Fiche de Biostatistique

Couplages de tableaux

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

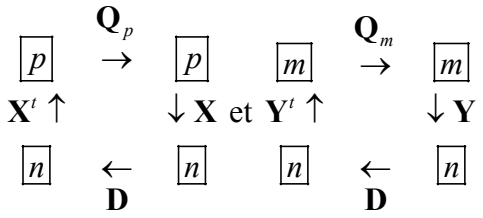
Résumé

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 :

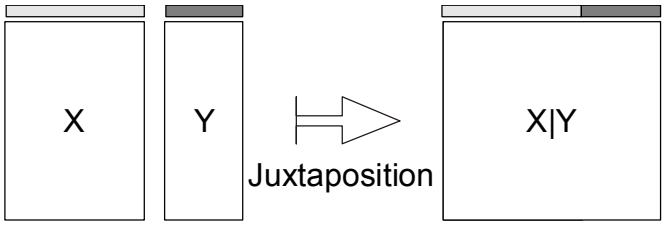


Il y a trois stratégies principales d'association de deux triplets. Elles recouvrent, à cause des nombreux jeux de paramètres, un vaste ensemble de pratiques. Il suffit de saisir ces trois principes pour faire un choix, voire pour construire des associations originales dont on peut avoir besoin.

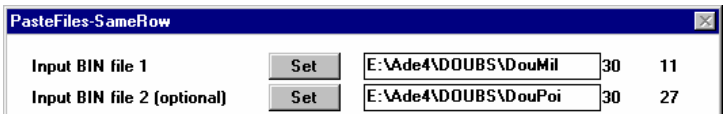
Le couplage de deux tableaux de données est une opération fondamentale en écologie statistique. On dispose d'une énorme littérature sur le sujet. Parmi les bases historiques on doit rappeler quelques opérations fondamentales.

1.1. Juxtaposition

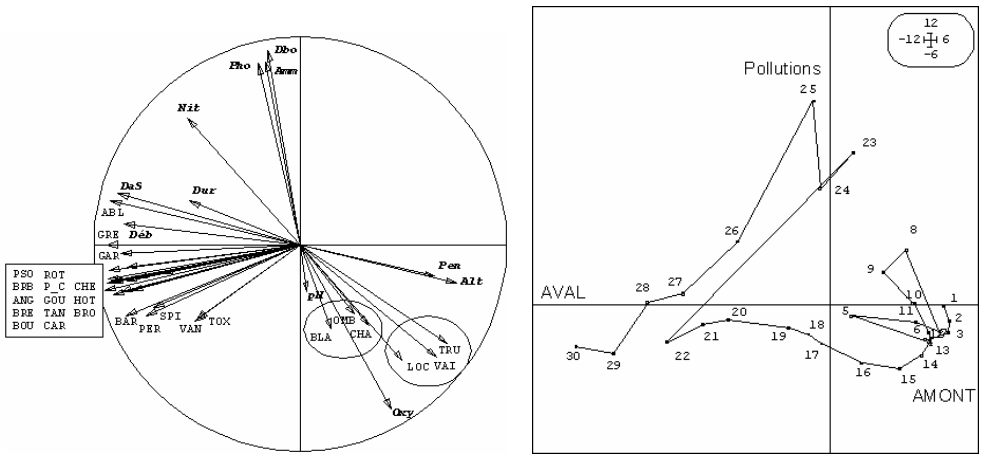
Deux tableaux ayant les mêmes lignes sont simplement accolés pour former un nouveau tableau qui appelle une analyse simple. La voie a été ouverte par P. Dagnélie [1].



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 de [2] (Carte Doubs dans ADE-4) :



On soumet le résultat à une ACP normée :

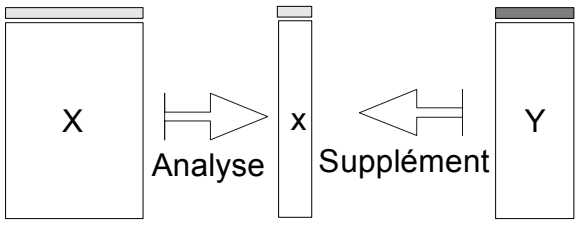


$$\begin{array}{ccccccc} \boxed{p} & \xrightarrow{\mathbf{I}_p} & \boxed{p} & \boxed{m} & \xrightarrow{\mathbf{I}_m} & \boxed{m} & \\ \mathbf{X}_*^t \uparrow & & \downarrow \mathbf{X}_* & \mathbf{Y}_*^t \uparrow & & \downarrow \mathbf{Y}_* & \mapsto \\ \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} & \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} & \end{array} \qquad \begin{array}{ccc} \boxed{p+m} & \xrightarrow{\mathbf{I}_{p+m}} & \boxed{p+m} \\ \begin{bmatrix} \mathbf{X}_*^t \\ \mathbf{Y}_*^t \end{bmatrix} \uparrow & & \downarrow [\mathbf{X}_* \quad \mathbf{Y}_*] \\ \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} \end{array}$$

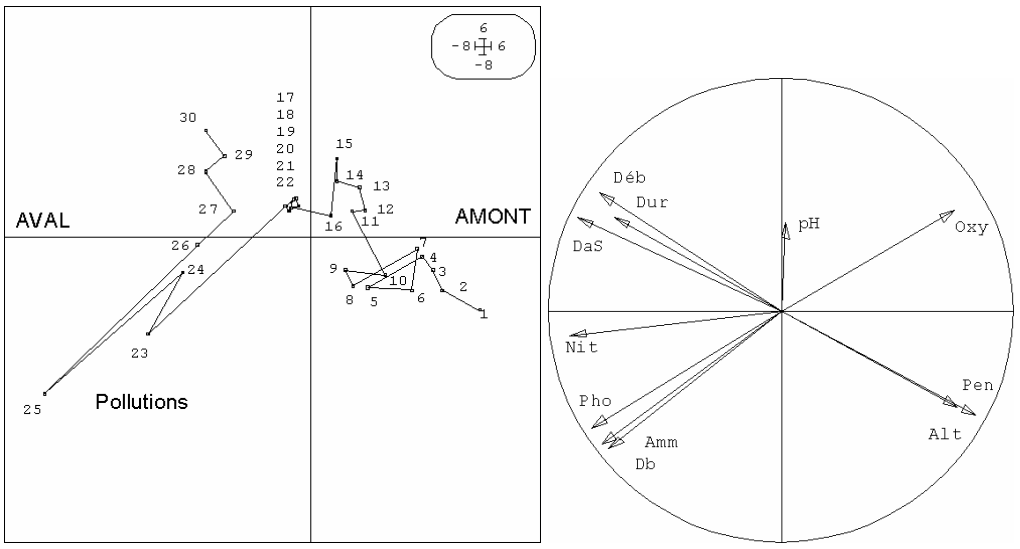
$\mathbf{X}_*$ et $\mathbf{Y}_*$ désignent les tableaux de variables normalisées. C'est le premier pas vers l'analyse factorielle multiple (AFM voir la fiche K-tableaux). Cette approche ne fonctionne que si les inerties des deux tableaux sont comparables. Si l'un des deux l'emporte largement, il imposera 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

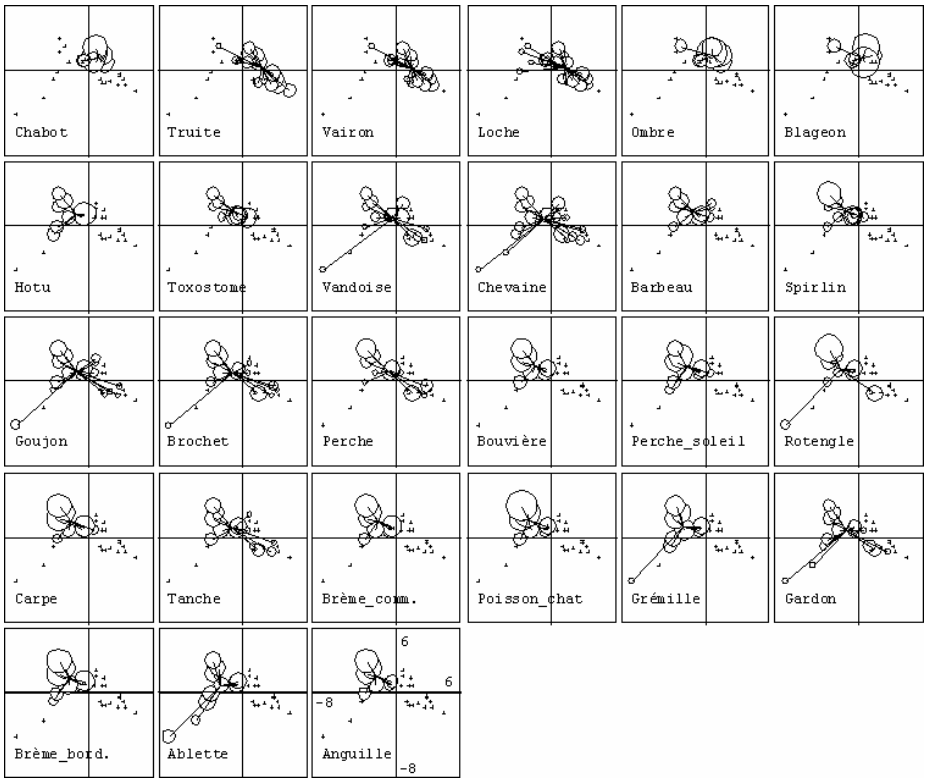
On pratique l'analyse d'un des deux tableaux et on introduit dans l'interprétation des éléments issus de l'autre. R. H. Whittaker, théoricien fondateur de l'analyse des données écologiques distingue ainsi l' ordination directe et l'ordination indirecte [3].



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



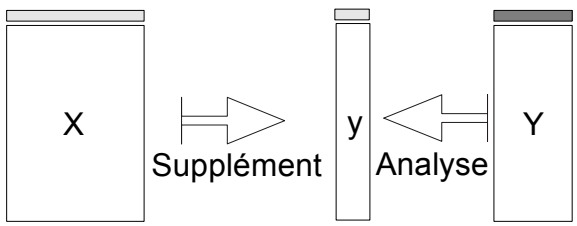
PCA: Correlation matrix PCA

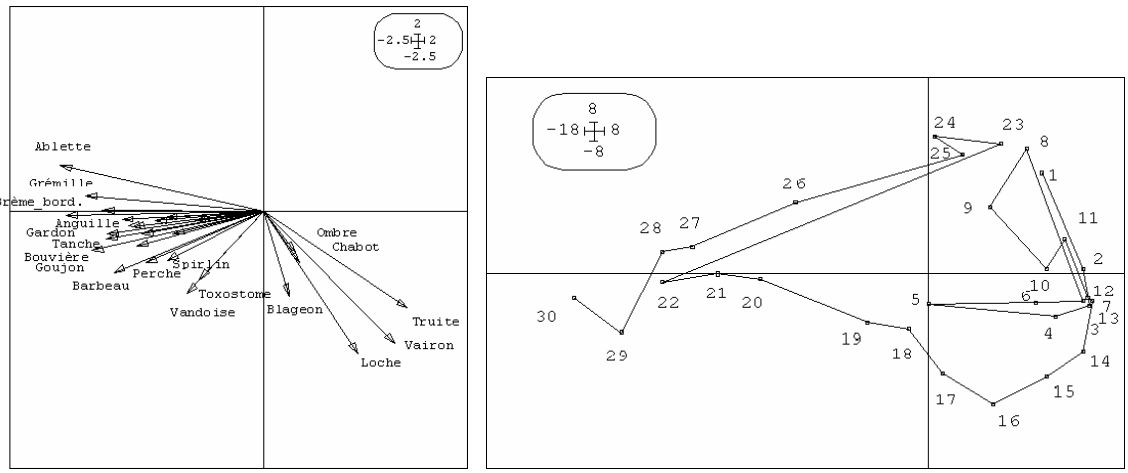


ScatterDistri: Frequencies ScatterDistri: Stars

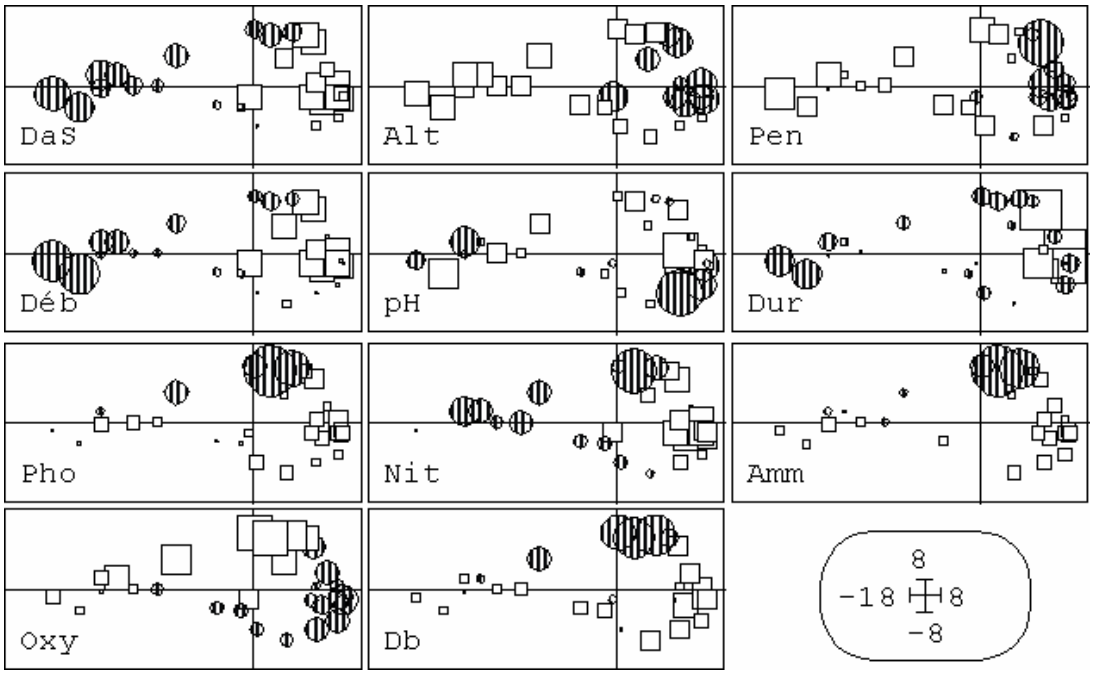
Ceci est une représentation d'informations supplémentaires.

Dans la seconde (*Ordination indirecte*[4]), on analyse le tableau relevés-espèces et on étudie la répartition des variables.





PCA: Covariance matrix PCA



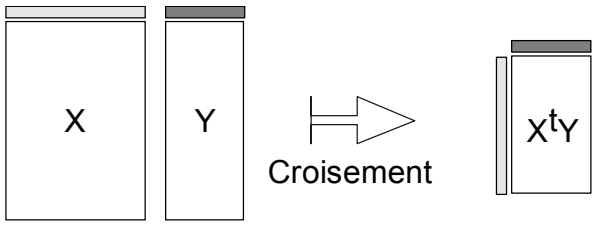
Scatters: Values

Ceci est une représentation d'informations supplémentaires.

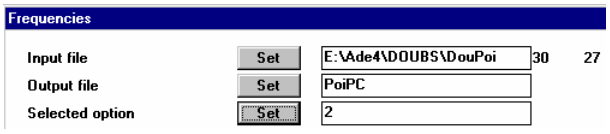
1.3. Croisement

Le tableau croisé utilise simplement un produit de matrices.

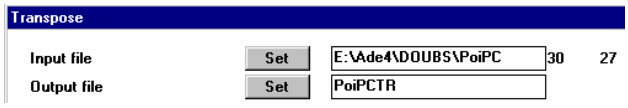
Analyse d'un tableau croisé



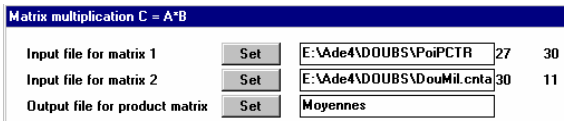
C'est la base des analyses de co-inertie que nous verrons plus loin. Il suffit que le produit de matrice ainsi défini ait un sens expérimental. Passer le tableau faunistique en pourcentage par espèce et transposer :




```
Bin->Bin: Frequencies
```




```
FilesUtil: Transpose
```

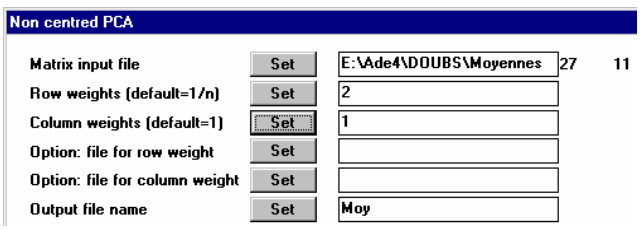



```
MatAlg: Matrix multiplication C =
A*B
```

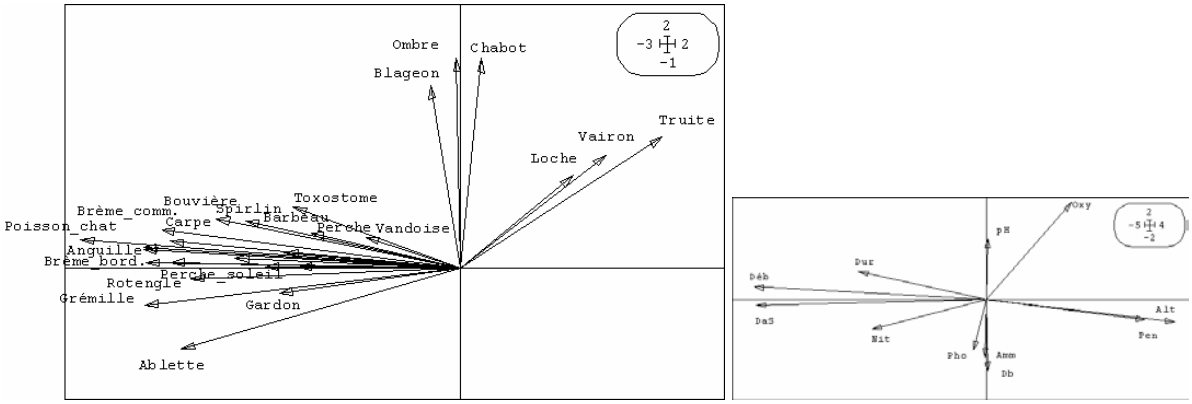
On obtient la position moyenne de chaque espèce sur chaque variable de milieu normalisée :

	DaS	Alt	Pen	Déb	pH	Dur	Pho	Nit	Amm	Oxy	Dbo
Chabot	-0.20	-0.23	-0.17	-0.06	0.80	0.33	-0.37	-0.28	-0.47	1.01	-0.57
Truite	-0.70	0.61	0.51	-0.53	0.25	-0.49	-0.47	-0.67	-0.47	0.79	-0.56
Vairon	-0.54	0.41	0.27	-0.43	0.14	-0.29	-0.41	-0.48	-0.41	0.64	-0.49
Loche	-0.41	0.32	0.15	-0.37	0.12	-0.28	-0.36	-0.36	-0.36	0.47	-0.44
Ombre	-0.11	-0.25	-0.18	0.14	0.64	0.51	-0.37	-0.36	-0.47	1.05	-0.52
Blageon	0.01	-0.37	-0.33	0.07	0.76	0.34	-0.31	-0.06	-0.41	0.77	-0.55
Hotu	0.87	-0.86	-0.66	0.78	-0.16	0.40	-0.05	0.48	-0.04	-0.22	-0.07
Toxostome	0.59	-0.72	-0.57	0.56	-0.14	0.35	-0.21	0.42	-0.15	0.12	-0.35
Vandoise	0.33	-0.34	-0.30	0.33	0.06	0.26	-0.09	0.18	-0.10	-0.05	-0.16
Chevaine	0.38	-0.34	-0.36	0.35	0.02	0.31	-0.01	0.20	-0.03	-0.20	-0.02
Barbeau	0.83	-0.81	-0.61	0.80	0.05	0.43	-0.12	0.44	-0.13	-0.07	-0.22
Spirlin	0.88	-0.83	-0.80	0.99	-0.12	0.55	-0.11	0.43	-0.11	-0.05	-0.26
Goujon	0.74	-0.66	-0.56	0.71	0.01	0.38	0.08	0.43	0.03	-0.24	-0.03
Brochet	0.63	-0.45	-0.43	0.68	-0.03	0.34	-0.01	0.27	-0.04	-0.29	-0.05
Perche	0.57	-0.44	-0.48	0.64	-0.08	0.36	-0.16	0.13	-0.17	-0.14	-0.20
Bouvière	1.13	-0.94	-0.78	1.18	-0.03	0.60	-0.03	0.53	-0.08	-0.28	-0.15
Perche_soleil	1.13	-0.94	-0.73	1.13	-0.07	0.62	0.00	0.56	-0.04	-0.38	-0.05
Rotengle	1.01	-0.76	-0.76	1.10	0.10	0.61	0.16	0.47	0.10	-0.46	0.04
Carpe	1.17	-0.94	-0.80	1.21	0.15	0.65	-0.05	0.51	-0.12	-0.31	-0.16
Tanche	0.70	-0.55	-0.52	0.66	0.06	0.30	-0.08	0.33	-0.11	-0.28	-0.11
Brème_comm.	1.28	-0.99	-0.76	1.32	0.02	0.65	-0.01	0.46	-0.13	-0.43	-0.07
Poisson_chat	1.51	-1.06	-0.91	1.71	0.13	0.91	0.03	0.45	-0.14	-0.51	-0.07
Grémille	1.23	-0.98	-0.78	1.19	-0.03	0.64	0.21	0.66	0.11	-0.61	0.16
Gardon	0.66	-0.56	-0.57	0.60	-0.06	0.38	0.07	0.41	0.06	-0.42	0.07
Brème_bord.	1.26	-0.99	-0.78	1.24	0.10	0.63	0.02	0.58	-0.04	-0.48	-0.01
Ablette	1.03	-0.91	-0.76	0.92	-0.12	0.54	0.39	0.76	0.36	-0.61	0.36
Anguille	1.23	-0.97	-0.84	1.30	0.05	0.68	0.01	0.57	-0.07	-0.38	-0.10

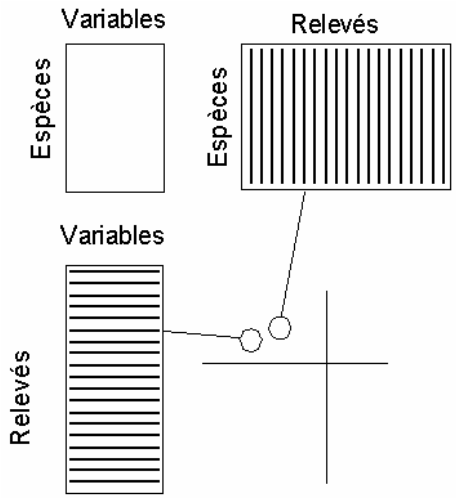
Faire l'ACP non centrée de ce tableau :



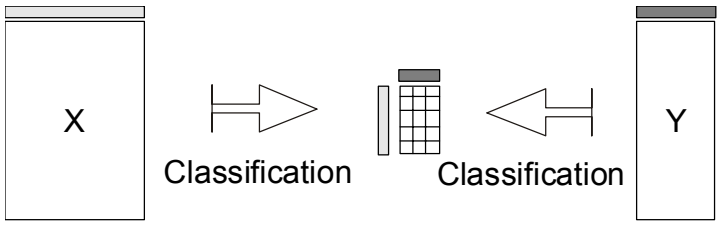
Fondamentalement, dans l'analyse du tableau croisé, les lignes sont les variables d'un tableau et les colonnes sont les variables de l'autre.



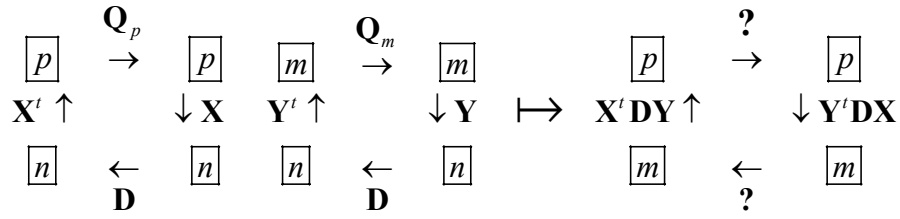
On a alors trois tableaux et on comprend que les relevés sont supplémentaires de deux manières dans l'analyse du tableau croisé :



Le cas le plus célèbre d'un croisement de tableau signifiant est celui des profils écologiques[5]. 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 $\mathbf{Y}^t\mathbf{X}$ 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 [6] a eu l'idée d'envoyer un tel tableau dans l'analyse des correspondances, idée retrouvée dans[7]. On obtient un cas particulier de l'analyse de co-inertie[8]. 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 :



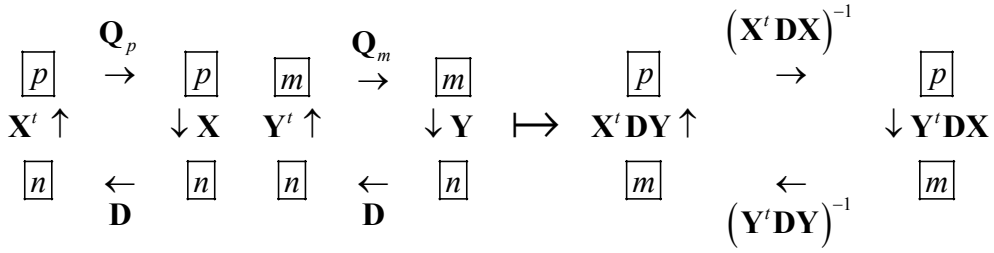
Le choix des métriques associées au croisement va définir les principales familles de méthodes.

2. Analyses canoniques

2.1. Analyse canonique des corrélations

L'analyse canonique des corrélations [9] est la plus ancienne et la plus connue des méthodes de couplage (revue pour l'écologie dans [10]). Le fondement considère deux ACP normées. On appellera simplement **X** et **Y** les deux tableaux normalisés (moyennes nulles et variances unitaires par colonnes).

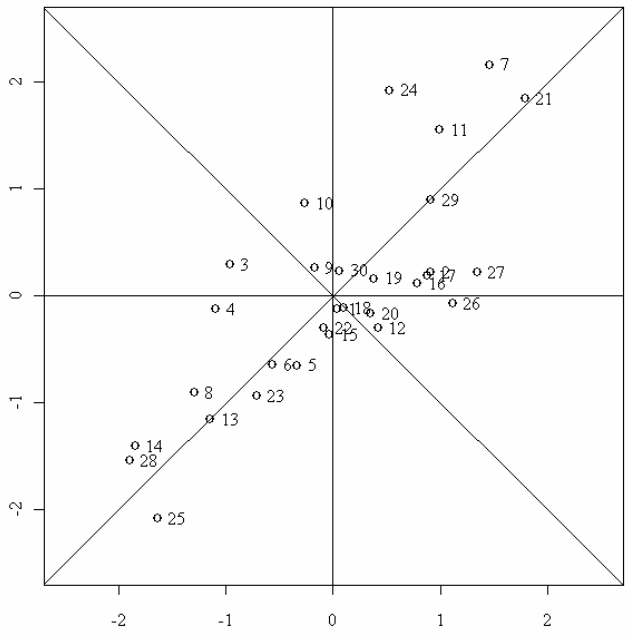
On suppose que les deux paquets de variables (colonnes de **X** et **Y**) sont sans redondances (la régression d'une variables de **Y** sur **X** ou d'une variable de **X** sur **Y** est définie sans problème) donc que les matrices de corrélations des deux paquets sont inversibles. L'analyse canonique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

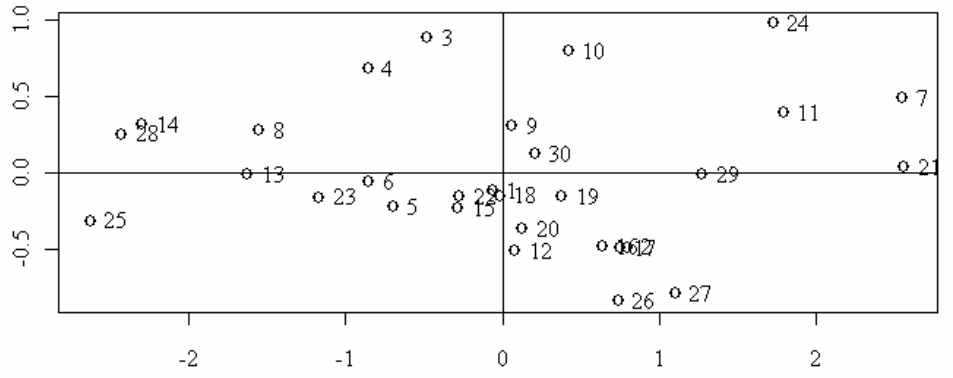
```
> cor <- matrix(c(1,0.8,0.8,1),nrow=2)
> cor
     [,1] [,2]
[1,]  1.0  0.8
[2,]  0.8  1.0
> X <- mvrnorm(30,c(0,0),cor)
> X <- scale(X)
```

```
> 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

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

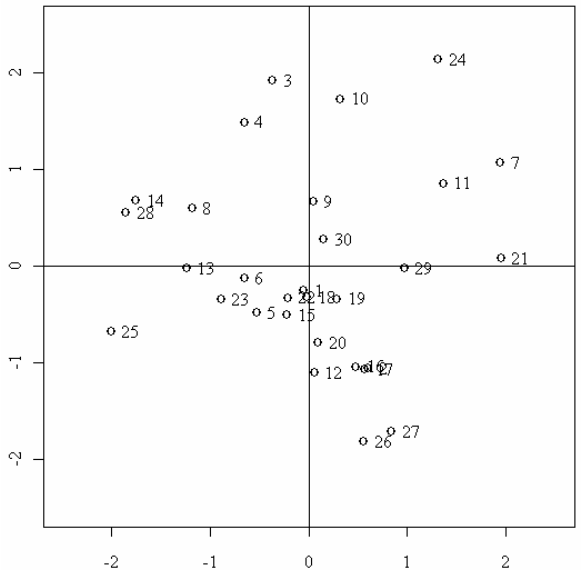


En se mettant dans la base des axes principaux sans changer de métrique, on conserve les distances :

```
> pr0 <- princomp(X)
> 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 :

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

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

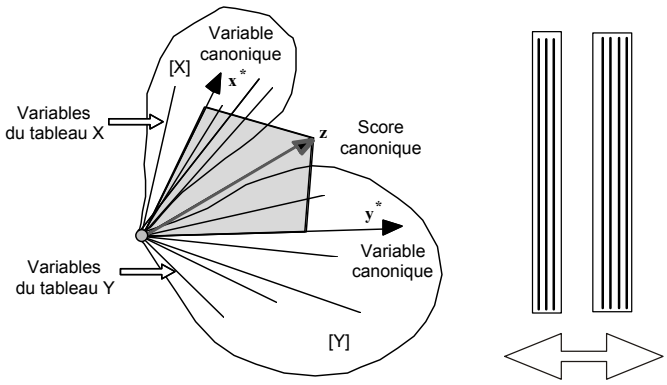
Les cofacteurs de ce schéma sont $\left(\left(\mathbf{Y}^t\mathbf{DY}\right)^{-1}\right)^{-1} = \mathbf{Y}^t\mathbf{DY}$ normés donc vérifient $\mathbf{b}^{*t}\,\mathbf{Y}^t\mathbf{DYb}^* = \left\|\mathbf{Yb}^*\right\|_{\mathbf{D}}^2 = 1$. Ce sont des coefficients de combinaisons linéaires de variables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

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

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

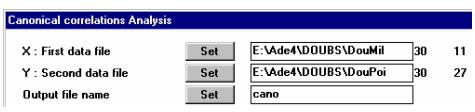
On obtient donc une combinaison de variables du tableau **X** de variance 1 (le vecteur $\mathbf{Xa}^*$) et une combinaison de variables du tableau **Y** de variance 1 (le vecteur $\mathbf{Yb}^*$) de corrélation maximale (la quantité $\left\langle \mathbf{Xa}^* \middle| \mathbf{Yb}^* \right\rangle_{\mathbf{D}}$). $\mathbf{Xa}^*$ 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.

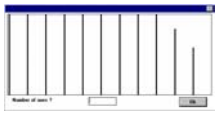
L'analyse canonique prend sa signification dans l'espace des variables. On peut la résumer par la figure :



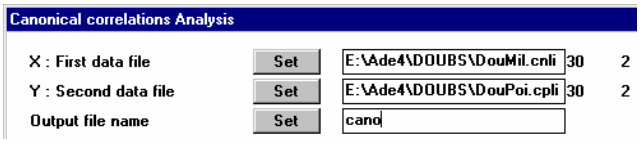
La variable canonique du tableau **X** est la combinaison des variables de **X** la mieux prédite par une régression multiple sur les variables de **Y** tout comme la variable canonique du tableau **Y** est la combinaison des variables de **Y** la mieux prédite par une régression multiple sur les variables 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.

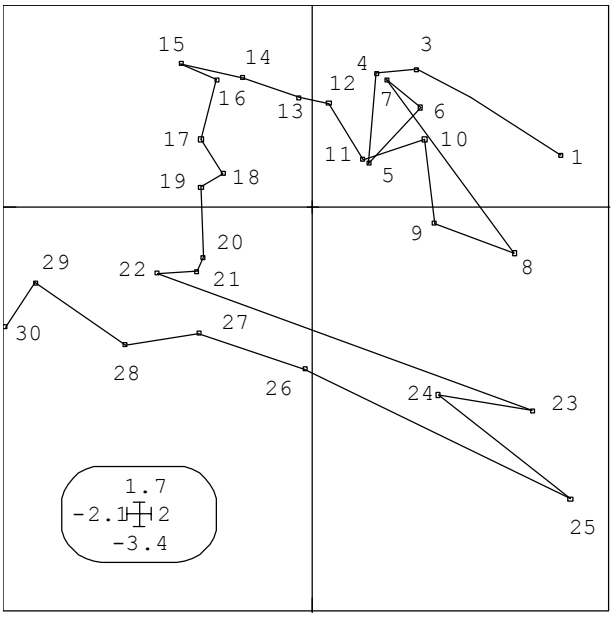




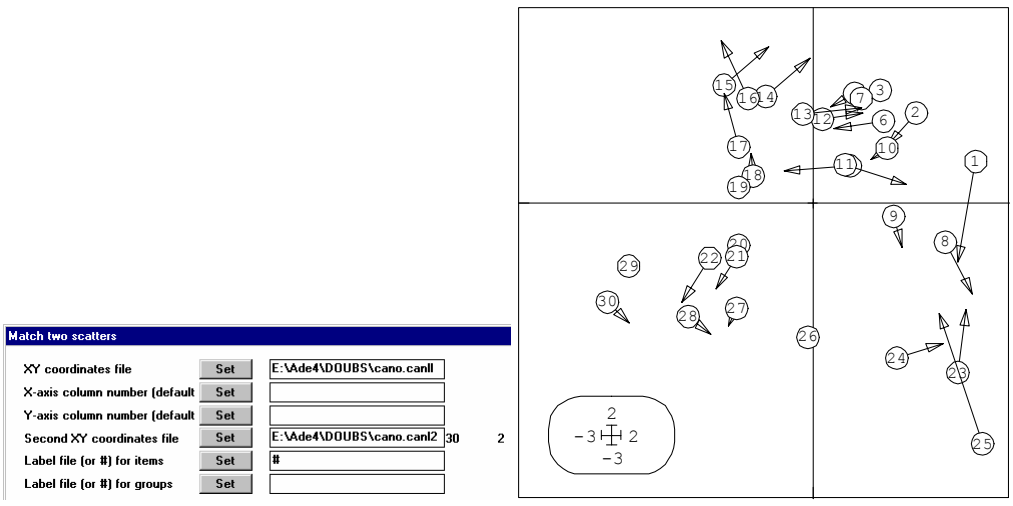
Ce n'est pas la peine de continuer. Il existe une multitude de combinaisons linéaires des variables faunistiques parfaitement corrélées à des combinaisons linéaires de variables de milieu, mais cela ne peut rien nous apprendre. Par contre, on l'utilise avec un tableau et les coordonnées de l'analyse de l'autre ou avec deux ensembles de coordonnées [11] :



On garde ici deux variables canoniques pour deux variables (coordonnées) dans chaque tableau. Les scores canoniques sont dans .canll. Ce sont des variables de synthèse, de moyenne 0 et de variance 1 qui expriment la corrélation entre les deux tableaux :



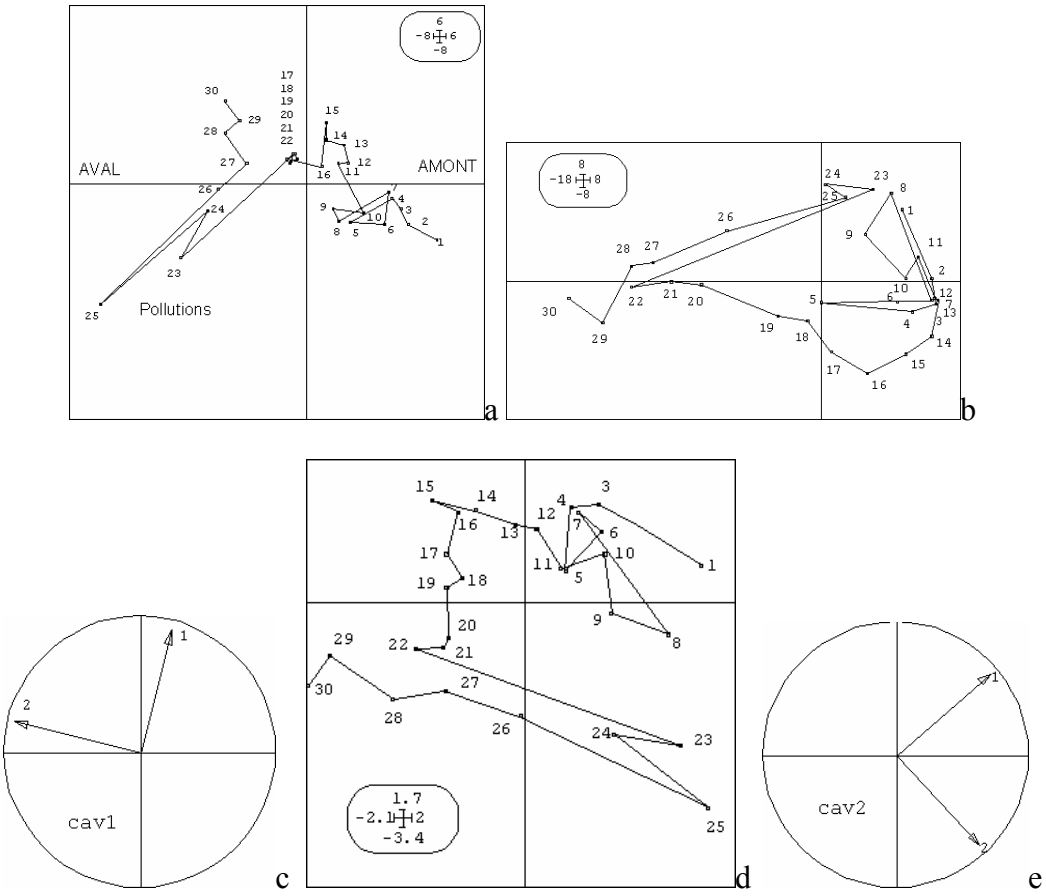
On peut évidemment utiliser ces codes comme fond de carte pour toute expression directe des données. Pour obtenir les scores canoniques, on a fait une moyenne normalisée de variables canoniques, combinaisons de variables normalisées de chacun des tableaux. On peut représenter ces deux ensembles et leur corrélation par :




```
Likelihood ratio tests of dimensionality
Barlett 1938, see Ch. 3.4.2 of Gittins, R. (1985) Canonical analysis,
```

```
a review with applications in ecology. Springer-Verlag, Berlin. 1-351
-----------------------------------------------------
k= 0 Khi2 = 4.1260e+01  ddl =  4   proba = 8.5582e-08
k= 1 Khi2 = 1.5586e+01  ddl =  1   proba = 1.3171e-04
-----------------------------------------------------

Canonical correlation coefficients
k= 1 rk = 7.9279e-01  rk2 = 6.2852e-01
k= 2 rk = 6.5775e-01  rk2 = 4.3263e-01
```

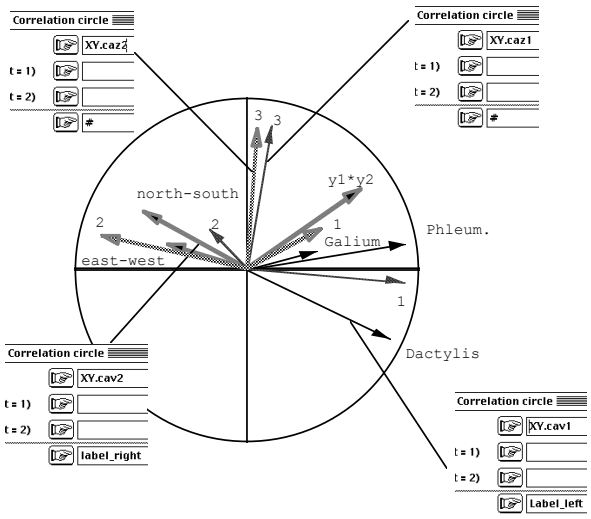
La géométrie des nuages de variables s'exprime sur des cercles de corrélation :



a - carte factorielle de l'analyse 1. b - carte factorielle de l'analyse 2. d - Représentation des deux premiers scores canoniques. c - projection des variables (composantes principales servant de variables) du tableau 1 sur le plan des scores canoniques. e - projection des variables (composantes principales servant de variables) du tableau 2 sur le plan des scores canoniques. Cette figure mélange une projection d'un nuage de points dans $\mathbb{R}^p$ (a), une projection d'un nuage de points dans $\mathbb{R}^q$ (b), deux projections dans $\mathbb{R}^n$ (c et e) et une figuration simple de deux scores normés non corrélés (d) non associés à une projection de nuages.

Il faut faire une rotation de 90° pour le milieu et une rotation de 45° suivie d'une symétrie autour de l'axe des x pour la faune pour recaler les deux nuages de points. C'est en bas à droite que l'ajustement sera le moins bon, ce qui est parfaitement exprimé dans le graphe double. En fait cette analyse marche bien parce qu'on a deux systèmes de covariances. Le premier concerne le gradient amont-aval avec covariances de signe opposé des espèces de tête de bassin et celles de plaine. Le second concerne la pollution qui élimine toutes les espèces en haut comme en bas. L'analyse canonique sur coordonnées factorielles est voisine de l'analyse de co-inertie.

Pour approfondir l'analyse canonique, la référence [10] s'impose. Retrouver les illustrations de cet ouvrage sur la carte Angles :



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 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.

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
> 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 :

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

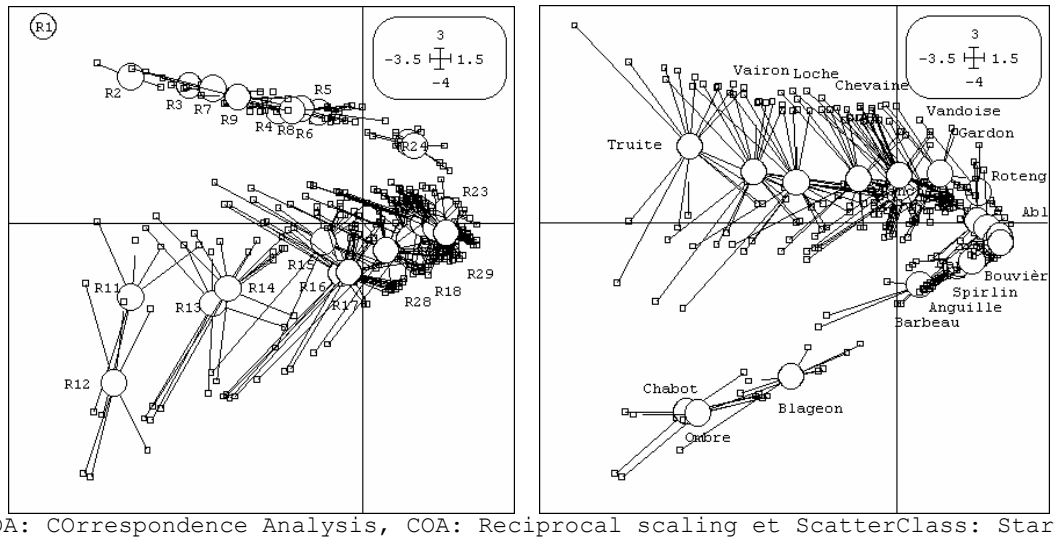
Le vecteur $\mathbf{1}_n$ est dans chacun des sous-espaces engendrés, d'où la corrélation canonique de 1 dans celle de R et la valeur propre de 0 dans celle de ADE-4. On trouve dans [12] l'affirmation de l'importance fondamentale de l'AFC comme analyse canonique.

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 pourrait se trouver dans certaines stations mais n'y sont pas pour des contingences historiques. Le raisonnement est dans [13]. 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 *3* ou plutôt 3/1004 (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 [14]. Voir OccurData: Array_to_Occur.



COA: COrrespondence Analysis, COA: Reciprocal scaling et ScatterClass: Stars.

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) [15].

2.3. 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).

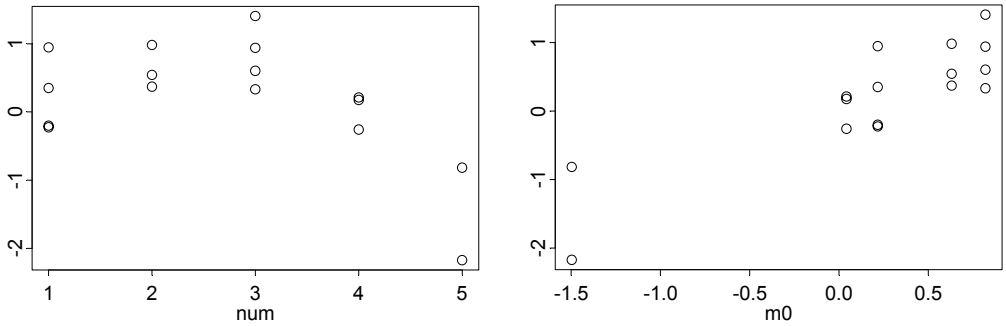
```
> num
 [1] 1 3 2 2 2 3 3 3 1 1 4 4 4 5 5 1
> tabnum
   X1 X2 X3 X4 X5
1   1  0  0  0  0
2   0  0  1  0  0
3   0  1  0  0  0
4   0  1  0  0  0
5   0  1  0  0  0
```

```
 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
```

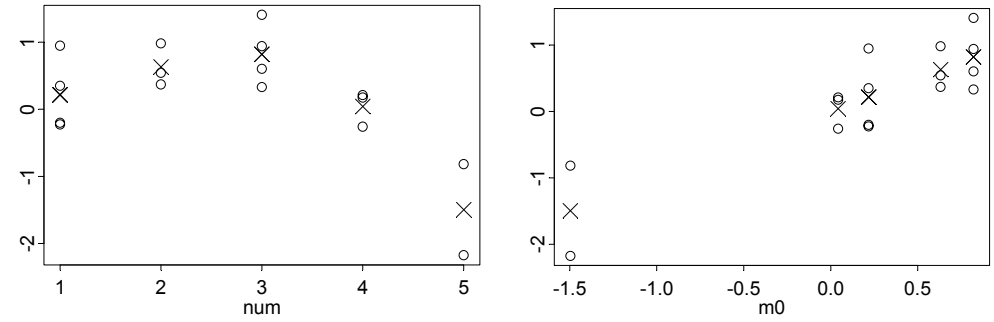
Le minimum nécessaire à la manipulation du lien entre une variable qualitative et une variable quantitative tient dans le rapport de corrélation.

```
> 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 :

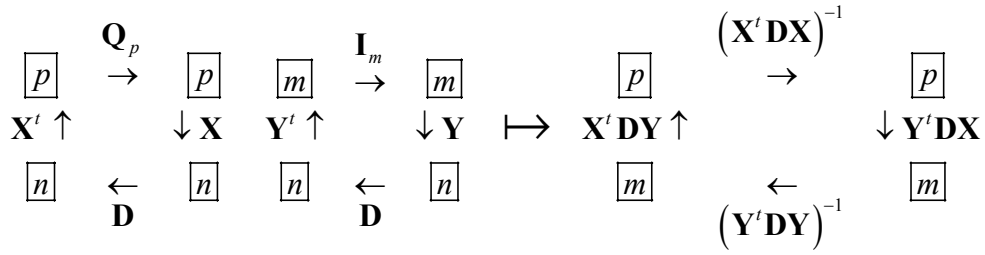



```
> tapply(y,num,mean)
      1      2    3       4      5
 0.2175 0.6311 0.82 0.04226 -1.496   Les moyennes par classe
> tapply(y,num,var,un=F)
      1      2      3      4     5
 0.2301 0.0663 0.1614 0.0455 0.461   Les variances par classe
> var(y,un=F)
[1] 0.6717                          La variance 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
```

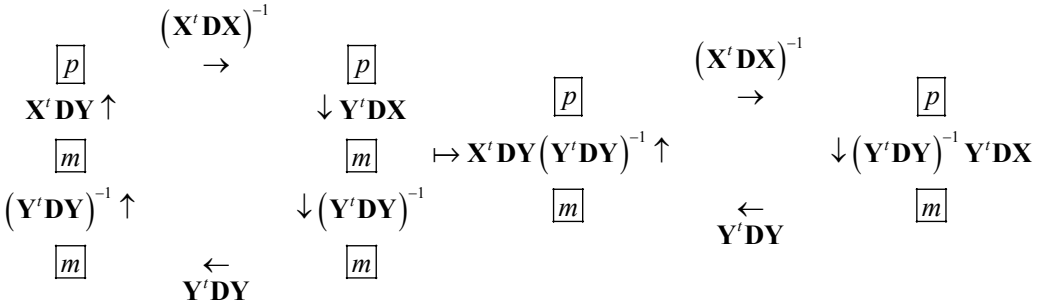
```
> sum(tapply(y,num,var,un=F)*table(num))/16
[1] 0.1765                La moyenne des variances par classe (variance Intra)
> sum(tapply(y,num,var,un=F)*table(num))/16+var(m0,un=F)
[1] 0.6717
> var(y,un=F)
[1] 0.6717                La variance totale = Inter + Intra

> 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 [16].

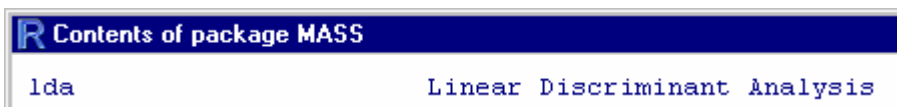
Les schémas de l'analyse discriminante sont donc :



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 :

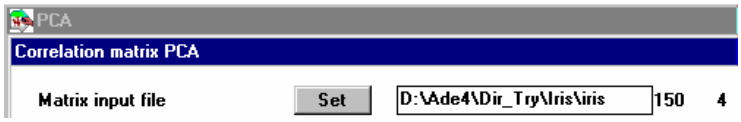


Dans **D,** on trouve les poids des points. $\mathbf{Y}^t\mathbf{DY}$ est une matrice diagonale (deux indicatrices n'ont aucune valeur non nulle à la même place) qui contient les poids des classes (somme des poids des points de la classe). $\mathbf{Y}^t\mathbf{DX}$ fait les sommes pondérées des valeurs 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.

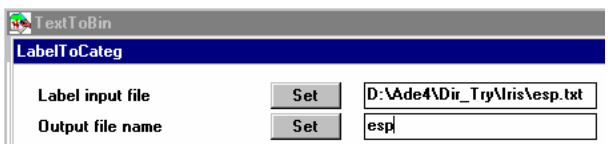


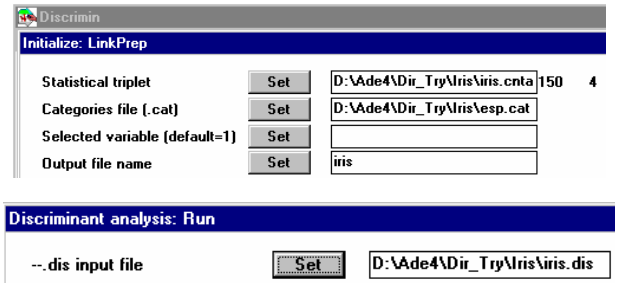
```
> library(MASS)
> ?lda
> data(iris)
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> dim(iris)
[1] 150   5
> iris[1:5,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         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
```

Les données sont reproduites dans la carte Iris.




```
---------------------- Correlation matrix -------------------
[  1] 1000
[  2] -118 1000
[  3]  872 -428 1000
[  4]  818 -366  963 1000
-------------------------------------------------------------
Total inertia:        4
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum   |
01   +2.9185E+00 +0.7296 +0.7296   |02   +9.1403E-01 +0.2285 +0.9581 |
03   +1.4676E-01 +0.0367 +0.9948   |04   +2.0715E-02 +0.0052 +1.0000 |
```






```
File D:\Ade4\Dir_Try\Iris\iris.dima contains the parameters:
   input file: D:\Ade4\Dir_Try\Iris\iris.cnta
   categorical variable file: D:\Ade4\Dir_Try\Iris\esp
   n. of categorical variable used: 1

Discriminant analysis
Groups are defined by column 1 of file D:\Ade4\Dir_Try\Iris\esp
Input statistical triplet: table D:\Ade4\Dir_Try\Iris\iris.cnta
Number of rows: 150, columns: 4
total inertia (norm C- generalised inverse) =  matrix rank:  4.000e+00

between-class inertia (norm C-):  1.192e+00 (ratio:  2.980e-01)

Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +9.6987E-01 +0.8137 +0.8137   |02   +2.2203E-01 +0.1863 +1.0000 |
03   +0.0000E+00 +0.0000 +1.0000

File  D:\Ade4\Dir_Try\Iris\iris.divp  contains  the  eigenvalues  and  relative
inertia for each axis
It has 3 rows and 2 columns

File   D:\Ade4\Dir_Try\Iris\iris.difa   contains   the   coefficient   of   the
discriminant scores
It has 4 rows and 2 columns
File :D:\Ade4\Dir_Try\Iris\iris.difa
|Col.|   Mini    |   Maxi    |
|----|----------|----------|
|   1|-1.200e-01| 6.790e-01|
|   2|-1.461e+00| 1.922e+00|
|----|----------|----------|

File D:\Ade4\Dir_Try\Iris\iris.dili contains the canonical scores of row (unit
norm)
It has 150 rows and 2 columns

File :D:\Ade4\Dir_Try\Iris\iris.dili
|Col.|   Mini    |   Maxi    |
|----|----------|----------|
|   1|-1.727e+00| 1.608e+00|
|   2|-2.355e+00| 2.439e+00|
|----|----------|----------|
File D:\Ade4\Dir_Try\Iris\iris.diax contains the principal axes
It has 4 rows and 2 columns
File :D:\Ade4\Dir_Try\Iris\iris.diax
|Col.|   Mini    |   Maxi    |
|----|----------|----------|
|   1|-5.308e-01| 9.850e-01|
|   2| 4.604e-02| 7.580e-01|
|----|----------|----------|
File  D:\Ade4\Dir_Try\Iris\iris.dicp  contains  the  correlations  between  PCA
scores
and DA scores. It has 4 rows and 2 columns
File :D:\Ade4\Dir_Try\Iris\iris.dicp
|Col.|   Mini    |   Maxi    |
|----|----------|----------|
|   1|-1.501e-01| 9.815e-01|
|   2|-8.347e-01| 2.991e-01|
|----|----------|----------|
```

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

```
> 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
```

Jusque là tout va bien.

```
> ld <- lda(iris[,1:4],iris$Species)
> names(ld)
[1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"
[8] "call"

> ld$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.

```
------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\IRIS\iris.difa - 4 rows, 2 cols.
   1 | -0.1200  0.0177
   2 | -0.1169  0.8378
   3 |  0.6790 -1.4609
   4 |  0.3744  1.9218
```

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 $\mathbf{X}$ 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 :

```
> ld$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 :

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

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 maximise 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 :

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

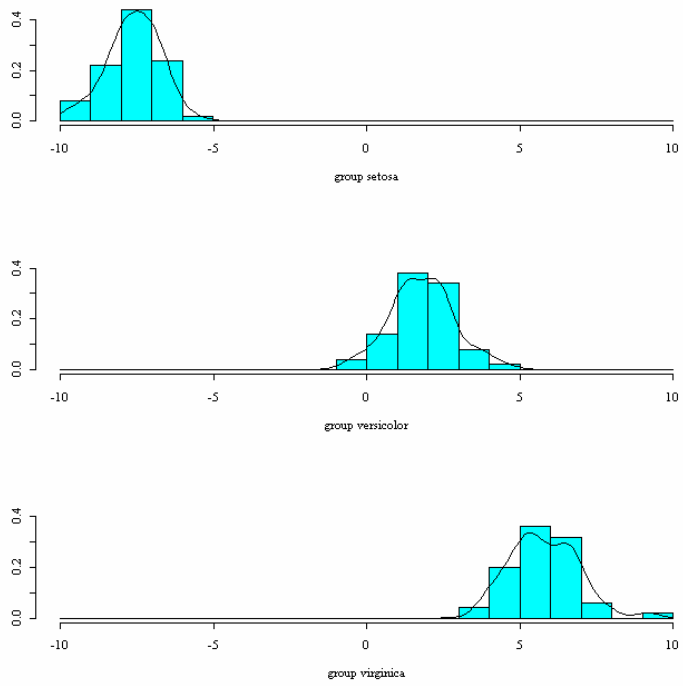
On y trouve les combinaisons linéaires de variance intra-classe unité et de covariance intra-classe nulle deux à deux qui maximise la variance inter-classe. C'est un problème très voisin car $\mathbf{G}^t\mathbf{D}_m\mathbf{G} = \mathbf{X}^t\mathbf{P}^t\mathbf{D}\mathbf{P}\mathbf{X} = \mathbf{B}$. Si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

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

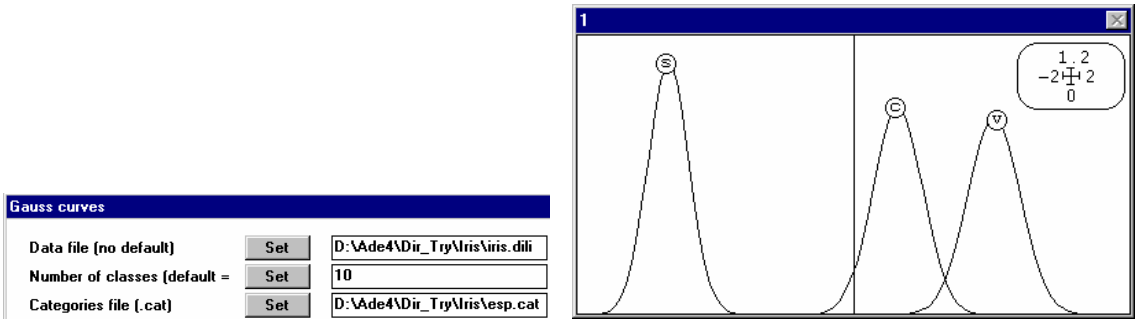
$$\Rightarrow \left(1-\lambda_k\right)\mathbf{B}\mathbf{u}_k = \lambda_k\mathbf{W}\mathbf{u}_k \Rightarrow \mathbf{W}^{-1}\mathbf{B}\mathbf{u}_k = \frac{\lambda_k}{1-\lambda_k}\mathbf{u}_k \Rightarrow \mu_k = \frac{\lambda_k}{1-\lambda_k} \Rightarrow \lambda_k = \frac{\mu_k}{1+\mu_k}$$

La trace du premier $Trace\left(\mathbf{T}^{-1}\mathbf{B}\right)$ s'appelle le critère de Pillai [17] et celle du second $Trace\left(\mathbf{W}^{-1}\mathbf{B}\right)$ s'appelle le critère généralisé de Hotelling dû à Lawley [18]. Ces quantités donnent des tests d'hypothèses dans le modèle gaussien (toute l'information est dans [19]).

```
> ld <- lda(iris[,1:4],iris$Species)
> plot(ld,dimen=1, type="both")
```



La fonction de R donne la combinaison de variables de variance intra-class unité (la variance par classe en moyenne vaut 1) qui maximise la variance inter-classe. La fonction d'ADE-4 donne la combinaison de variables de variance unité qui maximise la variance inter-classe.



On a donc la solution d'un même problème sous deux contraintes différentes. Il n'y a pas de contradictions mais pas d'identité des solutions. Le choix d'ADE-4 permet d'introduire dans une analyse discriminante un triplet de départ arbitraire, ce qui donne par exemple, une analyse discriminante des correspondances[20].

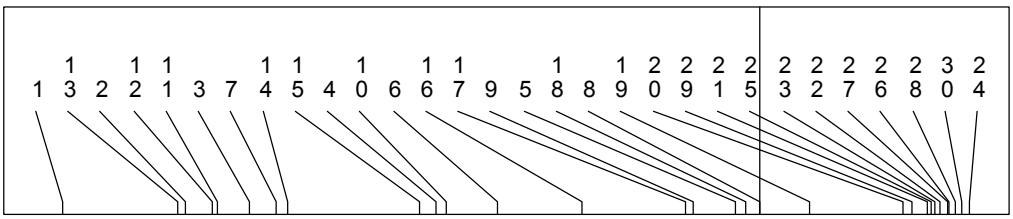
On retrouve le lien entre l'analyse discriminante et l'analyse canonique par :

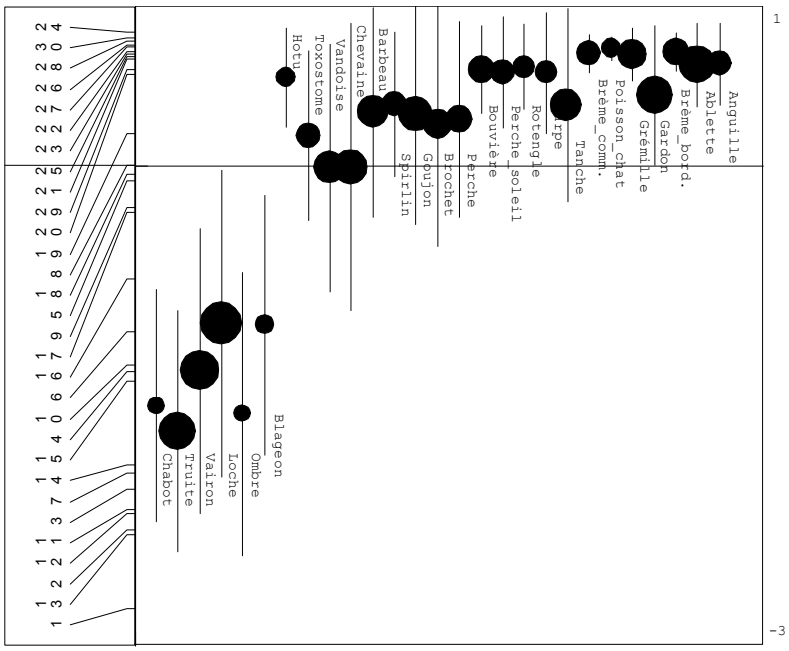
```
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +9.6987E-01 +0.8137 +0.8137   |02   +2.2203E-01 +0.1863 +1.0000 |
03   +0.0000E+00 +0.0000 +1.0000
```

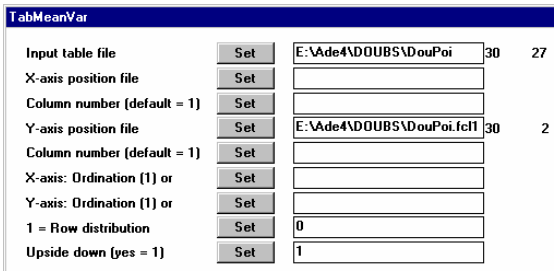
```
> x <- iris[,1:4]
> i1 <- as.numeric(iris$Species==levels(iris$Species)[1])
> i2 <- as.numeric(iris$Species==levels(iris$Species)[2])
> y <- cbind.data.frame(i1,i2)
> can0 <- cancor(x,y)
> can0
$cor
[1] 0.9848 0.4712

$xcoef
```

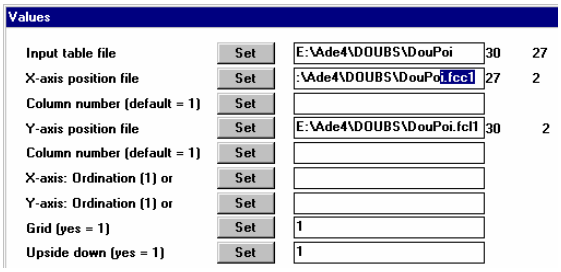
```
         [,1]      [,2]     [,3]     [,4]
[1,]  0.01187 -0.001753 -0.25105  0.07782
[2,]  0.02197 -0.157466  0.12503 -0.18250
[3,] -0.03151  0.067796  0.12746 -0.21329
[4,] -0.04023 -0.206546 -0.03786  0.37437

$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
1.15030
…
[145] -1.20059 -0.98977 -0.90816 -0.87102 -1.03205 -0.82112
>
------------------------------------------------
Binary input file: D:\ADE4\DIR_TRY\IRIS\iris.dili - 150 rows, 2 cols.
   1 | -1.4135  0.2677
   2 | -1.2499 -0.7009
   3 | -1.3132 -0.2365
   4 | -1.1946 -0.5975
   5 | -1.4259  0.4584
…
 145 |  1.2006  2.1642
 146 |  0.9898  1.4948
 147 |  0.9082 -0.3239
 148 |  0.8710  0.7316
 149 |  1.0321  2.0894
 150 |  0.8211  0.2958
```

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 [21], par exemple de la discrimination des espèces par les relevés.









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 $A \cap B^{\perp}$ [22] 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 [23].

2.4. 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 auxquel 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 maximise la variance des positions moyennes des espèces [24]. 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 lui 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.

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

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 :

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

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 $(2\times1.5+1\times1.8)/3$ 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 :

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

En passant aux occurrences, on aura :

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

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 :

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

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

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

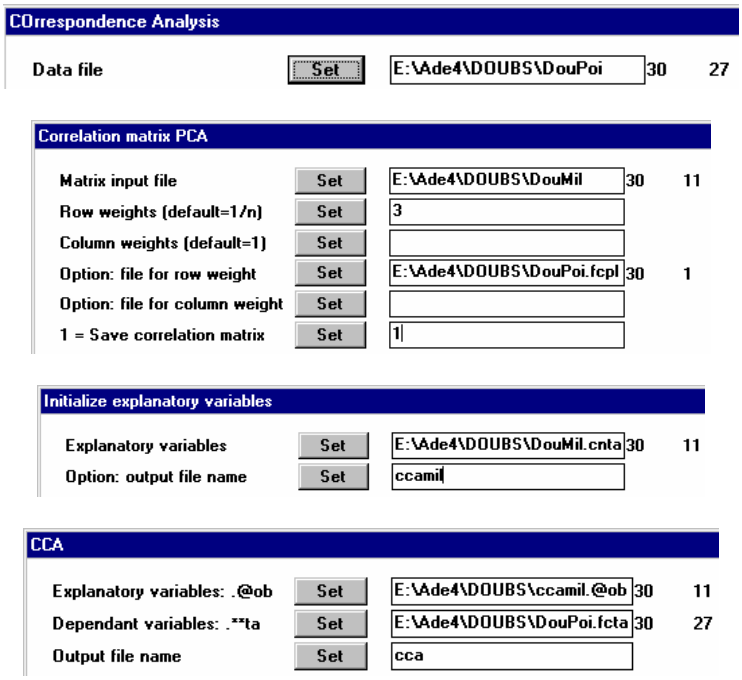
C'est-à-dire exactement la même chose qui s'écrit :

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

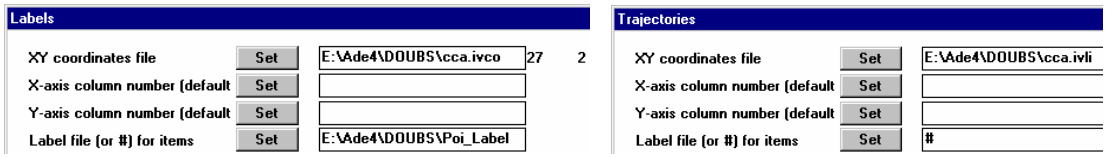
On obtient, comme cas particulier du modèle général, que les facteurs de cette analyse donnent des combinaisons de variables de milieu qui maximisent la variance des

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 [25]) et on retiendra des justificatifs le plus adéquats à défendre un point de vue. Plusieurs autres modèles se cachent dans ce schéma complexe et on trouvera des synthèses dans [26] et [27].

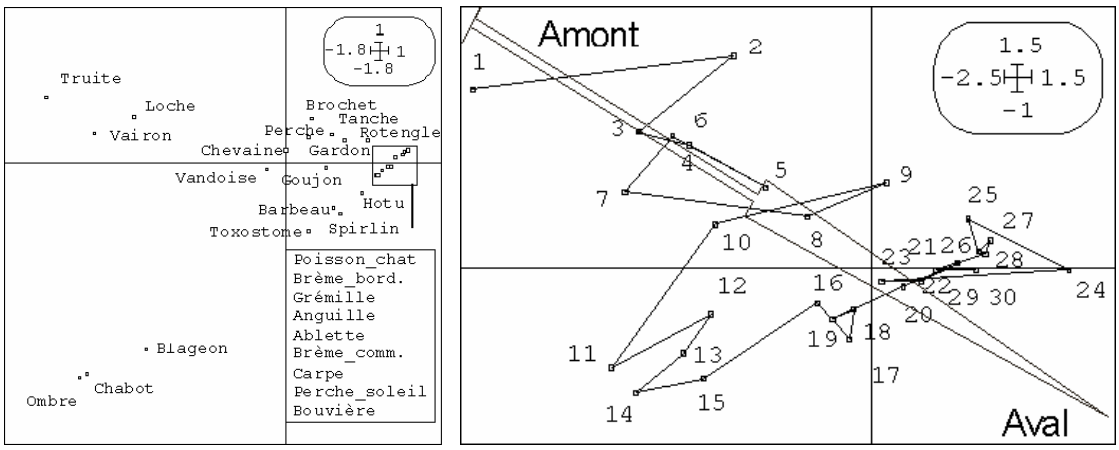
L'important à retenir : l'ACC se pratique après une AFC du tableau faune et une ACP normée du tableau de milieu *utilisant la pondération des relevés issue de l'AFC* du tableau faune. De ce point de vue encore, le milieu enregistré dans un relevé est d'autant plus important qu'il y a plus de taxons présents (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.




```
|-------------------------------------------
| files cca.ivta
|       cca.ivpc
|       cca.ivpl
|       cca.ivpa
|       cca.ivvp
|       cca.ivco
|       cca.ivli
| are those of the complete analysis
| of the projected table (DDUtil can been used)
```

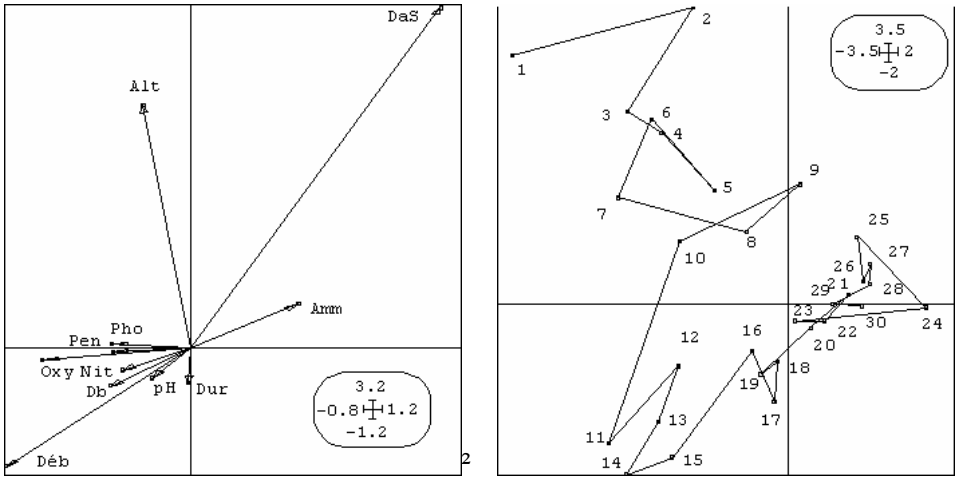


L'analyse est vue comme l'ACP du tableau d'AFC prédit par les variables de milieu [26].

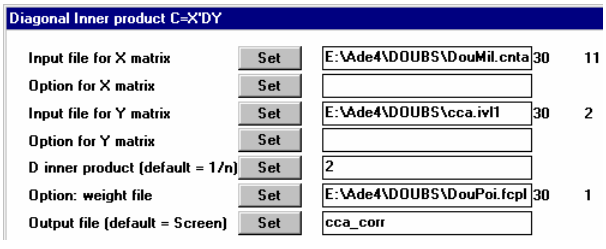


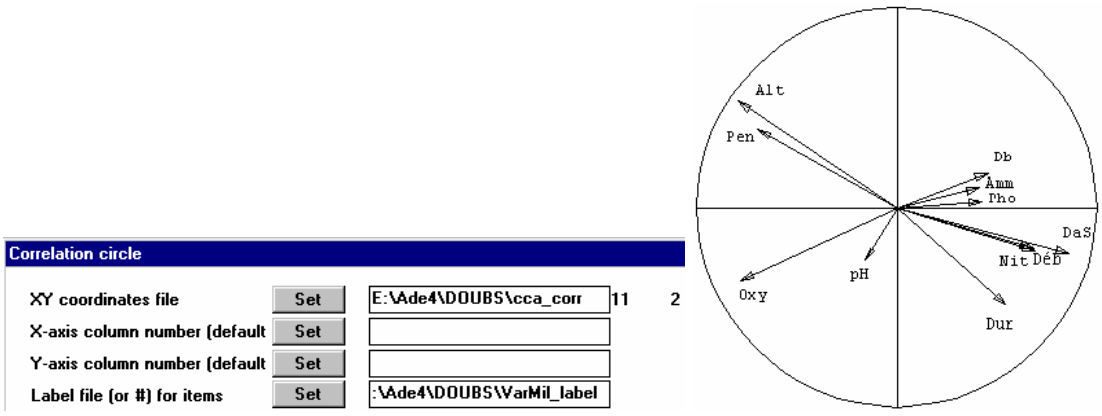

```
|-----------------------------------------------
| files cca.ivfa
|       cca.ivl1
|       cca.ivco
| allow a convenient interpretation
|-----------------------------------------------
```

ivfa contient les coefficients des variables de milieu (les loadings d'une ACP). Ces coefficients définissent des variables de variance unité (ivl1) :



qui maximisent la variance des positions des espèces (ci-dessus). Les coefficients sont bizarres. On calcule les corrélations entre scores et variables :



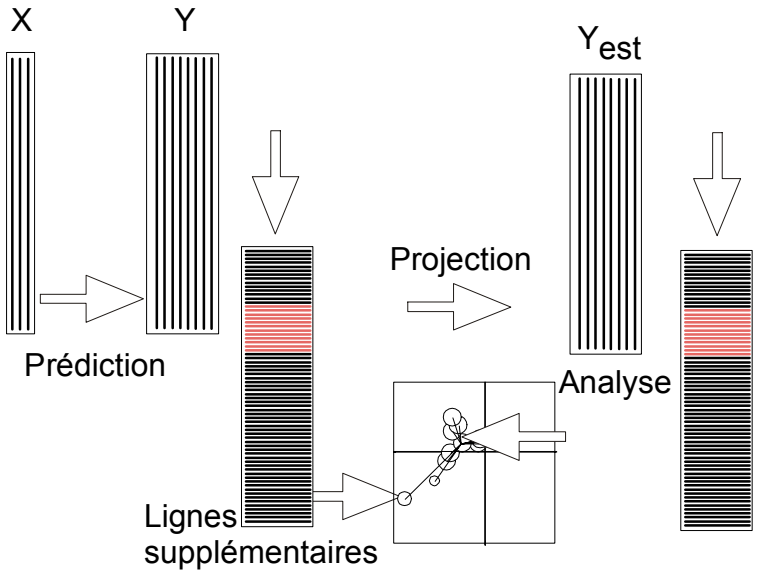


Les très fortes incohérences entre les coefficients qui fabriquent les scores à partir des variables et les corrélations entre ces mêmes scores et ces mêmes variables indiquent que les conditions d'emploi ne sont pas bonnes. C'est un problème très général en régression multiple et toutes les méthodes dérivées [28]. A utiliser avec la plus grande prudence. L'ACC fait souvent guère plus que l'AFC sans qu'on s'en rende compte.

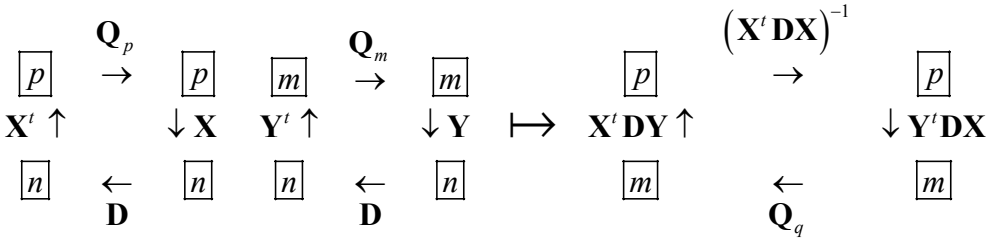
A retenir : on maximise une corrélation entre combinaisons de variables quantitatives en *analyse canonique des corrélations*. On maximise la corrélation entre scores des lignes et des colonnes dans une *analyse des correspondances*. On maximise un pourcentage de variance expliquée par une partition 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

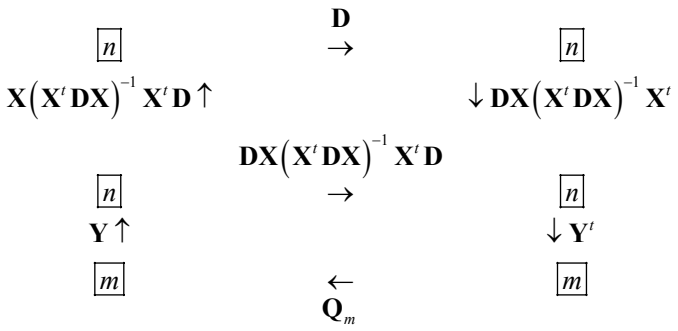
Ce sont des méthodes dissymétriques. Un tableau est formé d'explicatives (variables instrumentales) et un tableau est formé de variables à étudier. Dans ces méthodes on étudie le second en utilisant le premier. On parle en général d'Analyses en Composantes Principales sur Variables Instrumentales ou ACPVI. Fondations dans [29]. Le principe de base se trouve dans le schéma :



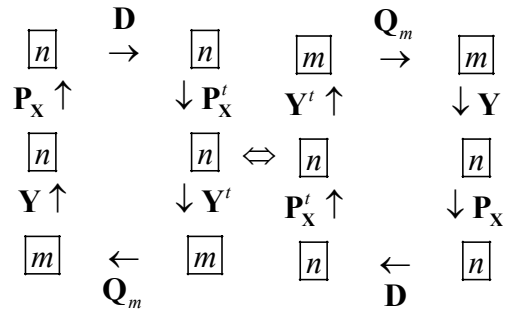
X est le tableau des variables instrumentales. **Y** est le tableau à analyser. Chacune des variables de **Y** est prédite par une régression multiple sur les variables de **X**. Les modèles sont rangés dans le tableau $\mathbf{Y}_{\mathbf{est}}$. Les variables de $\mathbf{Y}_{\mathbf{est}}$ sont obtenues 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 **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 **Y**. On fait une analyse de **Y** sous contrainte de **X**. Suivant **X** et **Y** on a un ensemble de méthodes dites d'ordination sous contraintes ou d'analyses sur variables instrumentales. Les schémas utiles sont :



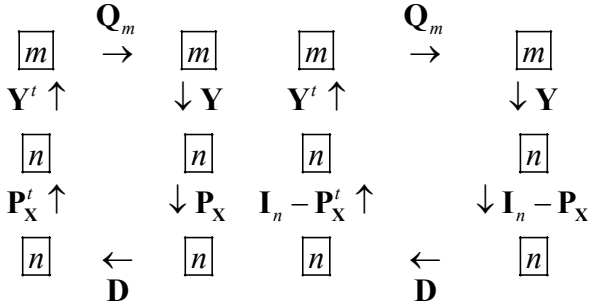
Le schéma de droite est celui d'une ACPVI. On le réécrit sous la forme :



ou encore :

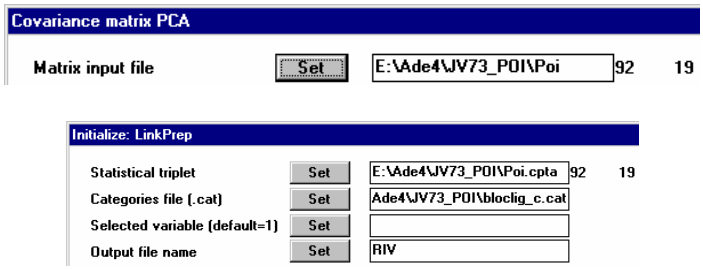


qui montre $\mathbf{Y}_{\mathbf{est}} = \mathbf{P}_{\mathbf{X}}\mathbf{Y}$. Le tableau **X** peut être remplacé par une base orthonormée d'un sous-espace de projection (module Projectors d'ADE-4) ce qui permet de distinguer les ACPVI directes et les ACPVI orthogonales qui décomposent une analyse simple avec deux projecteurs complémentaires :

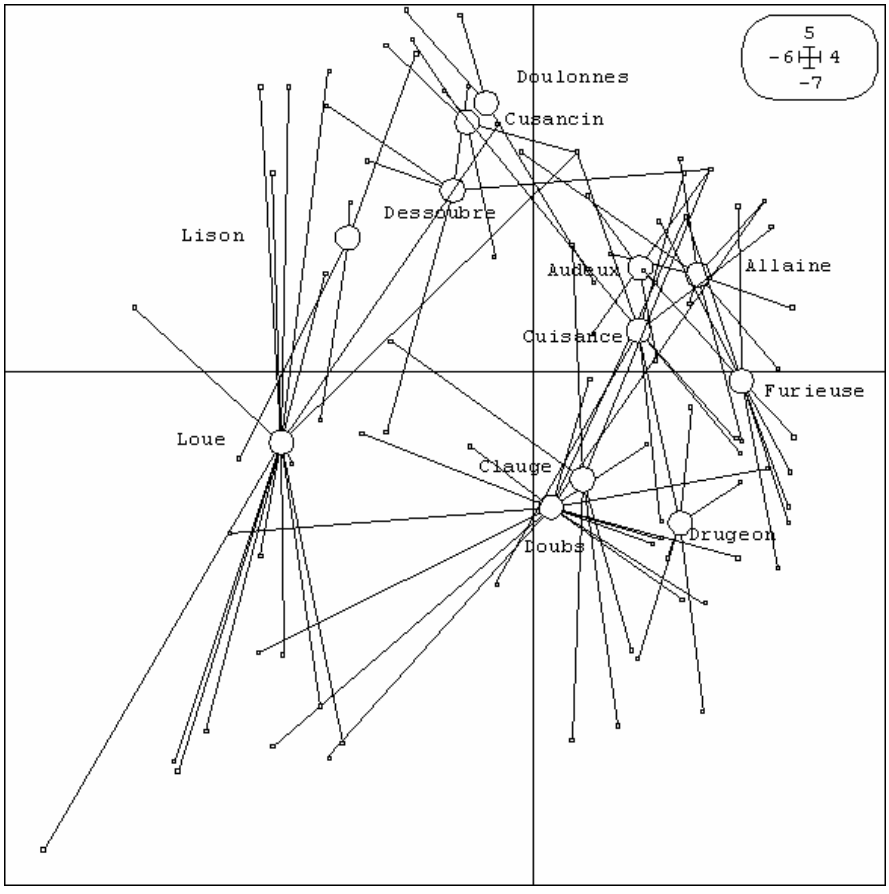


3.1. Analyses inter-classes

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. Utiliser la carte JV73_poi. Faire une typologie de rivière à partir d'une typologie de stations.

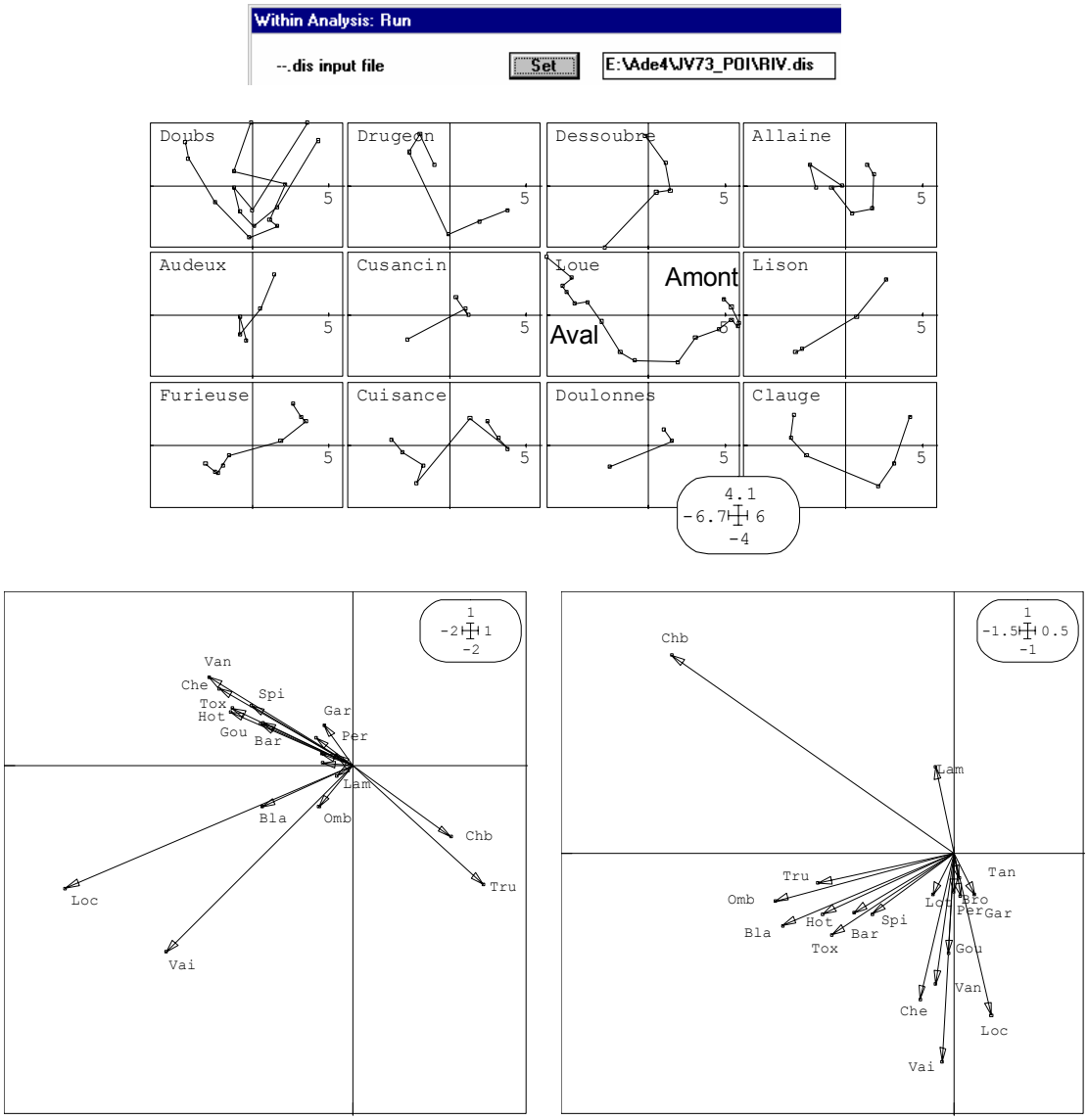



```
Between-class inertia  8.733e+00 (ratio:  3.135e-01)
31% de variabilité entre rivières
Within-class inertia  1.912e+01 (ratio:  6.865e-01)
69 % de variabilité interne aux rivières
```



3.2. Analyses intra-classes

Si on utilise, en lieu de l'analyse du tableau estimée 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.



A gauche, structure intraclasse, à droite structure interclasse.

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-classe 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 [30]. Retenir les schémas de principe :

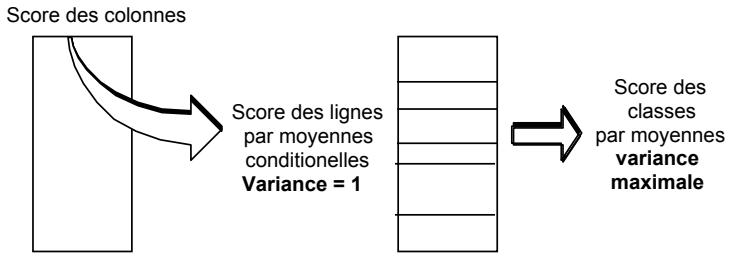


Schéma de principe de l'analyse discriminante des correspondances [31].

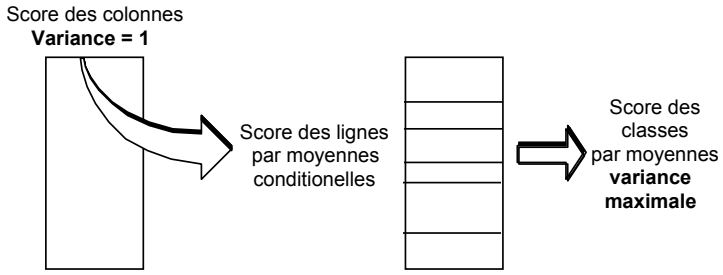
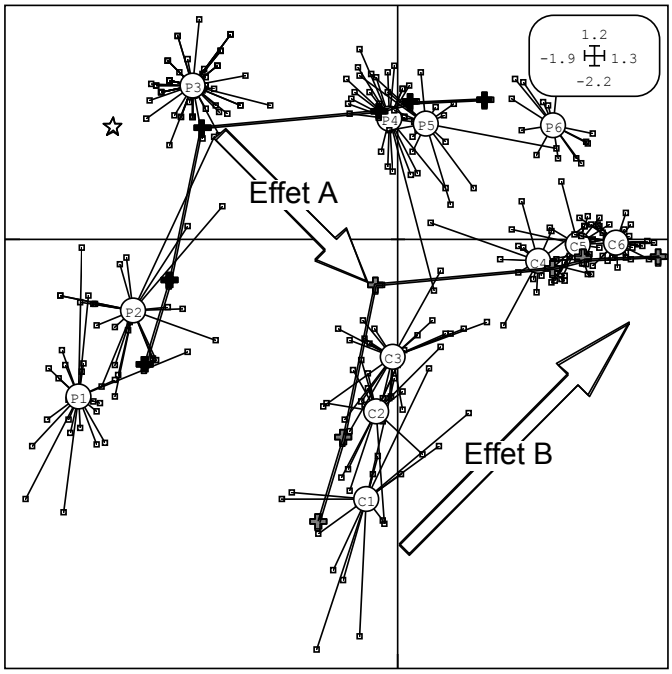


Schéma de principe de l'analyse des correspondances inter-classes.

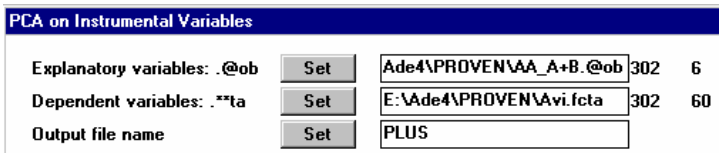
L'intra-classes en AFC peut s'étendre aux blocs de colonnes et aux doubles contraintes (COA: Internal COA) [32].

3.3. Ordinations sous contraintes

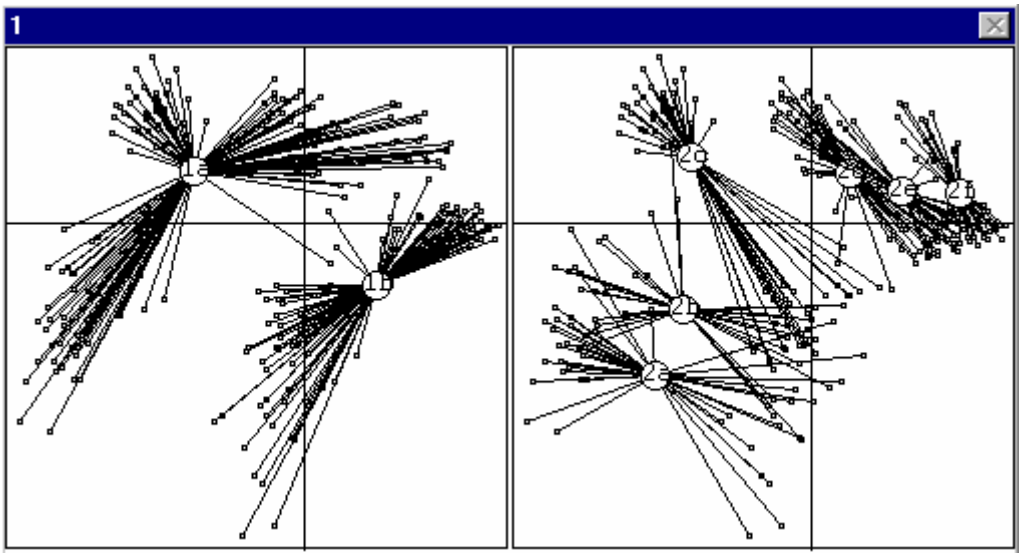
Les contraintes en inter et intra-classes sont simples. On peut les rendre assez complexes en introduisant des sous-espaces de projection (module Projectors). Utiliser la carte Proven [33] pour faire une analyse sous contrainte **A**+**B** :



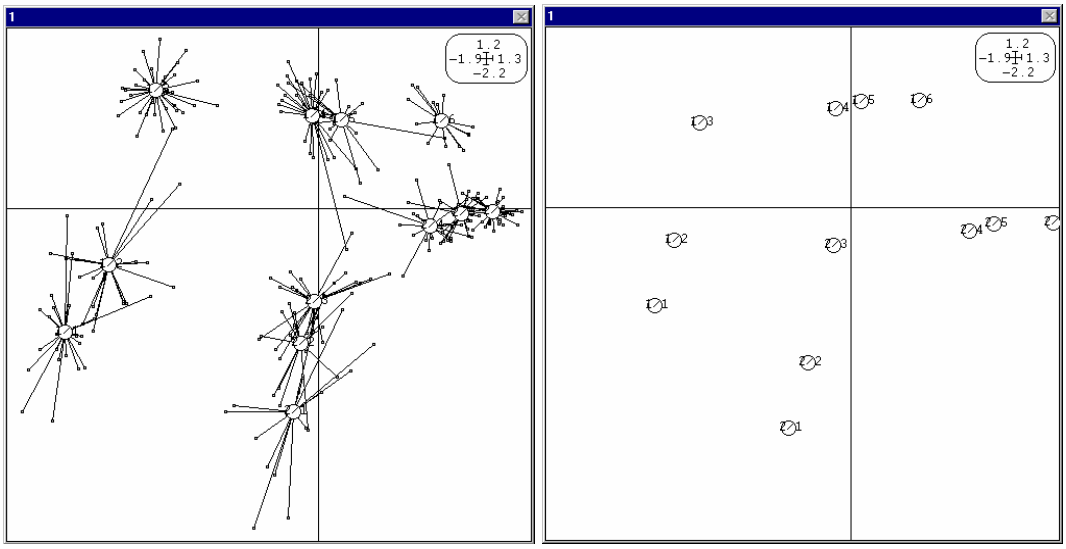
Lire le plan d'expérience (CategVar: Read Categ File), faire l'AFC du tableau faunistique (COA: COrrespondence Analysis), définir les espaces de projections (Projectors: Two Categ Var->Orthonormal Bases) en introduisant la pondération des lignes de l'analyse précédente et faire l'ACPVI sur le sous-espace A+B (Projectors: PCA on Instrumental Variables) :



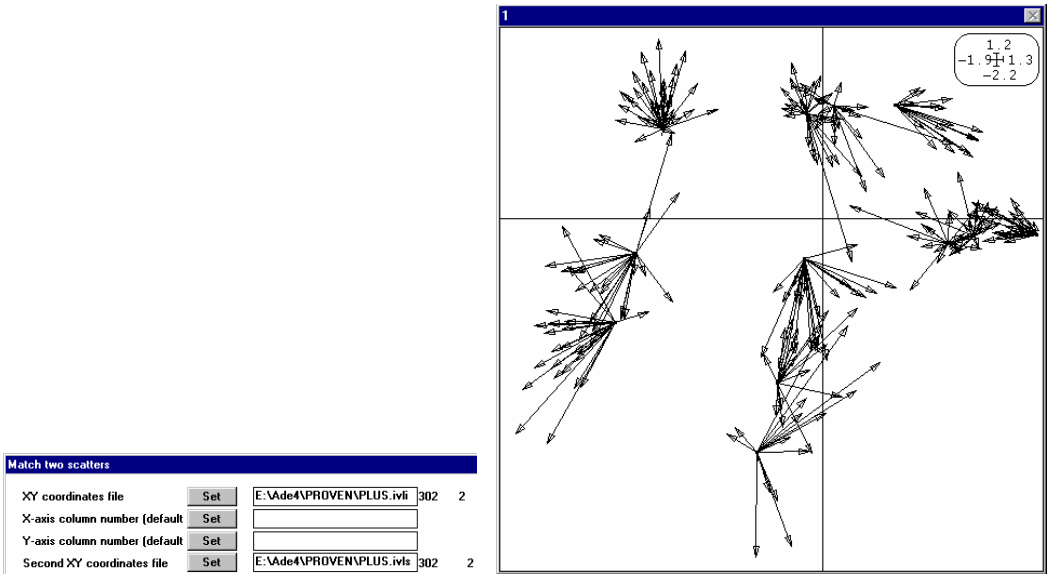
ScatterClass: Stars sur le fichier ivls donne avec le plan d'expériences :



Il faut aussi croiser les deux variables (TextToBin: LabelToCateg).



L'ensemble est obtenu après superposition du fichier ivli. On peut aussi faire la figure superposant ivls et ivli :

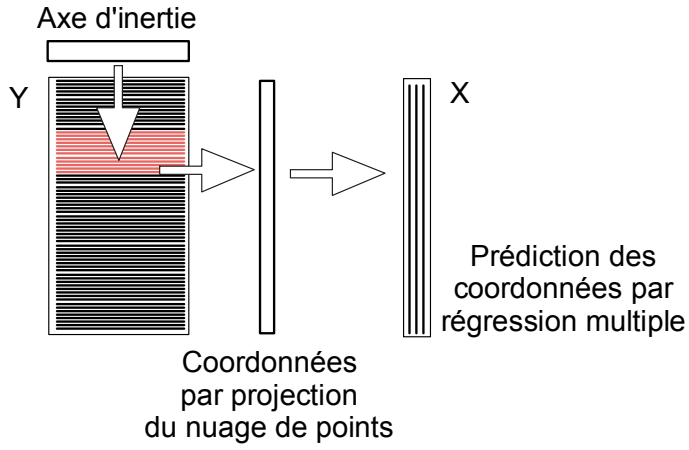


Superposer ivli et ivls signifie qu'on veut une analyse dont les coordonnées des lignes soient de la meilleure manière (aux moindres carrés pondérés) du type A+B. On peut imposer des contraintes emboîtées du type B sachant A ou $A \cap B^{\perp}$ [34] et décomposer l'inertie dans des plans complexes [35]. Ces méthodes ne sont pas d'un emploi ordinaire.

3.4. Analyse des Correspondances Non Symétriques

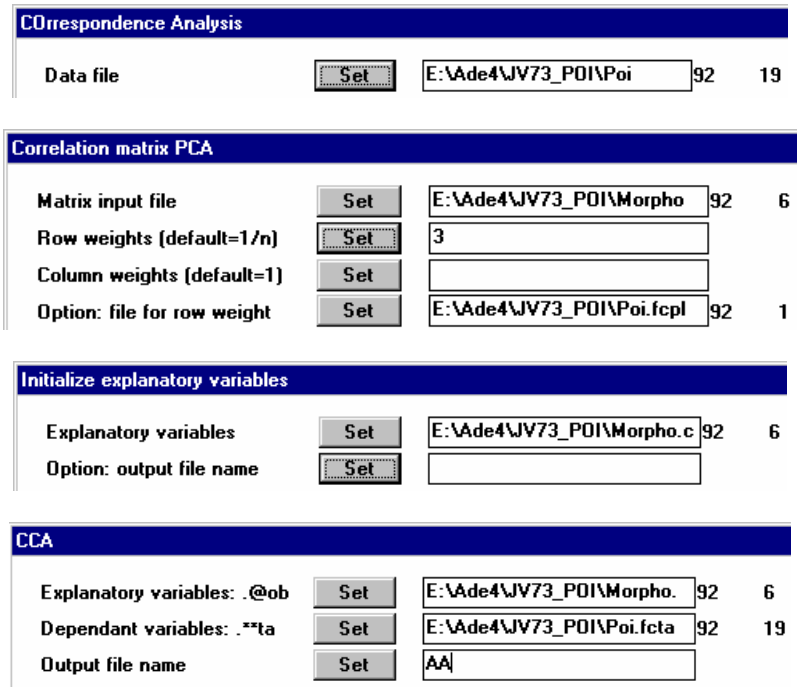
Pour avoir une vision coordonnée, on peut enregistrer que si l'AFC est l'analyse canonique entre les deux paquets d'indicatrices de classes, il existe deux ACPVI, une dans chaque sens, entre ces deux paquets d'indicatrices et qu'on obtient ainsi deux analyses non symétriques des correspondances. Origine dans [36], introduction en écologie dans [37] et en sciences sociales dans [38] (article vivement recommandé aux amateurs).

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

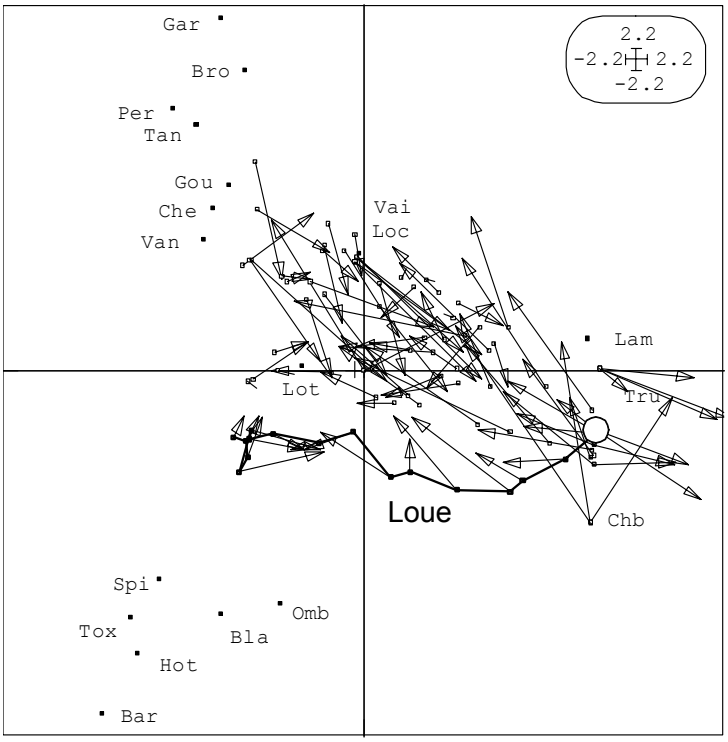


Par exemple, position des espèces avec des scores de variance unité, position des relevés à la moyenne et prédiction de ces positions par régression sur les variables de

milieu. Faire l'AFC du tableau faunistique, puis l'ACP normée du tableau de milieu en utilisant les poids de la précédente, puis utiliser CCA :

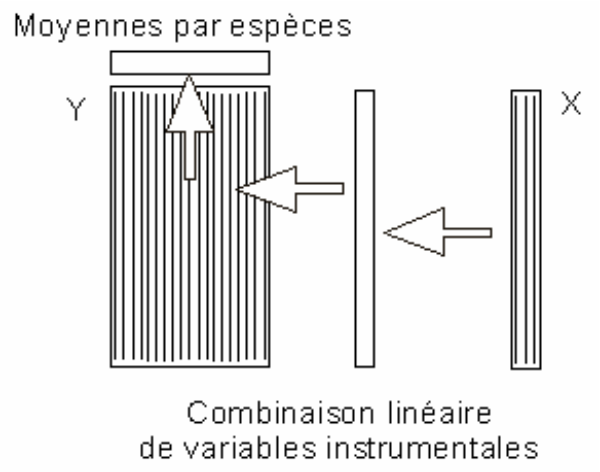



```
| files AA.ivc1
|       AA.ivls
|       AA.ivli
| allow a convenient interpretation
```

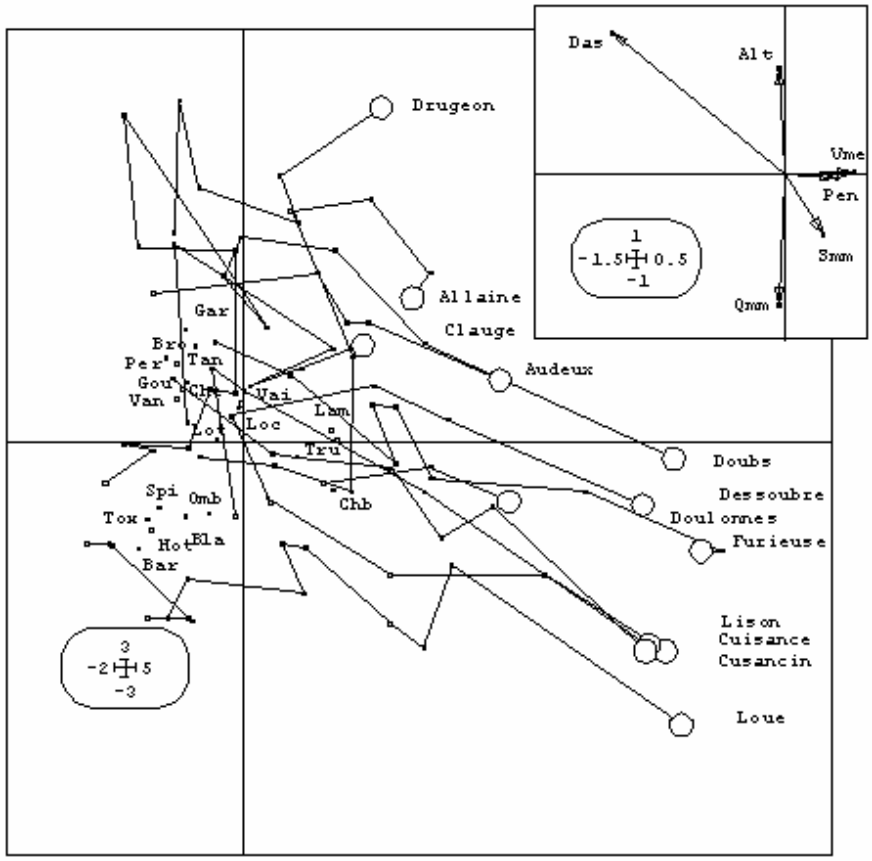


Double représentation des relevés en ACC. Carrés noirs: position des espèces avec des codes normalisés. Carrés blancs: position des relevés par averaging. L'extrémité du trait indique la prévision de la position précédente faite par régression multiple sur les variables de milieu.

Mais on a aussi le principe [11] :



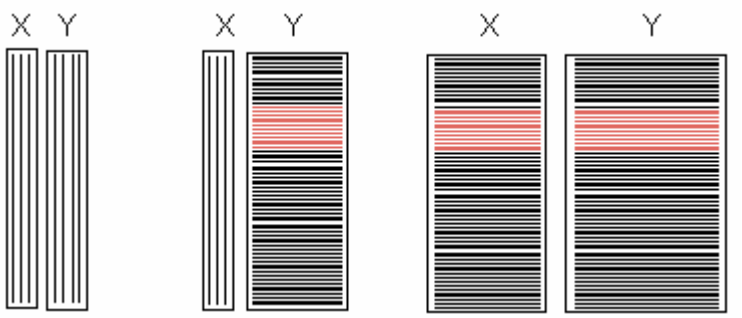

```
| files AA.ivfa
|       AA.ivl1
|       AA.ivco
| allow a convenient interpretation
|-----------------------------------------------
```



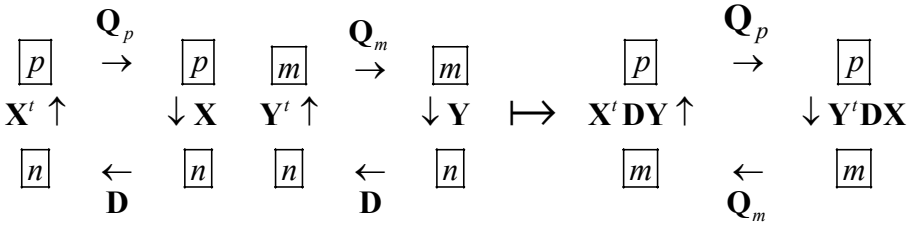
Représentation simultanée des objets en ACC. Flèches : coefficients des combinaisons linéaires (loadings). Carrés blancs: position des relevés avec des variances unité. Carrés noirs: position des espèces par averaging.

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.

4. Stratégie de la co-inertie



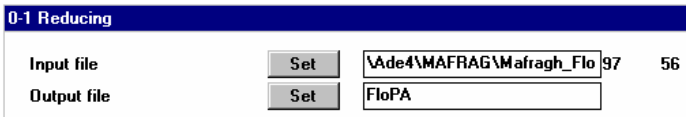
La troisième stratégie est la seule à tolérer des dimensions quelconques des deux côtés et est pratiquement la seule utilisable si les variables de milieu sont qualitatives (ce sont alors l'effectif des modalités qui définit la dimension du tableau **X**). C'est la plus simple puisqu'en gros elle fait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



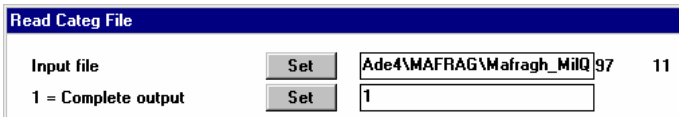
On peut étudier quatre exemples.

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 des cartes mafragh [39]. Le tableau floristique est passé en présence-absence :

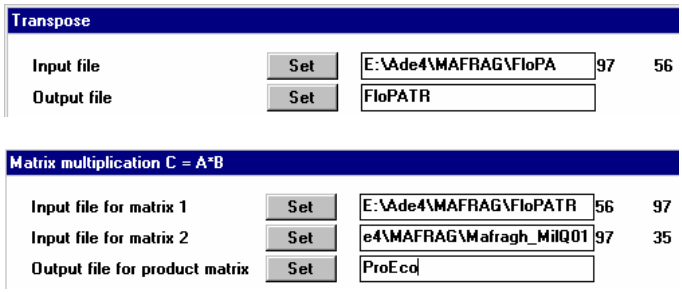


Le tableau de milieu qualitatif est lu puis passé en disjonctif complet :




```
Variable number 1 has 4 categories
-------------------------------------------------
[  1]Category:    1 Num:    21 Freq.:      0.216
[  2]Category:    2 Num:    18 Freq.:      0.186
[  3]Category:    3 Num:    22 Freq.:      0.227
[  4]Category:    4 Num:    36 Freq.:      0.371
...
```

Transposer le tableau faunistique et multiplier les matrices :

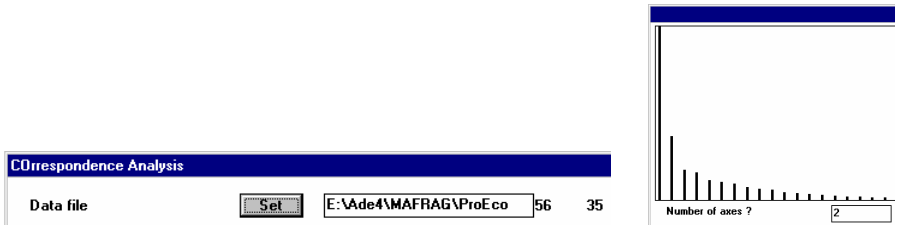



```
Output file: ProEco
--- Number of rows: 56, columns: 35
```

Éditer ce tableau :

01 Scirpus maritimus L.	6	3	3	22	13	11	6	4	29	5	25	9	3	23	8	1	11	14	8	14	20	2	11	13	8	19	15	4	19	11	1	22	8	3	0
02 Scirpus littoralis Schrad.	0	1	1	7	7	1	1	0	8	1	3	6	1	3	5	0	3	2	4	1	8	0	4	3	2	3	6	1	6	2	0	5	3	1	0
03 Juncus maritimus Lamk.	3	2	3	10	3	8	5	2	15	3	14	4	2	13	3	0	5	9	4	9	9	1	5	8	4	10	8	3	8	7	2	13	3	0	0
04 Aeluropus littoralis (Gouan) Parl.	7	5	2	18	9	12	6	5	26	6	25	7	7	19	6	5	9	12	6	14	18	6	7	13	6	20	12	7	16	9	1	15	10	3	3
05 Alisma Plantago-aquatica L.	0	0	2	6	3	3	2	0	8	0	7	1	2	4	2	0	1	6	1	4	4	0	5	2	1	2	6	3	3	2	0	5	3	0	0
06 Damasonium Alisma Mill.	3	1	1	3	1	3	1	3	6	2	7	1	0	7	1	1	4	3	0	4	4	2	3	2	1	5	3	2	5	1	0	5	2	1	0
07 Phragmites communis Trin.	2	2	0	2	1	1	2	2	5	1	4	2	1	5	0	1	2	3	0	5	1	1	1	1	3	5	1	1	2	3	0	5	0	0	1
08 Heleocharis uniglumis (Link.) Schult.	0	0	0	5	4	1	0	0	5	0	4	1	0	4	1	0	1	4	0	0	5	0	1	3	1	4	1	0	3	2	0	2	3	0	0
09 Salicornia arbica L.	2	3	6	0	0	2	7	2	10	1	9	2	0	6	5	0	5	4	2	7	4	1	5	2	3	4	7	1	6	4	4	6	0	0	1
10 Cressa cretica L.	0	2	1	3	2	1	2	1	6	0	5	1	0	5	1	0	1	4	1	4	2	0	1	2	3	4	2	0	2	4	0	5	1	0	0
11 Plantago coronopus L.	1	2	3	0	0	0	4	2	5	1	4	2	0	4	2	0	3	1	2	5	1	0	2	1	3	2	4	0	2	4	2	3	1	0	0
12 Lolium rigidum Gaud.	1	5	3	8	4	5	4	4	16	1	11	6	2	10	5	1	6	5	5	14	3	1	7	5	4	6	11	3	9	5	2	6	3	4	2
13 Medicago intertexta (L.) Mill.	10	9	15	11	7	12	17	9	37	8	36	9	11	27	7	12	21	9	3	41	4	18	15	7	5	27	18	19	20	6	4	7	8	11	15
14 Crypsis alopecuroides (Pill. et Mitt.) Schrad.	2	2	1	1	1	0	1	4	6	0	5	1	0	5	1	1	2	2	1	5	1	1	3	1	1	4	2	1	3	2	0	4	1	0	1
15 Carlina racemosa L.	4	1	4	4	1	3	4	5	11	2	11	2	3	9	1	2	6	4	1	11	2	2	8	2	1	7	6	4	7	2	2	3	4	1	3
16 Eryngium Barrelieri Boiss.	1	0	1	2	1	2	1	0	3	1	4	0	0	4	0	0	1	3	0	4	0	0	2	1	1	2	2	0	3	1	1	2	0	1	0
17 Carlina lanata L.	3	1	6	1	0	3	5	3	10	1	11	0	3	7	1	3	5	3	0	10	1	4	5	1	1	6	5	5	5	1	2	4	3	1	1
19 Hordeum maritimum Witth.	5	5	4	7	4	5	5	7	17	4	16	5	5	13	3	4	8	5	4	18	3	4	10	5	2	10	11	7	11	3	1	7	5	4	4
20 Phalaris caerulescens Desf.	7	7	11	8	5	12	9	7	27	6	27	6	8	20	5	9	12	9	3	29	4	13	10	7	3	18	15	14	16	3	2	9	7	6	9
21 Centaurium spicatum (L.) Fritsch.	5	7	7	10	7	8	8	6	24	5	22	7	5	20	4	6	14	5	4	27	2	10	11	6	2	17	12	12	13	4	1	6	6	5	11
22 Cynosurus cristatus Poiret non L.	1	0	0	0	0	1	0	0	0	1	1	0	0	1	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0
23 Urginea maritima (L.) Baker	1	1	4	3	2	3	4	0	8	1	9	0	0	9	0	3	4	2	0	9	0	4	4	1	0	7	2	5	4	0	0	2	4	0	3
24 Trifolium maritimum Huds.	2	2	3	1	0	2	4	2	6	2	6	2	2	3	3	1	5	1	1	5	3	2	4	0	2	2	6	2	4	2	1	4	2	1	0
25 Atriplexhastata L.	0	1	0	5	4	1	0	1	6	0	1	5	1	3	2	1	1	2	2	5	1	1	2	1	2	2	4	2	1	3	0	1	2	3	0
26 Narcissus Tazetta L.	0	0	3	1	0	1	3	0	4	0	4	0	0	4	0	0	2	2	0	4	0	0	2	1	1	2	2	0	3	1	1	1	1	0	1
27 Scilla autumnalis L.	1	5	1	5	2	4	1	5	11	1	7	5	1	8	3	1	3	3	5	10	2	1	4	5	2	5	7	2	6	4	0	5	1	4	2
28 Ormenis praecox (Link.) Briq.	5	3	3	3	2	3	6	3	9	5	13	1	4	10	0	4	9	1	0	14	0	8	5	1	0	10	4	8	6	0	0	3	3	2	6
30 Asphodelus microcarpus Salzm. et Viv.	0	0	1	2	1	1	1	0	3	0	2	1	0	2	1	0	2	0	1	3	0	0	2	1	0	1	2	1	2	0	0	0	1	2	0
31 Mentha Pulegium L.	2	1	0	0	1	1	0	1	1	2	3	0	1	2	0	1	2	0	0	3	0	0	3	0	0	2	1	1	2	0	0	0	2	1	0
32 Ridolfia segetum Moris.	3	3	1	2	3	2	3	1	4	5	7	2	2	7	0	6	2	0	1	9	0	8	0	0	1	8	1	8	0	1	0	2	4	1	2
33 Atriplex littoralis L.	2	2	1	5	4	1	3	2	9	1	5	5	2	6	2	0	3	5	2	9	1	1	3	3	3	5	5	1	5	4	0	4	1	3	2
34 Beta vulgaris ssp. maritima (L.) Batt.	9	4	4	4	1	5	6	9	17	4	21	0	6	14	1	6	8	5	2	20	1	9	8	2	2	12	9	10	8	3	1	4	4	2	10
35 Scolymus hispanicus L.	1	5	7	2	1	4	7	3	12	3	14	1	3	9	3	2	7	3	3	13	2	5	5	3	2	10	5	5	7	3	1	2	2	7	3
36 Cichorium Intybus L.	2	0	1	1	1	1	2	0	2	2	3	1	4	0	0	1	2	1	0	4	0	2	1	1	0	3	1	2	2	0	0	0	0	2	2
37 Inula crithmoides L.	2	4	1	4	3	4	1	3	8	3	8	3	4	4	3	3	5	2	1	9	2	4	5	1	1	5	6	5	5	1	0	2	4	2	3
38 Hedysarum coronarium L.	2	1	2	4	4	3	1	1	8	1	4	5	3	5	1	6	3	0	0	9	0	7	2	0	0	7	2	8	1	0	0	0	3	3	3
39 Picris echioides L.	7	7	5	4	3	5	11	4	17	6	20	3	10	11	2	7	11	4	1	22	1	12	5	6	0	18	5	12	9	2	0	3	3	8	9
40 Convolvulus arvensis L.	4	4	3	2	3	3	6	1	7	6	12	1	6	6	1	5	6	2	0	13	0	8	2	3	0	11	2	9	3	1	0	1	3	4	5
41 Galactites Duriaei Spach.	0	3	3	1	1	1	5	0	6	1	7	0	1	6	0	0	4	3	0	7	0	2	3	1	1	3	4	2	4	1	1	4	1	0	1
42 Arthrocnemum indicum (Willd.) Moq.	0	2	5	0	0	1	6	0	7	0	5	2	0	3	4	0	3	2	2	4	3	0	3	1	3	1	6	0	3	4	4	2	0	0	1
43 Sonchus oleraceus L.	4	5	4	3	3	4	7	2	11	5	14	2	5	8	3	4	9	2	1	16	0	10	2	4	0	12	4	10	5	1	0	2	3	6	5
44 Rumex Acetosella L. ssp. angiocarpus Murb.	5	2	5	1	3	1	8	1	8	5	12	1	4	8	1	3	5	5	0	13	0	8	2	2	1	11	2	8	3	2	1	1	1	4	6
45 Avenafatua L.	4	2	3	2	3	2	6	0	5	6	9	2	4	7	0	5	5	1	0	11	0	8	2	1	0	11	0	8	3	0	0	1	4	3	3
46 Diplotaxis erucoides (L.) DC	5	3	4	1	3	4	5	1	7	6	12	1	6	6	1	4	8	1	0	13	0	10	1	2	0	13	0	10	3	0	0	1	5	4	3
47 Stachys marrubifolia Viv.	1	2	3	1	1	2	3	1	7	0	6	1	0	6	1	3	4	0	0	7	0	6	1	0	0	6	1	6	1	0	0	1	3	2	1
48 Arisarum vulgare Targ. Tozz. ssp. exsertum	0	1	3	0	0	3	1	0	4	0	4	0	1	3	0	2	2	0	0	4	0	3	0	1	0	4	0	3	1	0	0	0	2	2	0
49 Borago off cinalis L.	3	1	6	1	2	4	4	1	9	2	10	1	4	6	1	5	5	1	0	11	0	8	1	2	0	11	0	8	3	0	0	1	4	6	0
50 Convolvulus tricolor L.	1	1	2	2	2	2	1	1	6	0	4	2	0	5	1	4	2	0	0	6	0	6	0	0	0	6	0	6	0	0	0	0	4	2	0
51 Narcissus pseudo-narcissus L.	1	0	1	0	1	0	1	0	1	1	2	0	0	2	0	0	1	1	0	2	0	1	1	0	0	2	0	1	1	0	0	0	0	1	1
53 Leotondon tuberosus L.	0	0	2	0	0	0	2	0	2	0	2	0	0	2	0	0	1	1	0	2	0	0	1	0	1	1	1	0	1	1	1	0	1	0	0
54 Ammi Visnaga Lamk.	0	1	1	0	0	0	2	0	2	0	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0	0	2	0	2	0	0	1	0	0	1
55 Lythrum tribracteatum Salzm.	0	2	0	0	0	1	1	0	1	1	2	0	0	1	1	0	1	1	0	2	0	0	1	1	0	1	1	0	1	1	0	0	1	0	1
56 Lolium multiflorum Lamk.	1	1	0	0	0	0	1	1	2	0	1	1	1	1	0	1	1	0	0	2	0	1	0	1	0	2	0	1	1	0	0	0	0	0	2
57 Cynodon dactylon (L.) Pers.	0	1	0	1	1	1	0	0	1	1	2	0	1	1	0	0	1	1	0	2	0	1	0	1	0	2	0	1	1	0	0	2	0	0	0
58 Chrozophora tinctoria L.	1	1	0	0	1	1	0	0	0	2	2	0	1	1	0	1	1	0	0	2	0	2	0	0	0	2	0	2	0	0	0	1	1	0	0
59 Juncus acutus L.	0	0	2	0	0	1	1	0	2	0	1	1	0	2	0	0	1	1	0	1	1	0	1	0	1	0	2	0	1	1	2	0	0	0	0

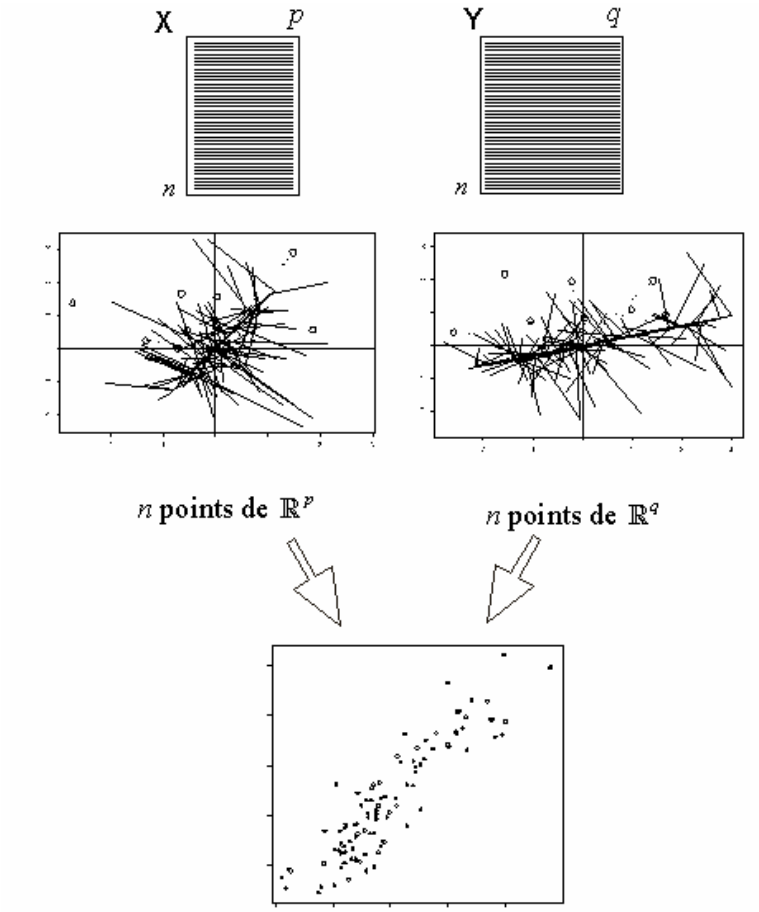
Il contient des profils écologiques bruts (nombre de stations contenant l'espèce dans chaque classe de milieu). On peut comparer chacun d'entre eux à la somme totale (nombre de présences d'espèces dans chaque classe de milieu) ou à la distribution des stations par classe de milieu. *Scirpus maritimus* a 34 présences dont 22 dans la classe 4 de la variable 1. Pour cette variable, le profil des stations est 21, 18, 22 et 36. Pour cette même variable, le décompte des présences totales est 132, 129, 158 et 189. C'est le profil moyen de toutes les espèces pour cette variable de milieu. On peut comparer tous les profils sur toutes les variables en faisant l'AFC de ce tableau (origine dans [6] et « redécouverte » dans [7]).



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 ([8]). 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 interbatterie

A l'origine de l'analyse de co-inertie, on trouve l'analyse interbatterie (résultats sur les mêmes individus de deux batteries de tests psychotechniques) de Tucker [40]. 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 maximisé 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 :

$$\begin{array}{ccccc} \boxed{p} & \xrightarrow{\mathbf{I}_p} & \boxed{p} & \xrightarrow{\mathbf{I}_m} & \boxed{m} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} & & \mathbf{Y}^t \uparrow \\ \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} \end{array} \quad \mapsto \quad \begin{array}{ccc} \boxed{p} & \xrightarrow{\mathbf{I}_p} & \boxed{p} \\ \frac{1}{n}\mathbf{X}^t\mathbf{Y} \uparrow & & \downarrow \frac{1}{n}\mathbf{Y}^t\mathbf{X} \\ \boxed{m} & \xleftarrow{\mathbf{I}_m} & \boxed{m} \end{array} \quad \begin{array}{ccccc} \boxed{p} & \xrightarrow{\mathbf{I}_p} & \boxed{p} & \xrightarrow{\mathbf{I}_m} & \boxed{m} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} & & \mathbf{Y}^t \uparrow \\ \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} & \xleftarrow{\frac{1}{n}\mathbf{I}_n} & \boxed{n} \end{array}$$

Le schéma de principe se résume dans la figure précédente.

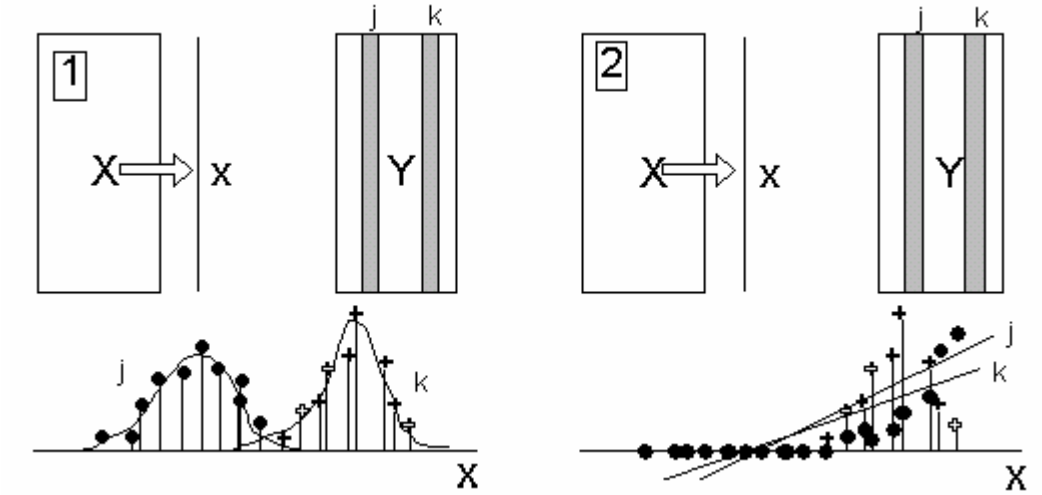
4.3. AFC des tableaux de Burt croisés

Un autre cas fondamental est celui de l'AFC des tableaux de Burt croisés que Cazes [41] 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 définit 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 [42], ce principe est aussi facile d'emploi que l'analyse d'un tableau.

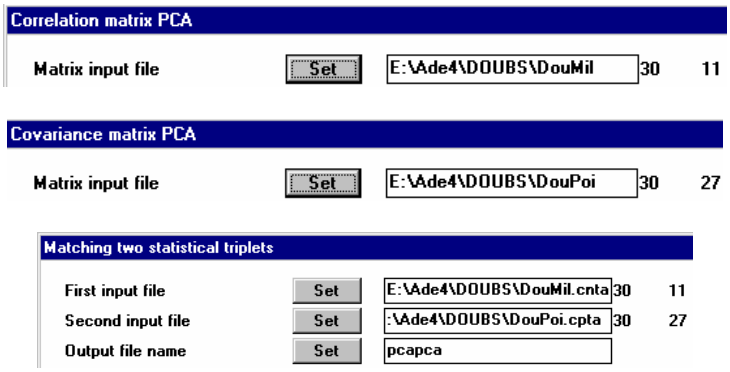
4.4. Analyse des niches écologiques

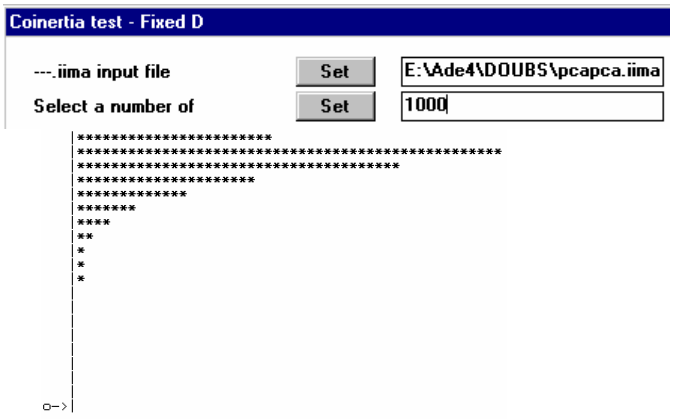
Pour croiser un tableau floro-faunistique et un tableau de variables mésologiques, on pensera surtout au modèle de liaison qu'on désire :



A gauche : Courbes de réponse non linéaire 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.

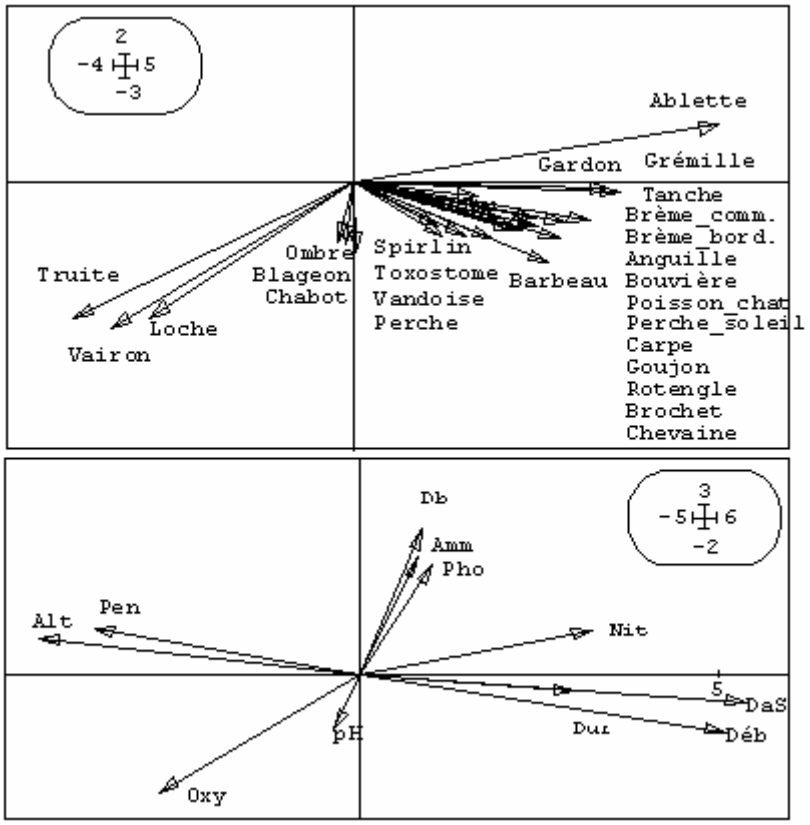
Reprendre l'exemple de l'introduction. Faire le couplage de deux ACP.





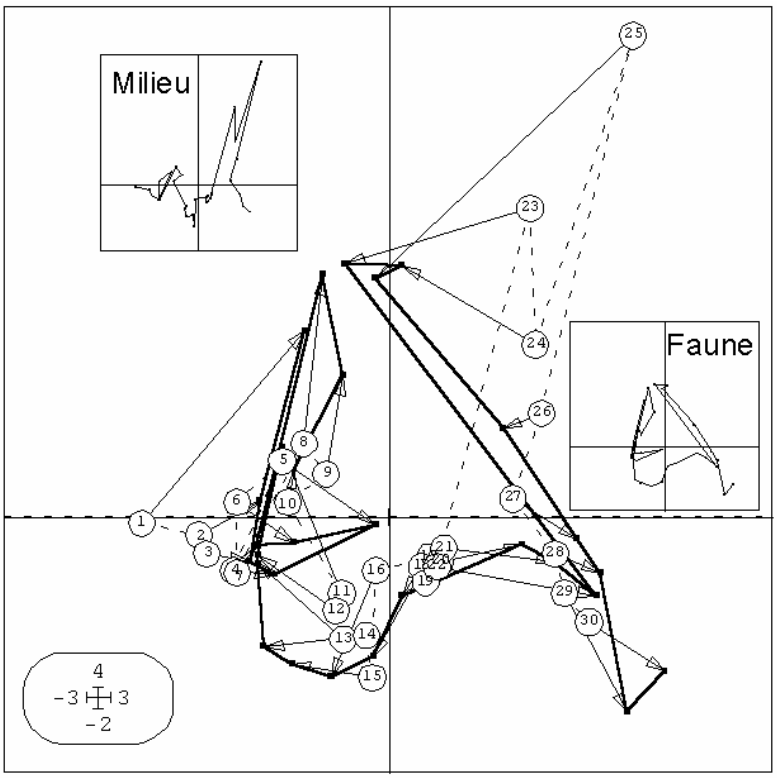

```
DiagoRC: General program for two diagonal inner product analysis
Input file: E:\Ade4\DOUBS\pcapca.iita
--- Number of rows: 27, columns: 11
-----------------------
Total inertia:  134.703
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +1.1902E+02 +0.8836 +0.8836   |02   +1.3871E+01 +0.1030 +0.9865 |
03   +7.5658E-01 +0.0056 +0.9922   |04   +5.2783E-01 +0.0039 +0.9961 |
05   +2.7089E-01 +0.0020 +0.9981   |06   +1.6462E-01 +0.0012 +0.9993 |
07   +6.6011E-02 +0.0005 +0.9998   |08   +1.8209E-02 +0.0001 +0.9999 |
09   +4.6041E-03 +0.0000 +1.0000   |10   +2.7934E-03 +0.0000 +1.0000 |
11   +0.0000E+00 +0.0000 +1.0000

File E:\Ade4\DOUBS\pcapca.iivp contains the eigenvalues and relative inertia
for each axis
--- It has 11 rows and 2 columns

File E:\Ade4\DOUBS\pcapca.iico contains the column scores
--- It has 11 rows and 2 columns

File E:\Ade4\DOUBS\pcapca.iili contains the row scores
--- It has 27 rows and 2 columns
```

Utiliser iili et iico pour aborder directement les deux paquets de variables :



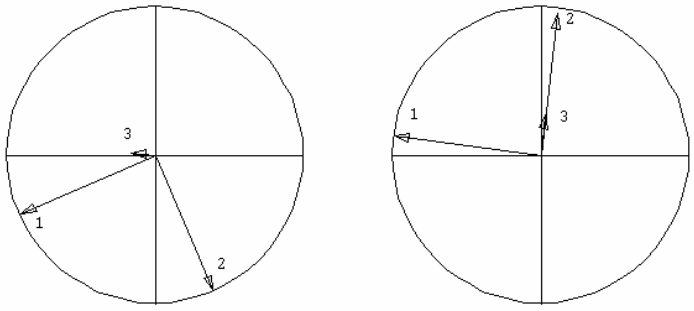

```
-------------------------------------------------------
Co-inertia analysis between two statistical triplets
   1  --->  E:\Ade4\DOUBS\DouMil.cnta  (rows:  30,  col:  11,  axes:  3,  inertia:
11.000000)
   2  --->  E:\Ade4\DOUBS\DouPoi.cpta  (rows:  30,  col:  27,  axes:  3,  inertia:
66.077797)
Co-inertia: 134.7, RV coefficient: 0.45056

E:\Ade4\DOUBS\pcapca.iiw1 is a binary file with 11 rows and 2 columns
It contains the canonical weights of variables of table 1

E:\Ade4\DOUBS\pcapca.iiw2 is a binary file with 27 rows and 2 columns
It contains the canonical weights of variables of table 2

E:\Ade4\DOUBS\pcapca.iil1 is a binary file with 30 rows and 2 columns
It contains the coordinates of the rows (table 1)

E:\Ade4\DOUBS\pcapca.iil2 is a binary file with 30 rows and 2 columns
It contains the coordinates of the rows (table 2)
```

Utiliser iim1 et iim2 pour aborder simultanément les deux nuages de points :




```
E:\Ade4\DOUBS\pcapca.iim1 is a binary file with 30 rows and 2 columns
It contains the normalized coordinates of the rows (table 1)

E:\Ade4\DOUBS\pcapca.iim2 is a binary file with 30 rows and 2 columns
It contains the normalized coordinates of the rows (table 2)

E:\Ade4\DOUBS\pcapca.iia1 is a binary file with 3 rows and 2 columns
It contains the coordinates of the projections of inertia
axes onto the co-inertia axes (table 1)

E:\Ade4\DOUBS\pcapca.iia2 is a binary file with 3 rows and 2 columns
It contains the coordinates of the projections of inertia
axes onto the co-inertia axes (table 2)
```

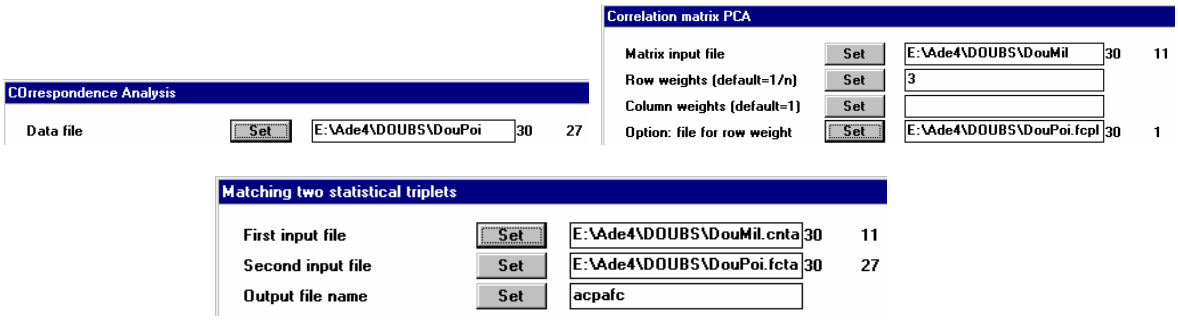
Utiliser iia1 et iia2 pour replacer les analyses simples par rapport au couplage (on voit comment se sont associées les deux analyses de base) :



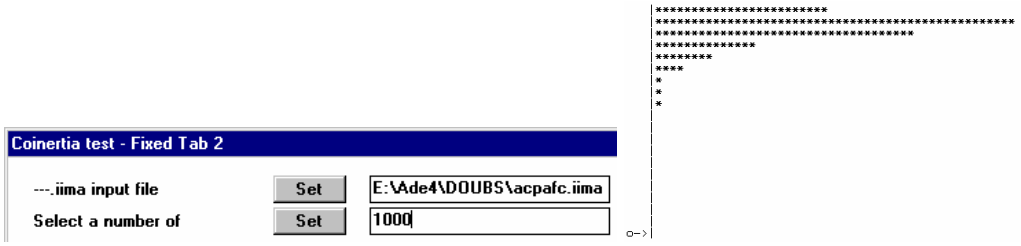
Num	Covaria.	Varian1	varian2	Correla.	INER1	INER2
1	10.91	5.412	41.25	0.7302	6.322	42.75
2	3.724	2.839	8.201	0.7718	2.232	8.158

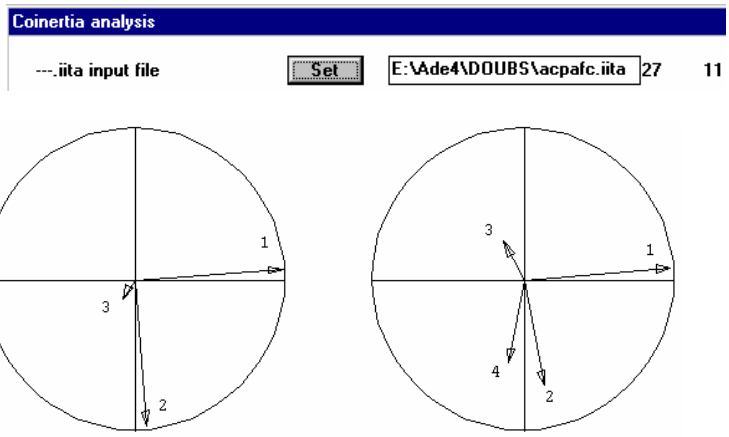
A commenter.

Coupler une ACP normée et une AFC :

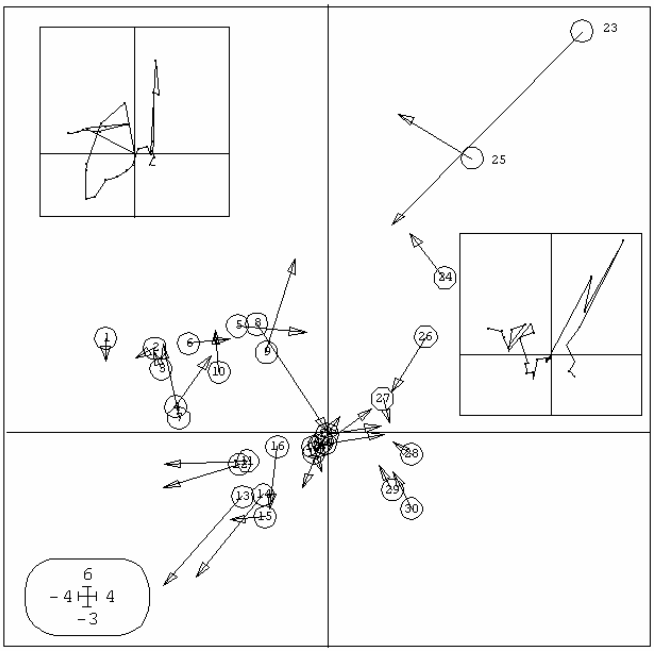


IMPORTANT : le test de permutation sur la trace

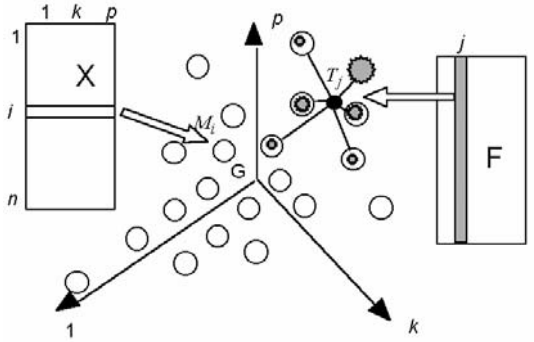




Num	Covaria.	Varian1	varian2	Correla.	INER1	INER2
1	1.53	5.598	0.5763	0.8519	5.728	0.601
2	0.4183	2.351	0.1113	0.8176	2.426	0.1444



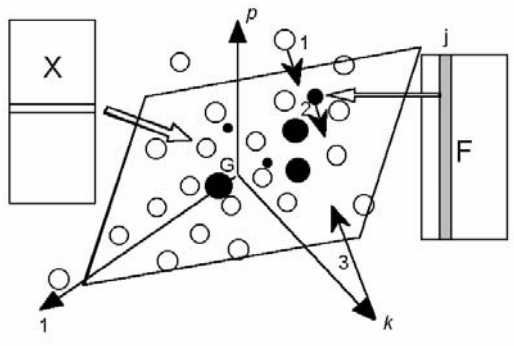
Il vaut mieux se méfier de l'axe 2 de cette analyse. L'analyse OMI est aussi une analyse de co-inertie. Son principe repose sur la figure :



*Eléments de base dans une analyse OMI. 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).*

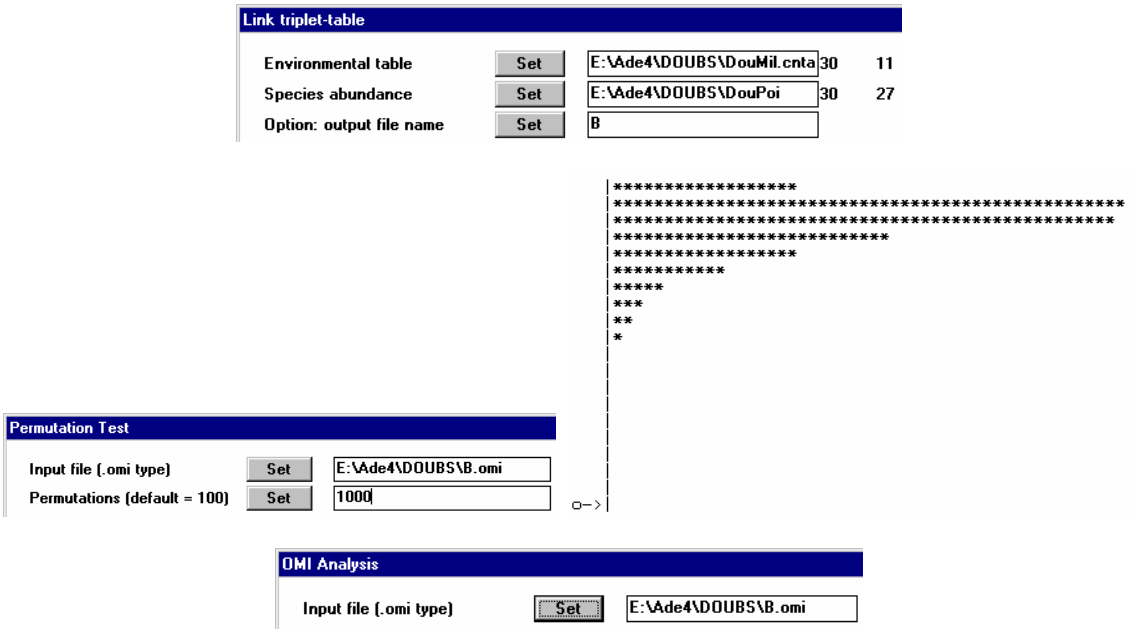
On évite ainsi, comme on l'a fait dans l'analyse précédente de centrer et réduire les variables de milieu avec les poids du tableau faunistique. Les stations sont d'abord considérées sans référence avec ce qu'on y trouve. Ce sont simplement des points d'échantillonnage du milieu.

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 [43].



*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.*

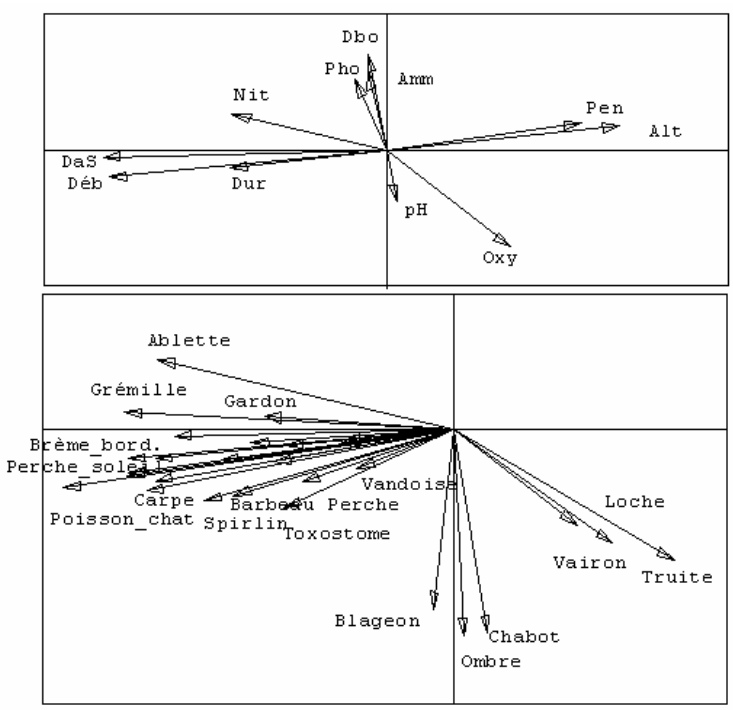
Dans le module Niche :



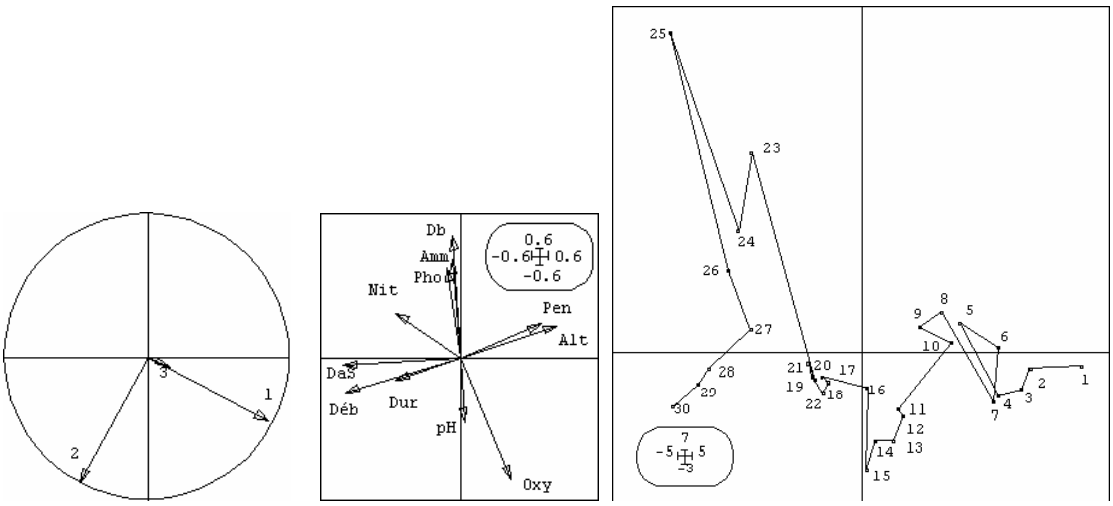

```
Environmental data: E:\Ade4\DOUBS\DouMil.cnta
Sites: 30   Variables: 11
Species abundance: E:\Ade4\DOUBS\DouPoi
Sites: 30   Species: 27

Mean position of each species on each variable in the file
E:\Ade4\DOUBS\B.nata
Row=species: 27   Col=variable: 11

Weights of species in file  E:\Ade4\DOUBS\B.napl
Row=species: 27   Col=1

Weights of variables in file  E:\Ade4\DOUBS\B.napc
(from input environmental triplet)
Row=variable: 11   Col=1
-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: E:\Ade4\DOUBS\B.nata
--- Number of rows: 27, columns: 11
-----------------------
Total inertia:  3.03502
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.6725E+00 +0.8805 +0.8805   |02   +3.0548E-01 +0.1007 +0.9812 |
```

```
03   +2.9327E-02 +0.0097 +0.9909   |04   +1.5938E-02 +0.0053 +0.9961 |
05   +5.2015E-03 +0.0017 +0.9978   |06   +4.3056E-03 +0.0014 +0.9992 |
07   +1.5703E-03 +0.0005 +0.9998   |08   +5.5805E-04 +0.0002 +0.9999 |
09   +1.0100E-04 +0.0000 +1.0000   |10   +7.1023E-05 +0.0000 +1.0000 |
11   +0.0000E+00 +0.0000 +1.0000

File E:\Ade4\DOUBS\B.navp contains the eigenvalues and relative inertia for
each axis
--- It has 11 rows and 2 columns

File E:\Ade4\DOUBS\B.naco contains the column scores
--- It has 11 rows and 2 columns

File E:\Ade4\DOUBS\B.nali contains the row scores
--- It has 27 rows and 2 columns
```

Utiliser nali et naco pour aborder directement les deux paquets de variables :

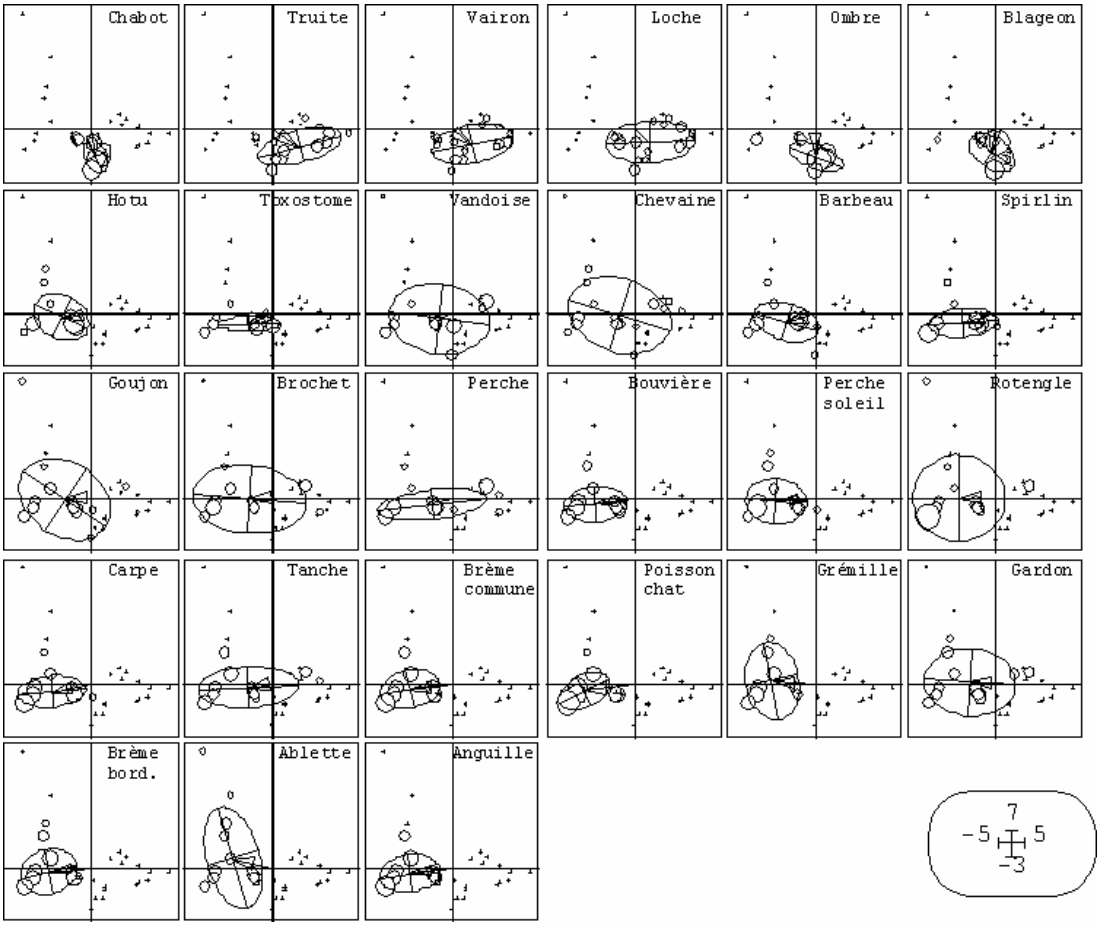



```
File E:\Ade4\DOUBS\B.nac1 contains the column scores with unit norm
It has 11 rows and 2 columns
File :E:\Ade4\DOUBS\B.nac1


File E:\Ade4\DOUBS\B.nals contains the coordinates of the sites
It has 30 rows and 2 columns
```



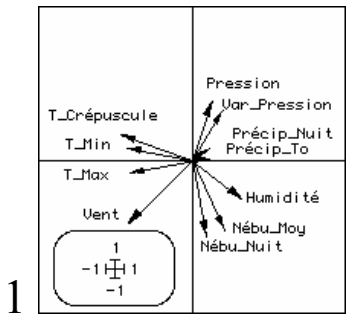

```
File E:\Ade4\DOUBS\B.naax contains the coordinates of the sites
It has 3 rows and 2 columns
```

L'espace des relevés de milieu est l'espace commun de représentation de tous les objets.



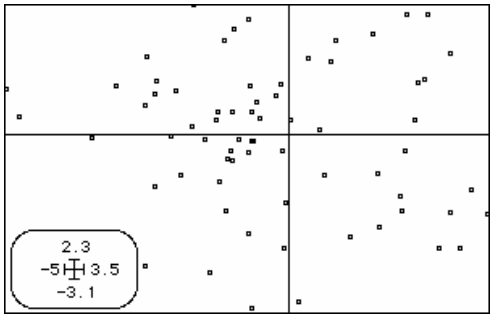
C'est sans doute la meilleure manière d'exprimer simultanément la pollution comme facteur limitant et le gradient amont-aval comme facteur de séparation de niche.

Refaire l'exercice sur la carte Light_tr [44]



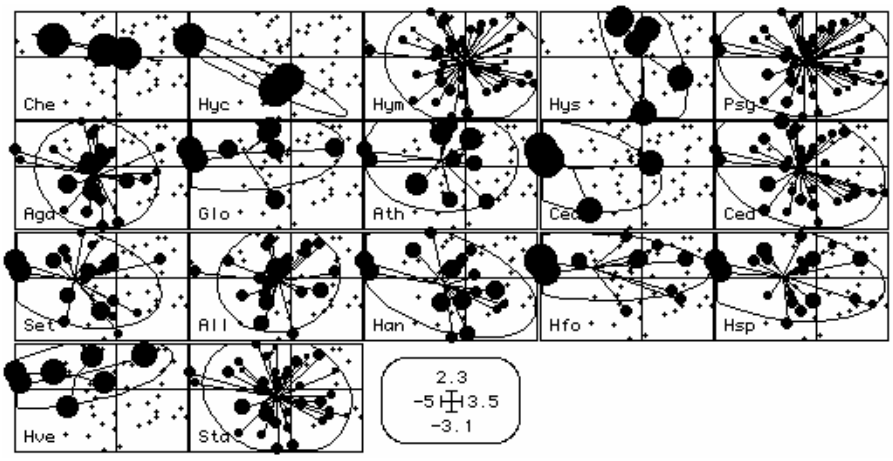
1

Coefficients des combinaisons linéaires de variables normalisées (comme dans une ACP normée)



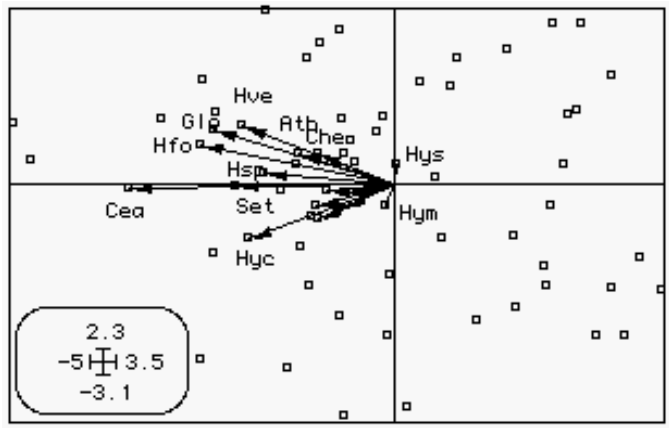
2

Positions des relevés (nuit de piégeage lumineux) par les combinaisons linéaires (comme dans une ACP normée)



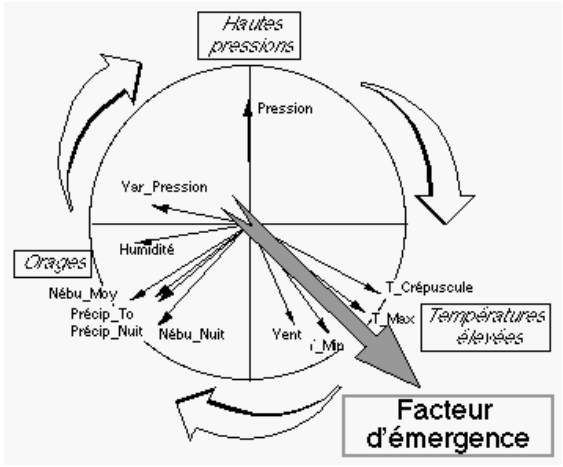
3

Représentation de la variation d'abondance des taxons sur le plan précédent. Les coefficients des variables optimisent la somme des carrés des distances à l'origine des positions moyennes des espèces.



4

Synthèse de la position moyenne des espèces dans le plan des relevés



5

Interprétation de l'axe de l'analyse OMI dans le plan des variables de l'ACP normée de départ.

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*[45].

5. Références

1 Dagnélie, 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.

2 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. 1-257.

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

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

5 Godron, M., Daget, P., Emberger, L., Le Floch, E., Poissonet, J., Sauvage, C. & Wacquant, J.P. (1968) *Relevé méthodique de la végétation et du milieu*. Editions du CNRS, Paris. 1-292. Gounot, M. (1969) *Méthodes d'étude quantitative de la végétation*. Masson, Paris. 1-314.

6 Romane, F. (1972b) Utilisation de l'analyse multivariable en Phytoécologie. *Investigaciön pesquera* : 36, 131-139.

7 Montana, C. & Greig-Smith, P. (1990) Correspondence analysis of species by environmental variable matrices. *Journal of Vegetation Science* : 1, 453-460.

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

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

10 Gittins, R. (1985) *Canonical analysis, a review with applications in ecology*. Springer-Verlag, Berlin. 1-351.

11 Barkham, J.P. & Norris, J.M. (1970) Multivariate procedures in an investigation of vegetation and soil relations of two beach woodlands, Costwold Hills, England. *Ecology* : 51, 4, 630-639.

12 Esteve, J. (1978) Les méthodes d'ordination : éléments pour une discussion. In : *Biométrie et Ecologie*. Legay, J.M. & Tomassone, R. (Eds.) Société Française de Biométrie, Paris. 223-250.

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

14 Fisher, R.A. (1940) The precision of discriminant functions. *Annals of Eugenics* : 10, 422-438. Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika* : 39, 274-289.

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

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

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

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

19 Tomassone, R., Danzard, M., Daudin, J.J. & Masson, J.P. (1988) *Discrimination et classement*. Masson, Paris. 1-173.

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

21 Hill, M.O. (1977) Use of simple discriminant functions to classify quantitative phytosociological data. In : *Proceedings of the First International Symposium on Data Analysis and Informatics*. Diday, E. (Ed.) INRIA Rocquencourt, France. 181-199.

22 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.

23 Pontier, J., Dufour, A.B. & Normand, M. (1990) *Le modèle euclidien en analyse des données*. SMA, édition Ellipses, Bruxelles. 1-428.

24 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.

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

25 Birks, H.J.B. & Austin, H.A. (1992) An annoted bibliography of canonical correspondence analysis and related constrained ordination methods (1986-1991). Botanical Institute, Allégaten 41, N-5007 Bergen, Norway. 1-29.

26 Lebreton, J.D., Sabatier, R., Banco, G. & Bacou, A.M. (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. In : Applied Multivariate Analysis in SAR and Environmental Studies. Devillers, J. & Karcher, W. (Eds.) Kluwer Academic Publishers. 85-114.

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

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

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

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

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

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

33 Blondel, J., Chessel, D. & Frochot, B. (1988) Niche expansion and density compensation of island birds in mediterranean habitats. A case study from comparison of two ecological successions. *Ecology* : 69, 6, 1899-1917.

34 Yoccoz, N. & Chessel, D. (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.

35 Sabatier, R., Lebreton, J.D. & Chessel, J.D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In : *Multiway data analysis*. Coppi, R. & Bolasco, S. (Eds.) Elsevier Science Publishers B.V., North-Holland. 341-352.

36 Lauro, N. & D'Ambra, L. (1984) L'analyse non symétrique des correspondances. In *: Data Analysis and Informatics III*. Diday, E. & Coll. (Ed.) Elsevier, North-Holland. 433-446.

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

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

39 Belair, G. de & Bencheikh-Lehocine, M. (1987) Composition et déterminisme de la végétation d'une plaine côtière marécageuse : La Mafragh (Annaba, Algérie). *Bulletin d'Ecologie* : 18, 4, 393-407.

40 Tucker, L.R. . (1958) An inter-battery method of factor analysis. *Psychometrika* : 23, 2, 111-136.

41 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.

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

43 Dolédec, S., Chessel, D. & Gimaret, C. (2000) Niche separation in community analysis: a new method. *Ecology,* 81, 2914-2927.

44 Usseglio-Polatera, P. & Auda, Y. (1987) Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie* : 23, 1, 65-79.

45 Green, R.H. (1993) Relating two sets of variables in environmental studies. In : *Multivariate environmental statistics*. Patil, G.P. & Rao, C.R. (Eds.) North-Holland, Amsterdam. 149-163.