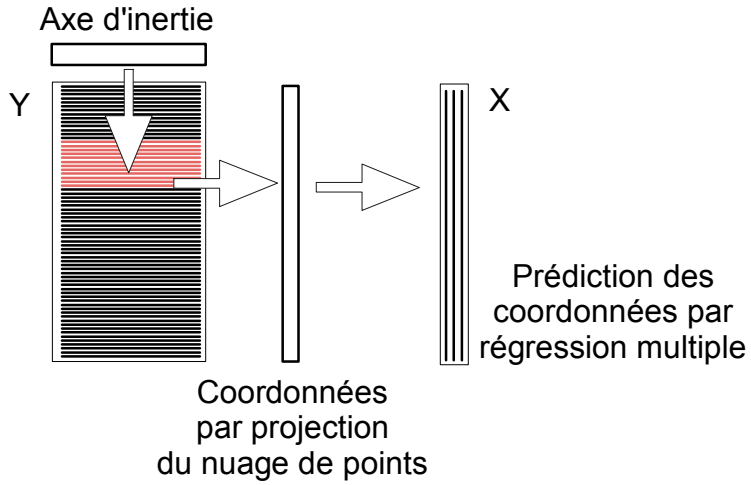
On peut imposer des contraintes emboîtées du type **B** sachant **A** ou $\mathbf{A} \cap \mathbf{B}^{\perp}$ (Yoccoz and Chessel 1988) et décomposer l'inertie dans des plans complexes (Sabatier et al. 1989). Ces méthodes ne sont pas d'un emploi ordinaire.

3.4. Analyse des Correspondances Non Symétriques

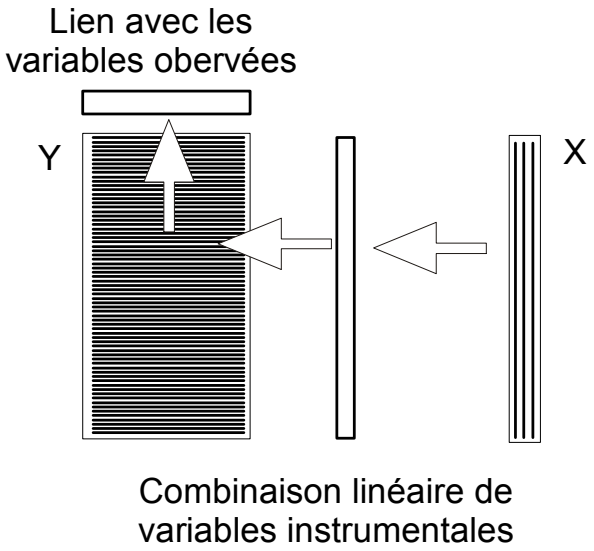
Pour avoir une vision coordonnée, on peut enregistrer que si l'AFC est l'analyse canonique entre les deux paquets d'indicatrices de classes, il existe deux ACPVI, une dans chaque sens, entre ces deux paquets d'indicatrices et qu'on obtient ainsi deux analyses non symétriques des correspondances. Origine dans Lauro et D'Ambra (1984),

introduction en écologie dans Gimaret-Carpentier et al. (1998). La meilleure présentation est celle faite en sciences sociales par Kroonenberg et Lombardo (1999).

Les ACPVI ont plusieurs modes d'interprétation qui peuvent être plus ou moins adaptées à certains types de couplage. Le premier est basé sur :



Par exemple, position des espèces avec des scores de variance unité, position des relevés à la moyenne et prédiction optimum de ces positions par régression sur les variables de milieu. Mais on a aussi le principe :

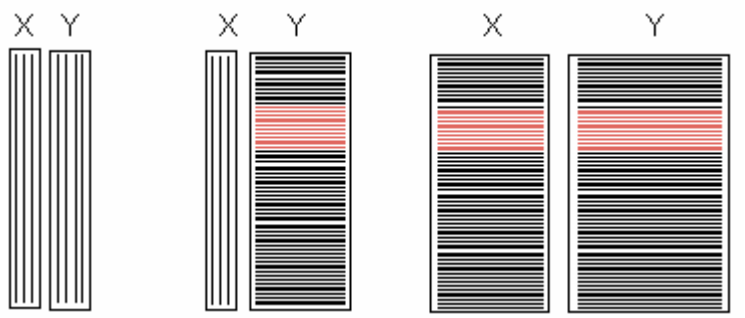


:

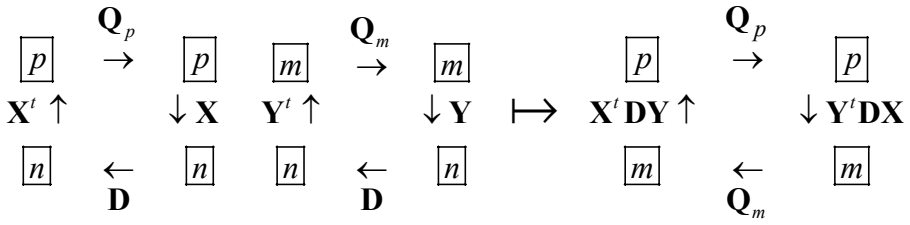
Par exemple, combinaison linéaire de variable de milieu optimisant la variance des moyennes par espèce, ou encore combinaison linéaire de variables instrumentales maximisant la somme des carrés de corrélation avec les variables dépendantes (Obadia 1978).

Il est important de reconnaître dans tous les cas que ces figures optimisent un critère donné sous des contraintes connues et que chaque figure est optimum de son point de vue. Il peut y avoir plusieurs points de vue. En première approche, il semble plus sûr de réserver ces méthodes lorsque les explicatives sont des facteurs contrôlés par l'utilisateur.

4. Stratégie de la co-inertie



La troisième stratégie est la seule à tolérer des dimensions quelconques des deux côtés et est pratiquement la seule utilisable si les variables de milieu sont qualitatives (c'est alors l'effectif des modalités qui définit la dimension du tableau **X**). C'est la plus simple puisqu'en gros elle fait une double analyse d'inertie des tableaux et garantit que les deux systèmes de coordonnées sont les plus cohérents possibles. Le schéma général est :



On peut étudier quatre exemples.

4.1. AFC des tableaux de profils écologiques

Une des plus anciennes pratiques issues de la théorie des profils écologiques consiste à faire l'AFC d'un tableau croisé. Les variables de milieu doivent être qualitatives. Prendre l'exemple la liste mafragh (Belair and Bencheikh-Lehocine 1987). Le tableau floristique est passé en présence-absence :

```
> veg01 = as.data.frame(apply(mafragh$flo, 2, function(x) as.numeric(x>0)))
```

Le tableau de milieu quantitatf est transformé en facteurs à 3 classes puis en disjonctif complet :

```
> milq = as.data.frame(apply(mafragh$mil,2,cut,breaks=3))
> mil01 = acm.disjonctif(milq)
```

Faire le produit de matrices et éditer une partie du résultat :

```
> tabeco = t(mil01)%*%as.matrix(veg01)
> tabeco [1:6,1:15]
```

```
          E1 E2 E3 E4 E5 E6 E7 E8 E9 E10 E11 E12 E13 E14 E15
Argile.X1  5  0  3  6  0  3  2  0  2   0   0   0   2   1   0
Argile.X2  4  1  2  7  0  1  2  0  5   2   5   8  23   3   6
Argile.X3 25  8 13 19  8  4  2  5  4   4   1   9  20   2   7
```

```
  Limon.X1  25  8 12 22  6  4  2  5  3   3   0   9  20   1   4
  Limon.X2   7  1  5  8  2  3  3  0  7   3   6   8  24   4   8
  Limon.X3   2  0  1  2  0  1  1  0  1   0   0   0   1   1   1
```

Il contient des profils écologiques bruts (nombre de stations contenant l'espèce dans chaque classe de milieu). On peut comparer chacun d'entre eux à la somme totale (nombre de présences d'espèces dans chaque classe de milieu) ou à la distribution des stations par classe de milieu. *Scirpus maritimus* (E1) a 34 présences dont 25 dans la classe 3 de la variable 1. Pour cette variable, le profil des stations est 11, 36, 50 :

```
> apply(mil01,2,sum)
    Argile.X1     Argile.X2     Argile.X3      Limon.X1      Limon.X2 . . .
           11            36            50            53            40
```

Pour cette même variable, le décompte des présences totales est 47, 274, 287 :

```
> apply(tabeco,1,sum)
    Argile.X1     Argile.X2     Argile.X3      Limon.X1      Limon.X2 . . .
           47           274           287           311           275
```

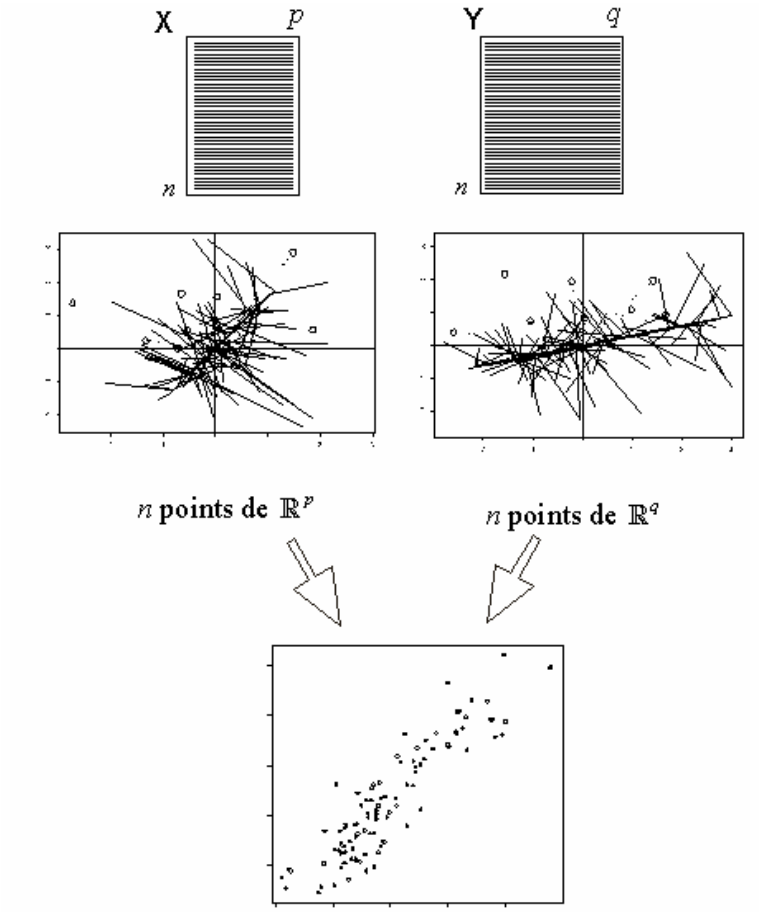
P. Romane (1972) a proposé de comparer tous les profils sur toutes les variables en faisant l'AFC de ce tableaux, idée qui sera "retrouvée" plus tard (Montaña and Greig-Smith 1990).

```
> dudi.coa(as.data.frame(tabeco),scan=F)$eig
 [1] 1.359e-01 5.042e-02 2.398e-02 1.804e-02 1.474e-02 1.297e-02 9.593e-03
 [8] 8.796e-03 7.195e-03 6.246e-03 5.511e-03 3.972e-03 3.594e-03 2.857e-03
[15] 2.489e-03 2.392e-03 1.358e-03 1.138e-03 8.223e-04 6.225e-04 2.882e-04
[22] 4.559e-05

> coa1 = dudi.coa(veg01,scann=F)
> for (j in 1:ncol(milq)) milq[,j] = as.factor(milq[,j])
> acm1 = dudi.acm (milq, scann=F, row.w = coa1$lw)
> coinertia(coa1,acm1,scann=F)$eig
 [1] 1.359e-01 5.042e-02 2.398e-02 1.804e-02 1.474e-02 1.297e-02 9.593e-03
 [8] 8.796e-03 7.195e-03 6.246e-03 5.511e-03 3.972e-03 3.594e-03 2.857e-03
[15] 2.489e-03 2.392e-03 1.358e-03 1.138e-03 8.223e-04 6.225e-04 2.882e-04
[22] 4.559e-05
```

On obtient un résultat complet par l'analyse de la co-inertie de l'AFC du tableau en présence-absence et de l'ACM du tableau de milieu pondéré par les poids des lignes de la précédente (Mercier et al. 1992). De même l'ACP non centrée de la page 6 donne les résultats de l'analyse de co-inertie de l'ACP normée du tableau et de l'ACP du tableau faunistique passé en fréquences par espèce.

4.2. Analyse inter-batteries

A l'origine de l'analyse de co-inertie, on trouve l'analyse inter-batterie (résultats sur les mêmes individus de deux batteries de tests psychotechniques) de Tucker (1958). Au lieu de chercher deux combinaisons de variables maximisant leur corrélation, on cherche deux combinaisons de variables maximisant leur covariance sous la contrainte $\sum_{j=1}^{p} \alpha_j^2 = 1$ et $\sum_{j=1}^{q} \beta_j^2 = 1$, p et q étant le nombre des variables de chaque tableau. On optimise ainsi un produit car $cov^2(\mathbf{x},\mathbf{y}) = cor^2(\mathbf{x},\mathbf{y}) var(\mathbf{x}) var(\mathbf{y})$. $cor^2(\mathbf{x},\mathbf{y})$ est maximisée dans l'analyse canonique, $var(\mathbf{x})$ est maximisée dans l'analyse du premier tableau et $var(\mathbf{y})$ est maximisée dans celle du second. La co-inertie est un compromis entre analyse canonique et double analyse simple.



Les propriétés dérivent directement du couplage simple de deux ACP normées :

$$\begin{array}{ccccccc} & \mathbf{I}_p & & & \mathbf{I}_m & & \\ \boxed{p} & \rightarrow & \boxed{p} & \boxed{m} & \rightarrow & \boxed{m} & \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} & \mathbf{Y}^t \uparrow & & \downarrow \mathbf{Y} & \mapsto \\ \boxed{n} & \leftarrow & \boxed{n} & \boxed{n} & \leftarrow & \boxed{n} & \\ & \frac{1}{n}\mathbf{I}_n & & & \frac{1}{n}\mathbf{I}_n & & \end{array} \qquad \begin{array}{ccc} & \mathbf{I}_p & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \frac{1}{n}\mathbf{X}^t\mathbf{Y} \uparrow & & \downarrow \frac{1}{n}\mathbf{Y}^t\mathbf{X} \\ \boxed{m} & \leftarrow & \boxed{m} \\ & \mathbf{I}_m & \end{array}$$

Le schéma de principe se résume dans la figure précédente.

4.3. AFC des tableaux de Burt croisés

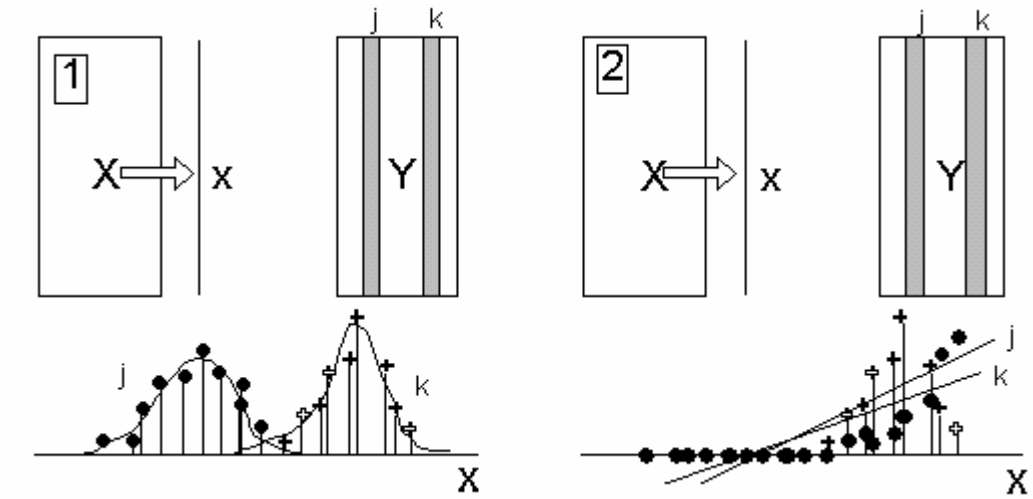
Un autre cas fondamental est celui de l'AFC des tableaux de Burt croisés que Cazes (1980, 1981) a appelé analyse canonique sur variables qualitatives et qui est en fait l'analyse de co-inertie de deux analyses des correspondances multiples. L'analyse de co-inertie est donc un procédé de couplage de tableaux très général. Son principe tient dans la figure ci-dessus.

Deux tableaux définissent deux nuages de points. Choisir un axe dans chaque espace et projeter les nuages définissent deux systèmes de coordonnées. La covariance des deux systèmes définit la co-inertie des deux axes. Trouver les deux axes qui maximisent la

co-inertie, c'est trouver les axes de co-inertie des deux nuages. Appliqué à tout couple d'analyses utilisant la même pondération des individus (Dolédec and Chessel 1994), ce principe est aussi facile d'emploi que l'analyse d'un tableau.

4.4. Analyse des niches écologiques

Pour croiser un tableau floro-faunistique et un tableau de variables mésologiques, on pensera surtout au modèle de liaison qu'on désire mettre en évidence :



A gauche : Courbes de réponse non linéaires des espèces sur les gradients environnementaux. Recherche des gradients séparant les profils. Cas typique de l'utilisation de l'ACC (AFCVI) ou de la co-inertie sur une méthode utilisant les profils espèces. A droite : Courbes de réponses monotones. Recherche des facteurs augmentant ou limitant l'abondance des espèces. Cas typique de l'utilisation de l'ACPVI ou de la co-inertie sur ACP.

Reprendre l'exemple de l'introduction. Faire le couplage de deux ACP.

```
> data(doubs)
> names(doubs)
[1] "mil" "poi" "xy"
> pca1 = dudi.pca(doubs$mil)
Select the number of axes: 2
> pca2 = dudi.pca (doubs$poi,scal=F)
Select the number of axes: 2
> pcapca = coinertia(pca1,pca2)
Select the number of axes: 2

Coinertia analysis
call: coinertia(dudiX = pca1, dudiY = pca2)
class: coinertia dudi

$rank (rank)     : 11
$nf (axis saved) : 2
$RV (RV coeff)   : 0.4506

eigen values: 119 13.87 0.7566 0.5278 0.2709 ...

  vector length mode    content
1 $eig   11     numeric eigen values
2 $lw    27     numeric row weigths (crossed array)
3 $cw    11     numeric col weigths (crossed array)

   data.frame nrow ncol content
1  $tab       27   11   crossed array (CA)
2  $li        27   2    Y col = CA row: coordinates
3  $l1        27   2    Y col = CA row: normed scores
4  $co        11   2    X col = CA column: coordinates
5  $c1        11   2    X col = CA column: normed scores
```

```
6  $lX        30   2    row coordinates (X)
7  $mX        30   2    normed row scores (X)
8  $lY        30   2    row coordinates (Y)
9  $mY        30   2    normed row scores (Y)
10 $aX        2    2    axis onto co-inertia axis (X)
11 $aY        2    2    axis onto co-inertia axis (Y)
```

Une analyse de co-inertie est un objet des classes 'dudi' et 'coi'. Les composantes de la liste ont une signification générale mais pourra prendre dans chaque type de couplage une signification particulière. En utilisant les notations :

$$\begin{array}{ccccc} \boxed{p} & \overset{\mathbf{Q}}{\rightarrow} & \boxed{p} & \overset{\mathbf{R}}{\rightarrow} & \boxed{q} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} & & \mathbf{Y}^t \uparrow \quad \downarrow \mathbf{Y} \\ \boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n} & \underset{\mathbf{D}}{\leftarrow} & \boxed{n} \end{array} \quad \mapsto \quad \begin{array}{ccc} \boxed{p} & \overset{\mathbf{Q}}{\rightarrow} & \boxed{p} \\ \mathbf{X}^t\mathbf{D}\mathbf{Y} \uparrow & & \downarrow \mathbf{Y}^t\mathbf{D}\mathbf{X} \\ \boxed{q} & \underset{\mathbf{R}}{\leftarrow} & \boxed{q} \end{array}$$

les éléments de la liste sont :

tab la matrice $\mathbf{Y}^t\mathbf{D}\mathbf{X}$ des produits scalaires entre colonnes de $\mathbf{X}$ et colonnes de $\mathbf{Y}$. Les éléments peuvent être des moyennes, des covariances, des corrélations, des cosinus suivant les tableaux d'origine.

cw la métrique diagonale de $\mathbb{R}^p$ (poids des colonnes de $\mathbf{X}$)

lw la métrique diagonale de $\mathbb{R}^q$ (poids des colonnes de $\mathbf{Y}$)

eig les valeurs propres de l'analyse, carrés des produits scalaires (en général des covariances) entre les coordonnées de co-inertie de même rang

c1 les axes de co-inertie dans $\mathbb{R}^p$, vecteurs normés en colonnes

l1 les xes de co-inertie dans $\mathbb{R}^q$, vecteurs normés en colonnes

co les produits scalaires entre colonnes de $\mathbf{X}$ et coordonnées de co-inertie dans $\mathbb{R}^q$

li les produits scalaires entre colonnes de $\mathbf{Y}$ et coordonnées de co-inertie dans $\mathbb{R}^p$

lX les coordonnées de co-inertie dans $\mathbb{R}^p$ donnant les projections des lignes de $\mathbf{X}$ sur les axes de co-inertie dans $\mathbb{R}^p$

lY les coordonnées de co-inertie dans $\mathbb{R}^q$ donnant les projections des lignes de $\mathbf{Y}$ sur les axes de co-inertie dans $\mathbb{R}^q$

mX les scores normés obtenus en normant dans $\mathbb{R}^n$ les coordonnées de lX

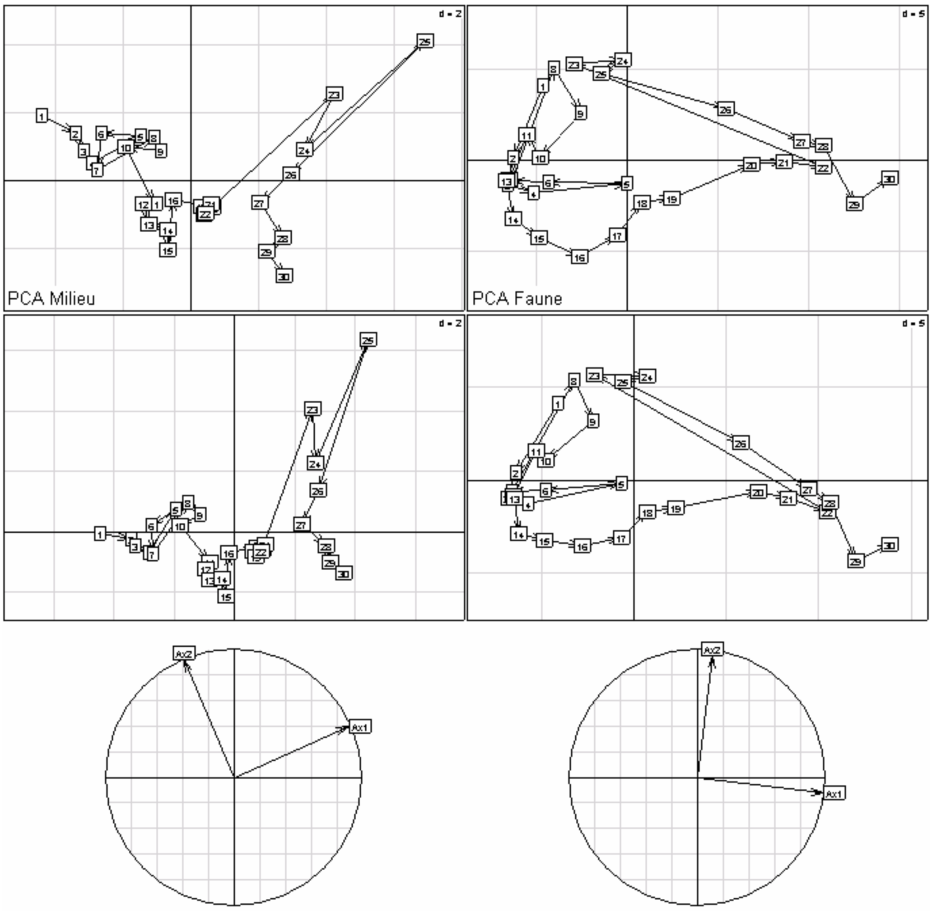
mY les scores normés obtenus en normant dans $\mathbb{R}^n$ les coordonnées de lY

aX les coordonnées de la projection des axes d'inertie dans $\mathbb{R}^p$ (analyse initiale de $\mathbf{X}$) sur les axes de co-inertie dans $\mathbb{R}^p$

aY les coordonnées de la projection des axes d'inertie dans $\mathbb{R}^q$ (analyse initiale de $\mathbf{Y}$) sur les axes de co-inertie dans $\mathbb{R}^q$

```
par(mfcol=c(3,2))
s.traject(pca1$li,clab=0, sub="PCA Milieu",csub=2)
s.label(pca1$li,clab=1,add.plot=T)
s.traject(pcapca$lX,clab=0)
s.label(pcapca$lX,clab=1,add.plot=T)
s.corcircle(pcapca$aX)
s.traject(pca2$li,clab=0, sub="PCA Faune",csub=2)
s.label(pca2$li,clab=1,add.plot=T)
s.traject(pcapca$lY,clab=0)
```

```
s.label(pcapca$lY,clab=1,add.plot=T)
s.corcircle(pcapca$aY)
```



Les plans de co-inertie sont pratiquement les plans d'inertie : les positions sont réajustées par une simple rotation. Ce peut être beaucoup plus compliqué dans d'autres exemples.

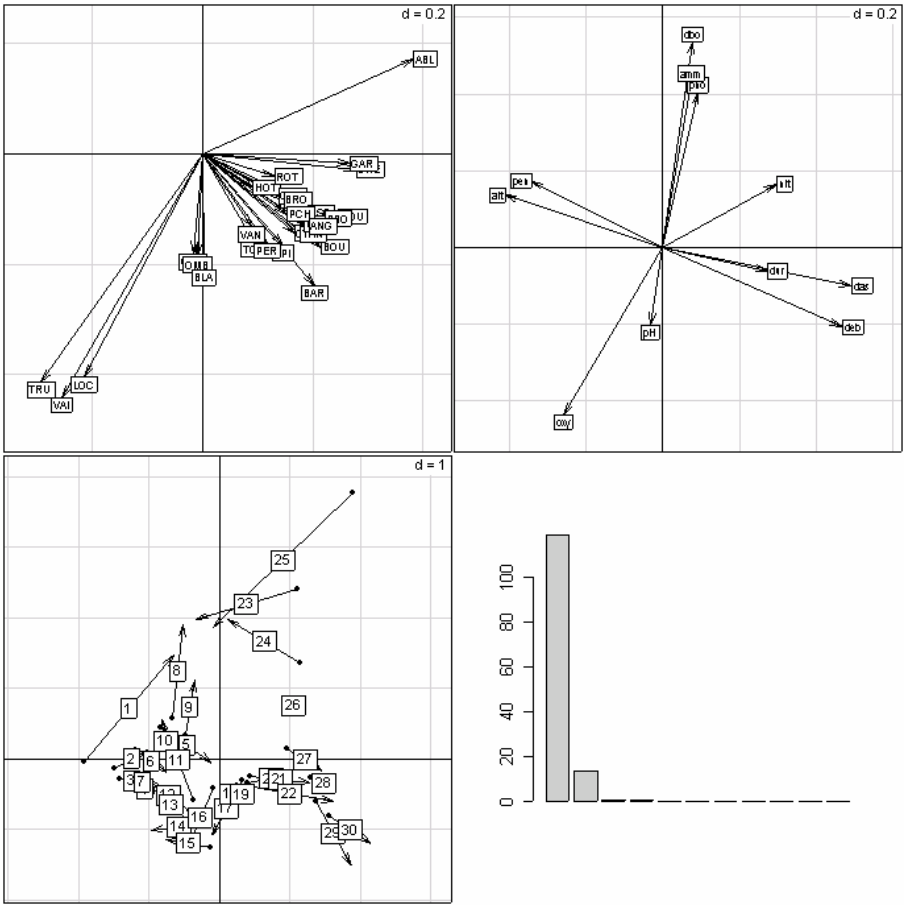
```
> summary(pcapca)

Eigenvalues decomposition:
     eig  covar   sdX   sdY   corr
1 119.02 10.910 2.326 6.423 0.7302
2  13.87  3.724 1.685 2.864 0.7718

Inertia & coinertia X:
   inertia   max  ratio
1    5.412 6.322 0.8561
12   8.251 8.553 0.9647
```

L'inertie projetée sur le plan de co-inertie vaut 96% du maximum possible

```
Inertia & coinertia Y:
   inertia   max  ratio
1    41.25 42.75 0.9650
12   49.45 50.90 0.9714
```

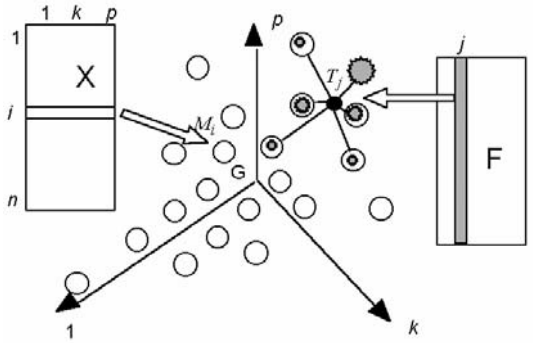
L'inertie projetée sur le plan de co-inertie vaut 97% du maximum possible

```
RV:
 0.4506 # voir le RV dans les méthodes k-tableaux

par(mfrow=c(2,2))
s.arrow(pcapca$l1,clab=0.75)
s.arrow(pcapca$c1,clab=0.75)
s.match (pcapca$mX,pcapca$mY,clab=1)
barplot(pcapca$eig,col=grey(0.8))
```



En haut, les projections des vecteurs des bases canoniques dans chaque espace, équivalent des biplots (Gabriel 1971, 1981) des ACP séparées. En bas, la superposition des deux nuages de points après normalisation sur chaque axe pour exprimer la partie corrélation de la co-inertie. Les valeurs propres des analyses de co-inertie sont en général très structurées : il faut en effet réunir de l'inertie projetée cohérente dans les deux espaces et les axes ayant un sens sont très apparents.

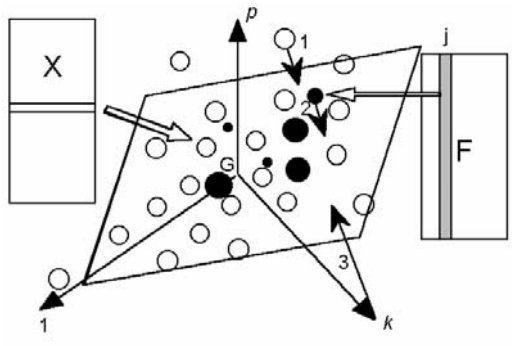
L'analyse OMI (**O**utlying **M**ean **I**ndex) est aussi une analyse de co-inertie. Son principe repose sur la figure :



*Eléments de base dans une analyse OMI dite aussi 'niche'. Les lignes du tableau **X** définissent un nuage de points et chaque taxon (colonne de **F** est une pondération de ces points qui définit un centre de gravité (position moyenne du taxon dans l'espace).*

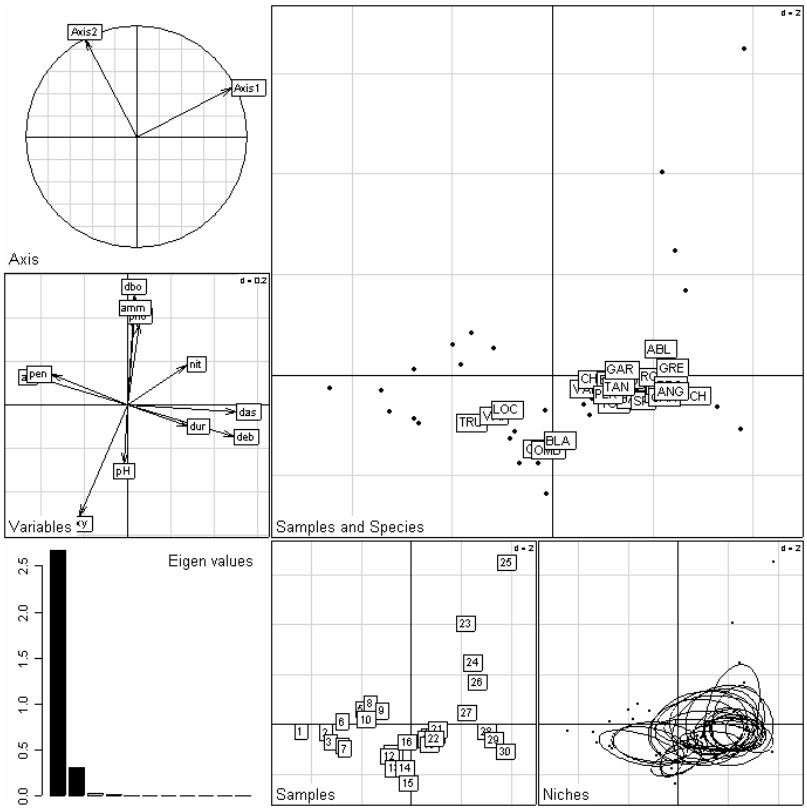
Les stations sont d'abord considérées sans référence avec ce qu'on y trouve. Ce sont simplement des points d'échantillonnage du milieu. Les variables de milieu définissent un nuage de *n* points de $\mathbb{R}^p$. Chaque espèce est une distribution de fréquences qui a un

centre de gravité (centre de la niche) et on fait l'ACP non centrée des points moyens. Sur les plans, on reprojette ensuite les vecteurs de la base canonique (variables), les relevés de départ et les positions moyennes des espèces. On peut ainsi prendre en compte les deux types de relation au milieu, séparer les niches ou mettre en évidence les parties de l'espace mésologique où se concentrent des groupes d'espèces (Dolédec et al. 2000).



*Représentation simultanée dans une analyse OMI. 1 - Les lignes du tableau **X** (relevés) 2 - les colonnes du tableau **F** (position moyenne du taxon dans l'espace) 3 - les vecteurs de la base canonique (variables) sont positionnés par projection orthogonale sur un même plan.*

```
> plot(niche(pca1,doubs$poi))
Select the number of axes: 2
```



L'interprétation est la même. Cette analyse repose sur une co-inertie entre l'analyse du tableau milieu et l'ACP centrée du tableau sites-espèces passé en pourcentage par espèce (tableau des profils taxons).

Refaire l'exercice avec la liste trichometeo (Usseglio-Polatera and Auda 1987) :

```
data(trichometeo)
names(trichometeo)
[1] "fau"   "meteo" "cla"
dim(trichometeo$meteo)
[1] 49 11
dim(trichometeo$fau)
[1] 49 17
```

49 nuits de piégeage d'un piège lumineux au bord du Rhône. 11 variables météorologiques. 17 espèces de trichoptères.

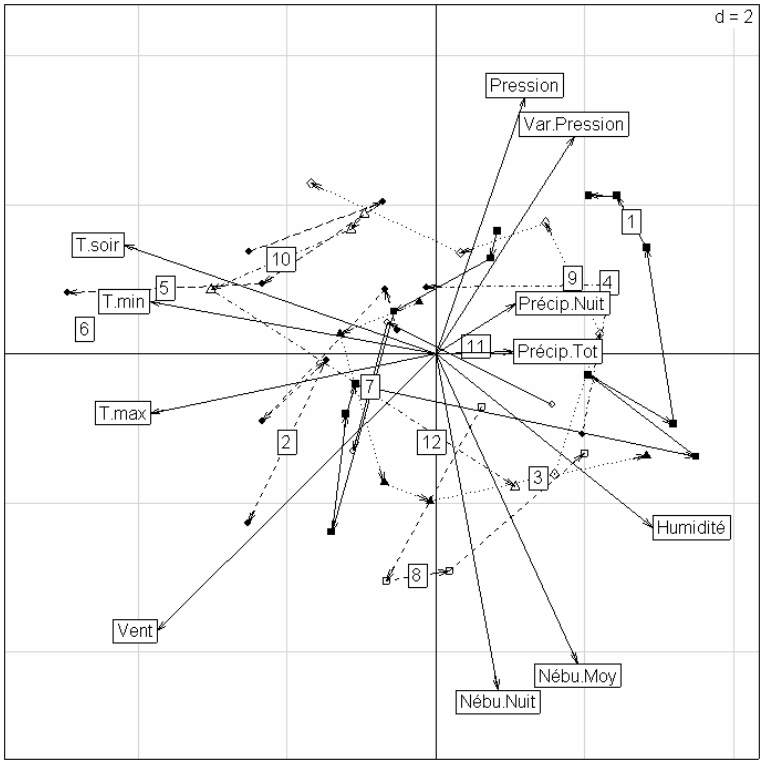
```
trichometeo$cla
 [1] 1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  3  3  3  3  3  4  4  4
[26] 4  5  5  5  5  6  7  7  7  8  8  8  8  9  9  9  9  9  10 10 10 10 11 12
Levels: 1 2 3 4 5 6 7 8 9 10 11 12
```

12 groupes de nuits consécutives au cours des mois d'été.

```
faulog <- log(trichometeo$fau + 1)
```

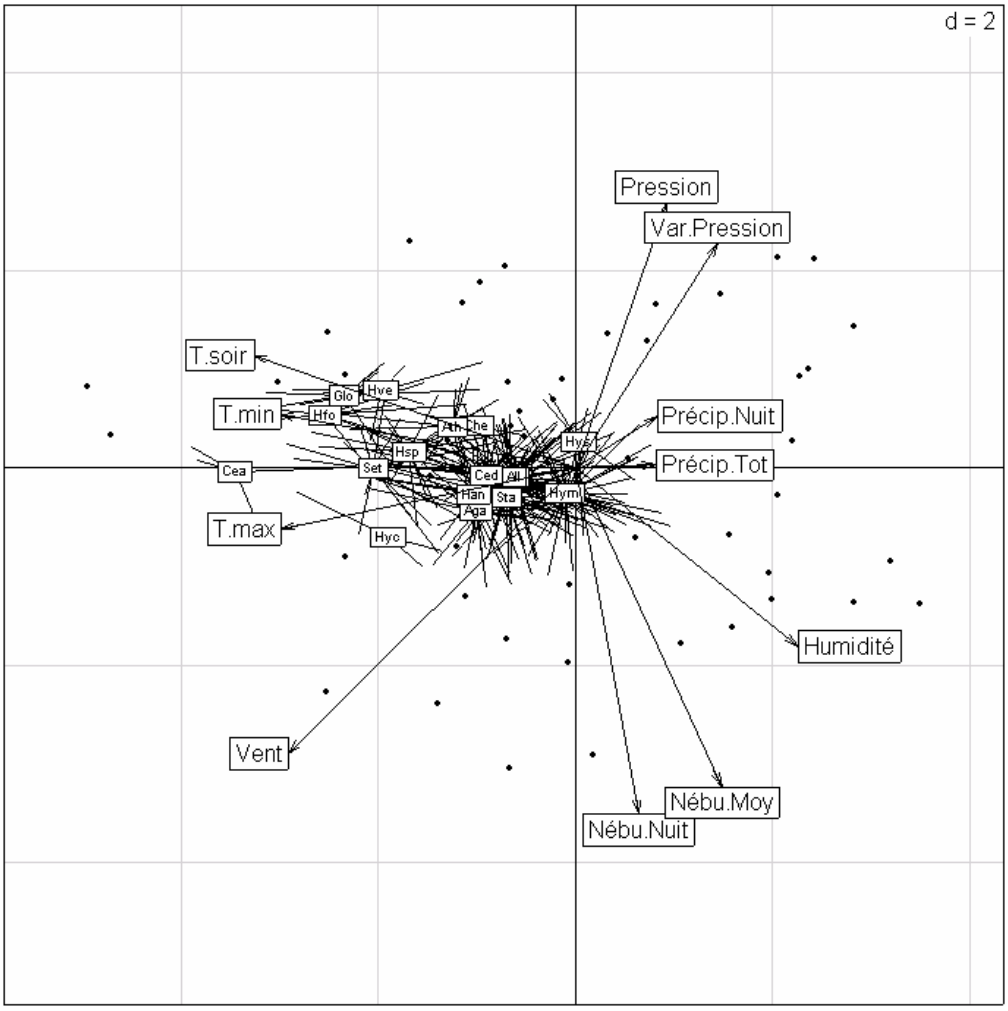
L'abondance des espèces est transformée par :

```
pca1 <- dudi.pca(trichometeo$meteo, scan = FALSE)
niche1 <- niche(pca1, faulog, scan = FALSE)
s.traject(niche1$ls,trichometeo$cla)
s.arrow(9 * niche1$c1, clab = 1, add.p = TRUE)
```



Biplot météorologique : les trajectoires tournent dans le sens trigonométriques. La pression augmente (beau temps), la température devient élevée (canicule), le vent se lève et l'orage éclate (humidité précipitations). C'est le cycle de ma météorologie estivale de la région lyonnaise.

```
s.label(niche1$ls, clab = 0)
s.distri(niche1$ls, faulog, clab = 0.6, add.p = TRUE, cell = 0, csta = 0.3)
s.arrow(9 * niche1$c1, clab = 1, add.p = TRUE)
```



L'émergence des adultes d'éphéméroptères se fait globalement à une position précise du cycle météorologique. L'optimisation de la somme des carrés des écarts à l'origine (facteur 1) a placé toutes les espèces d'un même côté.

Le couplage des tableaux est donc un univers assez complexe. Chacun des tableaux supporte sa propre analyse. Le couplage des deux peut se faire suivant trois principes généraux. Le schéma retenu pour le couple peut enfin être interprété de diverses façons. Il s'en suit une grande diversité d'expression. Fondamentalement, il y a plusieurs manières de voir et d'exprimer l'essentiel. Prévoir une réflexion préalable à toute analyse pour définir les objectifs et intégrer les propriétés connues des données, éviter les conseils impérieux de ceux qui savent ce qu'il faut faire et faire des essais préalables sans se soucier d'une expression définitive. Une méthode se révèle systématiquement bonne dans certains cas et mauvaise dans d'autres. Quand les données ne sont pas fameuses, éviter enfin "l'acharnement méthodologique". *What is not acceptable is to rummage around trying methods until the desired significance (or lack thereof) is obtained (Green 1993).*

Références

Afriat, S. N. 1957. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences 53:800-816.

Barkham, J. P., and J. M. Norris. 1970. Multivariate procedures in an investigation of vegetation and soil relations of two beach woodlands, Costwold Hills, England. Ecology **51**:630-639.

Belair, G. d., and M. Bencheikh-Lehocine. 1987. Composition et déterminisme de la végétation d'une plaine côtière marécageuse : La Mafragh (Annaba, Algérie). Bulletin d'Ecologie **18**:393-407.

Bergougnan, D., and C. Couraud. 1982. Pratique de la discrimination barycentrique. Les Cahiers de l'Analyse des Données **7**:341-354.

Birks, H. J. B., and H. A. Austin. 1992. An annoted bibliography of canonical correspondence analysis and related constrained ordination methods (1986-1991). Botanical Institute, Allégaten 41, N-5007 Bergen, Norway.

Blondel, J., D. Chessel, and B. Frochot. 1988. Bird species impoverishment, niche expansion, and density inflation in mediterranean island habitats. Ecology **69**:1899-1917.

Cazes, P. 1980. L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives. II Questionnaires : variantes des codages et nouveaux calculs de contributions. Les Cahiers de l'Analyse des Données **5**:145-161 & 387-406.

Cazes, P. 1981. L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. III Codage simultanné de variables qualitatives et quantitatives. IV Cas modèles. Les Cahiers de l'Analyse des Données **6**:9-18 & 135-143.

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

Chessel, D., J. D. Lebreton, and N. Yoccoz. 1987. Propriétés de l'analyse canonique des correspondances. Une utilisation en hydrobiologie. Revue de Statistique Appliquée **35**:55-72.

Dagnelie, P. 1965. L'étude des communautés végétales par l'analyse statistique des liaisons entre les espèces et les variables écologiques : principes fonadamentaux, un exemple. Biometrics **21**:345-361 & 890-907.

Dolédec, S., and D. Chessel. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology **31**:277-294.

Dolédec, S., D. Chessel, and C. Gimaret. 2000. Niche separation in community analysis: a new method. Ecology **81**:2914-1927.

Estève, J. 1978. Les méthodes d'ordination : éléments pour une discussion. Pages 223-250 *in* J. M. Legay and R. Tomassone, editors. Biométrie et Ecologie. Société Française de Biométrie, Paris.

Fisher, R. A. 1940. The precision of discriminant functions. Annals of Eugenics **10**:422-438.

Gabriel, K. R. 1971. The biplot graphical display of matrices with application to principal component analysis. Biometrika **58**:453-467.

Gabriel, K. R. 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. Pages 147-174 *in* V. Barnett, editor. Interpreting multivariate data. John Wiley and Sons, New York.

Gimaret-Carpentier, C., D. Chessel, and J. P. Pascal. 1998. Non-symmetric correspondence analysis: an alternative for community analysis with species occurrences data. Plant Ecology **138**:97-112.

Gittins, R. 1985. Canonical analysis, a review with applications in ecology. Springer-Verlag, Berlin.

Godron, M., P. Daget, L. Emberger, E. Le Floch, J. Poissonet, C. Sauvage, and J. P. Wacquant. 1968. Relevé méthodique de la végétation et du milieu. Editions du CNRS, Paris.

Gounot, M. 1969. Méthodes d'étude quantitative de la végétation. Masson, Paris.

Green, R. H. 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve Molluscs of Central Canada. Ecology **52**:543-556.

Green, R. H. 1993. Relating two sets of variables in environmental studies. Pages 149-163 *in* G. P. Patil and C. R. Rao, editors. Multivariate environmental statistics. North-Holland, Amsterdam.

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

Hill, M. O. 1977. Use of simple discriminant functions to classify quantitative phytosociological data. Pages 181-199 *in* E. Diday, editor. Proceedings of the First International Symposium on Data Analysis and Informatics. INRIA Rocquencourt, France.

Hotelling, H. 1936. Relations between two sets of variates. Biometrika **28**:321-377.

Kroonenberg, P. M., and R. Lombardo. 1999. Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. Multivariate Behavioral Research **34**:367-396.

Lauro, N., and L. D'Ambra. 1984. L'analyse non symétrique des correspondances. Pages 433-446 *in* E. C. Diday, editor. Data Analysis and Informatics III. Elsevier, North-Holland.

Lawley, D. N. 1938. A generalization of Fisher' Z-test. Biometrika **30**:180 sqq.

Lebreton, J. D., D. Chessel, R. Prodon, and N. Yoccoz. 1988a. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. Acta Œcologica, Œcologia Generalis **9**:53-67.

Lebreton, J. D., M. Richardot-Coulet, D. Chessel, and N. Yoccoz. 1988b. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances . II Variables de milieu qualitatives. Acta Œcologica, Œcologia Generalis **9**:137-151.

Lebreton, J. D., R. Sabatier, G. Banco, and A. M. Bacou. 1991. Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure-activity and species- environment relationships. Pages 85-114 *in* J. Devillers and W. Karcher, editors. Applied Multivariate Analysis in SAR and Environmental Studies. Kluwer Academic Publishers.

McIntosh, R. P. 1958. Plant communities. Science **128**:115-120.

Mercier, P., D. Chessel, and S. Dolédec. 1992. Complete correspondence analysis of an ecological profile data table: a central ordination method. Acta Œcologica **13**:25-44.

Montaña, C., and P. Greig-Smith. 1990. Correspondence analysis of species by environmental variable matrices. Journal of Vegetation Science **1**:453-460.

Obadia, J. 1978. L'analyse en composantes explicatives. Revue de Statistique Appliquée **24**:5-28.

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

Perrière, G., and J. Thioulouse. 2002. Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins. Computer Methods and Programs in Biomedicine **in press**.

Pillai, K. C. S. 1955. Some new test criteria in multivariate analysis. Annals of Mathematical Statistics **26**:117-121.

Pontier, J., A. B. Dufour, and M. Normand. 1990. Le modèle euclidien en analyse des données. SMA, édition Ellipses, Bruxelles.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya, A **26**:329-359.

Romane, F. 1972. Utilisation de l'analyse multivariable en Phytoécologie. Investigaciõn pesquera **36**:131-139.

Sabatier, R., J. D. Lebreton, and D. Chessel. 1989. Principal component analysis with instrumental variables as a tool for modelling composition data. Pages 341-352 *in* R. Coppi and S. Bolasco, editors. Multiway data analysis. Elsevier Science Publishers B.V., North-Holland.

Ter Braak, C. J. F. 1986. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167-1179.

Ter Braak, C. J. F. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio **69**:69-77.

Ter Braak, C. J. F. 1990. Interpreting canonical correlation analysis through biplots of structure correlations and weights. Psychometrika **55**:519-531.

Ter Braak, C. J. F., and P. F. M. Verdonschot. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences **57**:255-289.

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

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

Tucker, L. R. 1958. An inter-battery method of factor analysis. Psychometrika **23**:111-136.

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

Verneaux, J. 1973. Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie. Thèse d'état, Besançon.

Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Reviews **42**:207-264.

Yoccoz, N., and D. Chessel. 1988. Ordination sous contraintes de relevés d'avifaune : élimination d'effets dans un plan d'observations à deux facteurs. Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D **III**:307 : 189-194.