

Quand l'analyse canonique des correspondances est-elle une analyse canonique ?

Résumé

La fiche montre comment l'analyse canonique des correspondances est une analyse canonique quand on confronte une liste d'occurrences d'espèces avec un tableau de variables environnementales portant sur ces occurrences. Cela modifie sensiblement les conditions d'utilisation de la méthode quand elle est employée, comme c'est le cas général, sur des tableaux relevés-espèces.

Plan

| | |
|---|----|
| 1 — Introduction..... | 2 |
| 2 — Stratégies de couplages symétriques..... | 2 |
| 2.1 — Principe de l'analyse canonique | 4 |
| 2.2 — Pratique de l'analyse canonique..... | 7 |
| 3 — Stratégies de couplages dissymétriques..... | 11 |
| 3.1 — Dissymétrie de prédiction..... | 11 |
| 3.2 — Dissymétrie de centrage | 15 |
| 4 — Analyse canonique et liste d'occurrences..... | 20 |
| 4.1 — Présentation des données | 20 |
| 4.2 — L'AFC est une analyse canonique | 24 |
| 4.3 — L'ACC est une analyse discriminante | 24 |
| 4.4 — L'ACC est une analyse canonique..... | 27 |
| 5 — Conclusion | 29 |
| 5.1 — Quand les données sont des relevés..... | 29 |
| 5.2 — Quand les données sont des occurrences..... | 29 |
| Références | 30 |

D. Chessel & C. Gimaret

The most obvious procedures are multivariate regression analysis and canonical correlation analysis. Surprisingly, substantive applications of these methods in ecology are few.

Pourquoi donc une des principales méthodes d'analyse des données n'est-elle pas utilisée ? A la page 3, de la même introduction, la remarque surprenante qui suit est faite :

There is a further procedure for investigating relationships between two measurement domains. This is the inter-battery factor analysis of Tucker (1958)³ ... Browne (1979)⁴ has recently shown that, while the underlying rationales of inter-battery factor analysis and canonical analysis are entirely different, numerically the procedures are substantially equivalent.

On est à la source de bien des difficultés, surtout pour les pragmatistes : comment des principes très différents peuvent-ils conduire à des procédures très voisines ? L'analyse inter-batterie (IB) est la première version de l'analyse de co-inertie, l'analyse canonique des corrélations est la première version de l'analyse canonique. Les deux sont symétriques, sont voisines et différentes.

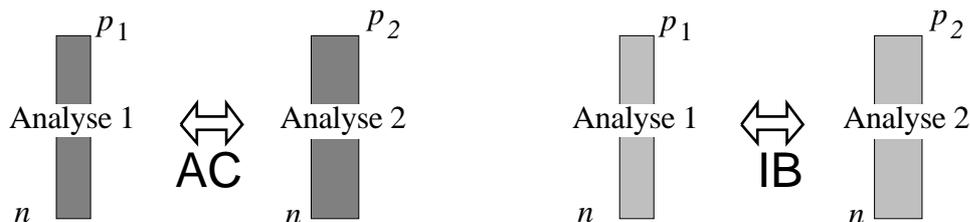
Au plan théorique les deux sont des cas particuliers de la théorie du schéma de dualité⁵. L'AC se définit entièrement par le schéma :

$$\mathbf{Y}^t \mathbf{D}_n \mathbf{X}, (\mathbf{X}^t \mathbf{D}_n \mathbf{X})^{-1}, (\mathbf{Y}^t \mathbf{D}_n \mathbf{Y})^{-1}$$

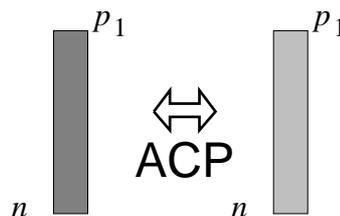
L'IB s'écrit simplement :

$$(\mathbf{Y}^t \mathbf{D}_n \mathbf{X}, \mathbf{I}_{p_1}, \mathbf{I}_{p_2})$$

Dans cette écriture \mathbf{X} et \mathbf{Y} sont des tableaux centrés et réduits d'ACP normée. Les deux sont des méthodes aux valeurs propres qui diagonalisent une matrice et c'est pourquoi les procédures sont voisines. La différence essentielle vient du point de vue. Chacun des tableaux contient des lignes (individus) et des colonnes (variables). On a le choix entre deux directions :



L'AC voit les deux tableaux comme deux ensembles de variables dans le même espace, l'IB voit les deux tableaux comme deux ensembles de n individus dans deux espaces. Il n'est possible de réunir les deux points de vue dans une même approche que dans le cas d'un seul tableau :



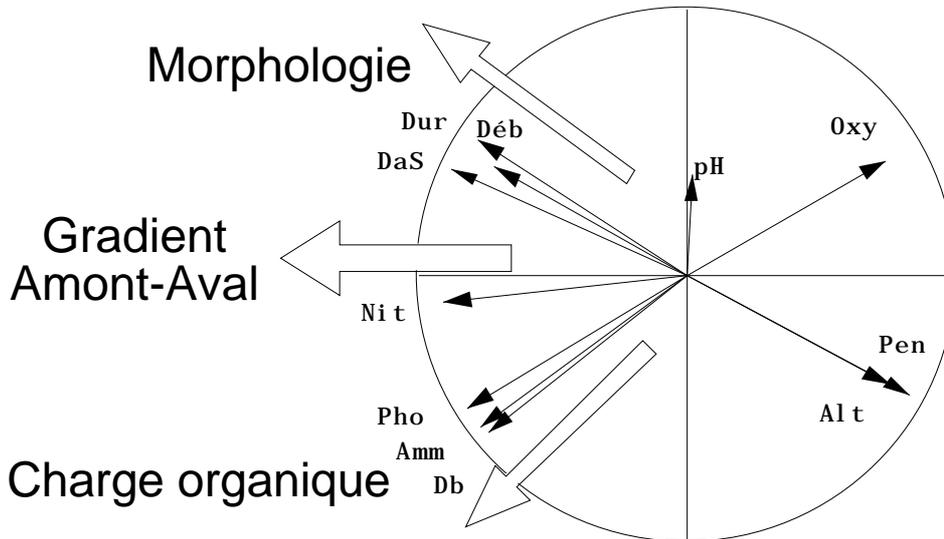
Quand il y a deux tableaux, on est obligé de choisir un point de vue de référence.

2.1 — Principe de l'analyse canonique

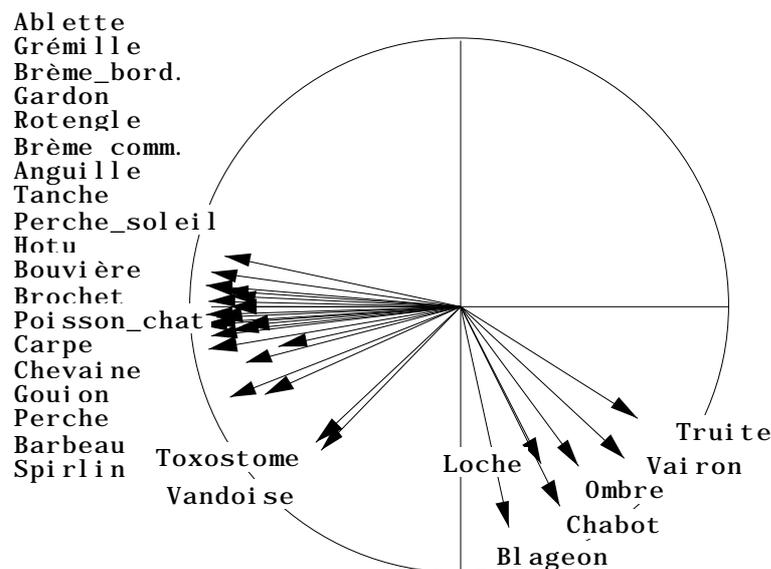
Quand on privilégie d'abord les variables, surtout sur des tableaux d'ACP, la corrélation joue le rôle principal. Prenons le tableau de milieu de la carte Doubs :



Le cercle des corrélations s'appuie sur la géométrie du nuage de variables :



Les variables, vecteurs de longueur unité, sont sur la sphère et se projette sur le plan des composantes principales dans le cercle unité. Les coordonnées sont des corrélations (cercle des corrélations). S'exprime l'évolution des variables morphologiques (Distance à la source, Débit et charge ionique augmentent quand la pente et l'altitude diminuent) et l'évolution des variables biologiques (Phosphates, ammoniacque et demande biologique augmentent pendant que l'oxygène diminue). Ces évolutions ont une part commune (gradient amont-aval) et une part spécifique qui sont bien représentées par cette projection euclidienne. Avec le tableau faunistique, on évite cependant cette opération qui donne ici :

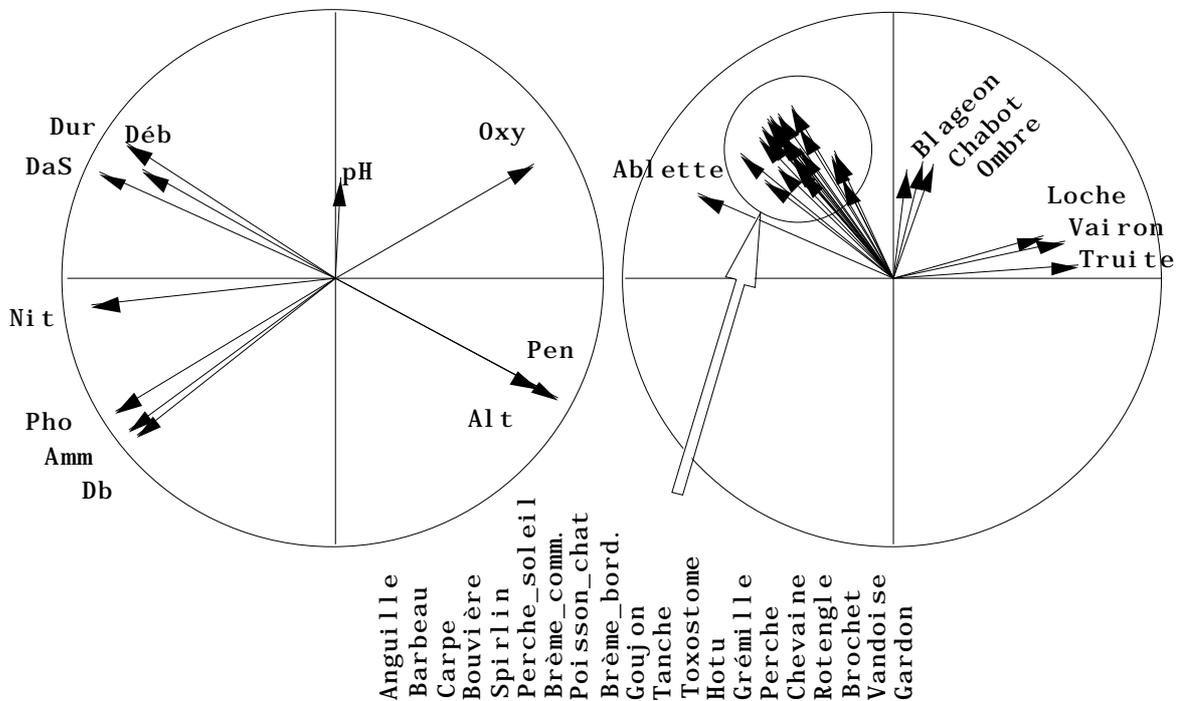
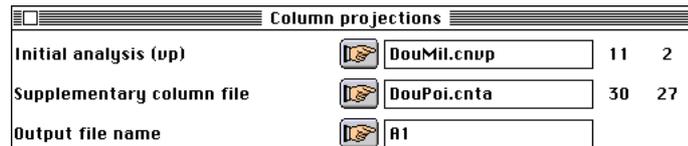


On n'utilise pas la normalisation des données faunistiques pour des raisons d'hétérogénéité des variances, elle est ici utilisée pour des raisons pédagogiques :

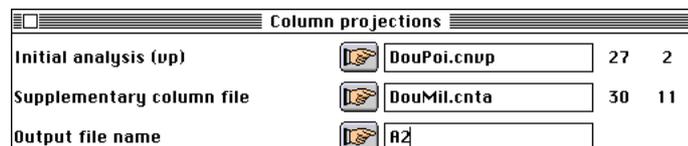


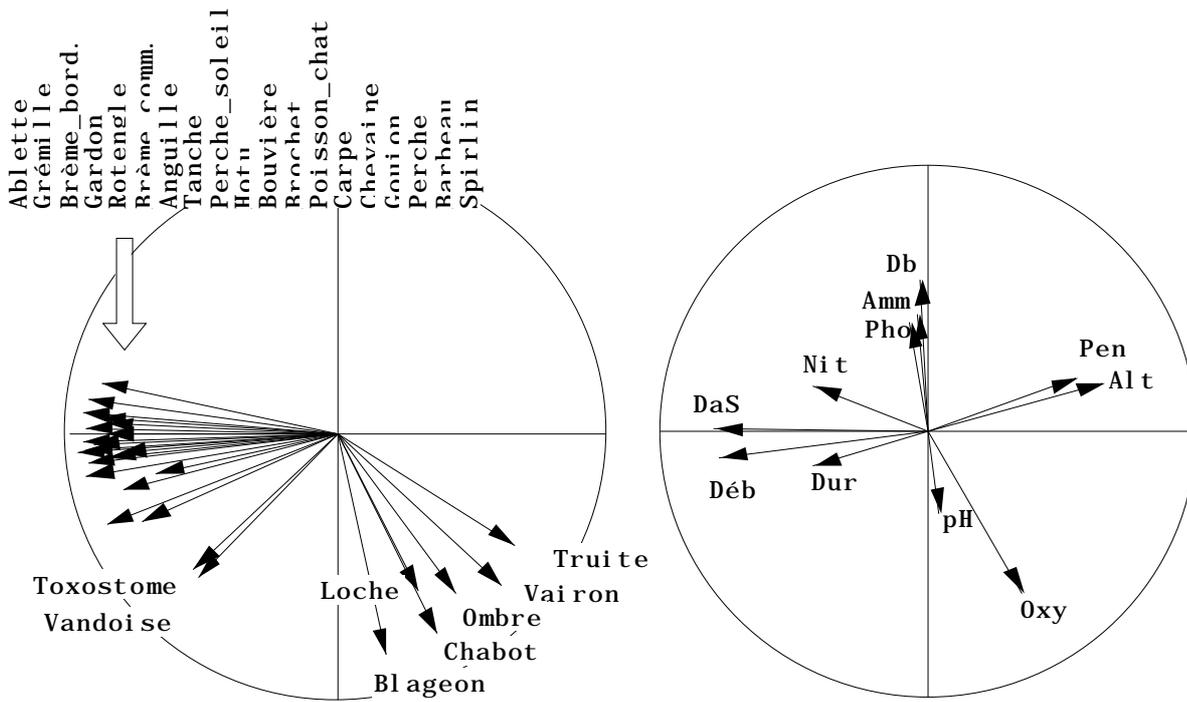
Il est clair que les deux ensembles de variables sont dans un même espace et peuvent donc être représentés ensemble. Par projection, ces vecteurs normés ont des projections sur des plans systématiquement inscrite dans des cercles. Quel plan choisir ?

On peut commencer par projeter le nuage des variables-espèces sur le plan des composantes milieu :

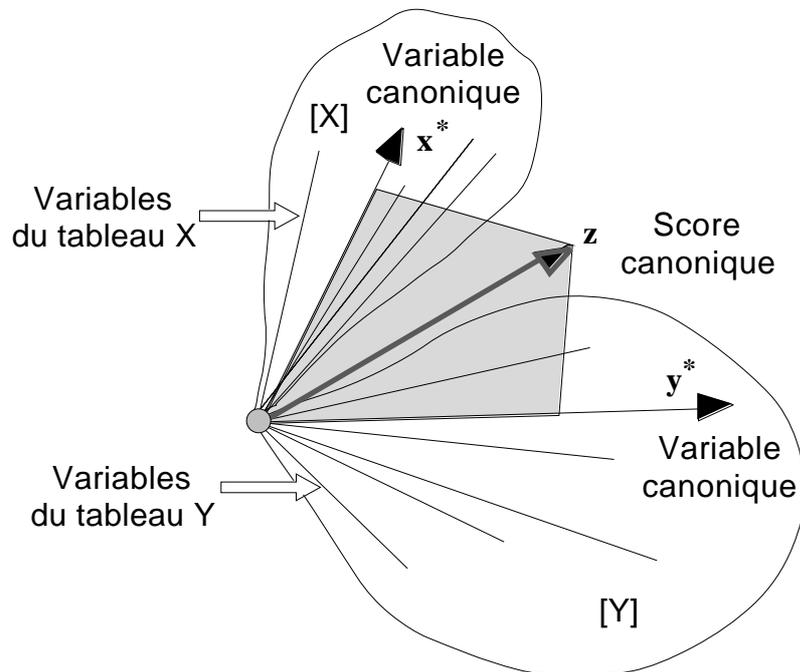


L'opération symétrique est possible. Elle consiste à projeter les variables-milieu sur le plan des composantes principales du tableau espèces :





Dans chacune des opérations, un groupe de variables est privilégié. Celui qui définit les composantes principales aura une meilleure représentation que l'autre. Le but de l'analyse canonique est de faire les deux simultanément :



L'analyse canonique résout plusieurs problèmes simultanément. Il suffit de comprendre qu'il est possible de calculer l'angle d'un vecteur avec un sous-espace. Les variables du tableau X engendrent un sous-espace $[X]$ alors que les variables du tableau Y engendrent un sous-espace $[Y]$. L'angle d'un vecteur z avec un sous-espace est l'angle minimum qu'on peut trouver en z et un vecteur quelconque du sous-espace. Notons $ang(z, [X])$ cette valeur. Alors le cosinus carré de cet angle est exactement le carré de la corrélation multiple de z avec les variables du tableau X :

$$\text{Cos}^2(\text{ang}(\mathbf{z}, [X])) = R_{\mathbf{z}; \mathbf{X}}^2$$

Le vecteur le plus proche simultanément de [X] et de [Y] définit la bissectrice entre les deux sous-espaces. Il maximise :

$$\text{Cos}^2(\text{ang}(\mathbf{z}, [X])) + \text{Cos}^2(\text{ang}(\mathbf{z}, [Y])) = R_{\mathbf{z}; \mathbf{X}}^2 + R_{\mathbf{z}; \mathbf{Y}}^2$$

On peut appeler \mathbf{z} le score canonique de rang 1 (il existe un vecteur, qui sous la contrainte d'être perpendiculaire à \mathbf{z} optimise la même quantité, on l'appellera le score canonique de rang 2, ...). Le vecteur unitaire \mathbf{x}^* de [X] qui définit l'angle minimum ainsi trouvé est porté par la projection orthogonale de \mathbf{z} sur le sous-espace engendré [X]. Le vecteur unitaire \mathbf{y}^* de [Y] qui définit l'angle minimum de l'autre côté est porté par la projection orthogonale de \mathbf{z} sur le sous-espace engendré [Y]. Les vecteurs unitaires \mathbf{x}^* et \mathbf{y}^* sont appelées variables canoniques.

Chaque score canonique est donc attaché à un couple de variables canoniques. Les vecteurs unitaires \mathbf{x}^* et \mathbf{y}^* définissent alors l'angle des deux espaces comme l'angle minimum que peuvent faire un angle d'un vecteur de [X] avec un angle d'un vecteur de [Y]. On notera la continuité entre (i) l'angle de deux variables défini par la corrélation :

$$\text{Cos}(\text{ang}(\mathbf{x}, \mathbf{y})) = \text{Cor}(\mathbf{x}, \mathbf{y})$$

(ii) l'angle d'une variable et d'un sous-espace défini par le carré de corrélation multiple:

$$\text{Cos}^2(\text{ang}(\mathbf{z}, [X])) = R_{\mathbf{z}; \mathbf{X}}^2 = \text{Max}_{\mathbf{x} [X]} [\text{Cos}^2(\text{ang}(\mathbf{z}, \mathbf{x}))]$$

et (iii) l'angle entre deux sous-espaces défini par :

$$\text{Cos}^2(\text{ang}([X], [Y])) = \text{Max}_{\mathbf{x} [X]} [\text{Cos}^2(\text{ang}(\mathbf{x}, [Y]))] = \text{Max}_{\mathbf{y} [Y]} [\text{Cos}^2(\text{ang}(\mathbf{y}, [X]))]$$

ou encore :

$$\text{Cos}^2(\text{ang}([X], [Y])) = \text{Max}_{\mathbf{x} [X], \mathbf{y} [Y]} [\text{Cos}^2(\text{ang}(\mathbf{x}, \mathbf{y}))]$$

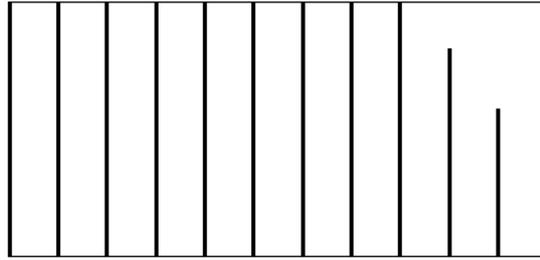
Cette dernière valeur est appelé carré de corrélation canonique et se calcule comme valeur propre de la matrice diagonalisée dans l'analyse canonique des deux tableaux.

2.2 — Pratique de l'analyse canonique

Elle est mise en œuvre dans le module Canonical :

| Canonical correlations | | | |
|------------------------|---|--------|-------|
| X : First data file |  | DouMil | 30 11 |
| Y : Second data file |  | DouPoi | 30 27 |
| Output file name |  | MilPoi | |

On s'aperçoit immédiatement que cette analyse est impossible :



Il existe 9 scores canoniques correspondant à un carré de corrélation canonique égal à 1. Les sous-espaces [X] et [Y] se chevauchent et l'intersection est un sous-espace de dimension 9. C'est exactement comme si dans \mathbb{R}^3 on cherchait l'angle de deux plans. On trouverait nécessairement 0 (1 pour le cosinus) car les deux plans ont une droite en commun.

La dimension des listes de variables-espèces exclue pratiquement des méthodes de couplage l'analyse canonique classique. Un nombre limité de variables dans chaque tableau s'impose. Retenons les espèces les plus abondantes :

SelectTaxa

Matrix input file 30 27

Option: labelling file

Output file name

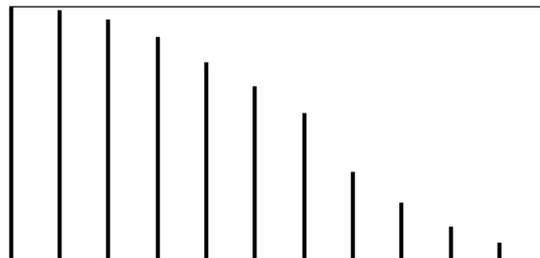
Minimal frequency

Canonical correlations

X : First data file 30 11

Y : Second data file 30 17

Output file name



Le graphe des valeurs propres n'est pas un graphe qui décompose l'inertie mais est formé des valeurs successives des carrés de corrélation canonique. Il est impossible de sélectionner un nombre pertinent de dimension mais le listing donne le test de dimensionalité explicité par Gittins :

| Num. | Eigenval. | R.Iner. | R.Sum | Num. | Eigenval. | R.Iner. | R.Sum |
|------|-------------|---------|---------|------|-------------|---------|---------|
| 01 | +9.9976E-01 | +0.1504 | +0.1504 | 02 | +9.8762E-01 | +0.1485 | +0.2989 |
| 03 | +9.4795E-01 | +0.1426 | +0.4415 | 04 | +8.8118E-01 | +0.1325 | +0.5740 |
| 05 | +7.7941E-01 | +0.1172 | +0.6912 | 06 | +6.8608E-01 | +0.1032 | +0.7944 |
| 07 | +5.8074E-01 | +0.0873 | +0.8817 | 08 | +3.4956E-01 | +0.0526 | +0.9343 |
| 09 | +2.3149E-01 | +0.0348 | +0.9691 | 10 | +1.3630E-01 | +0.0205 | +0.9896 |
| 11 | +6.8890E-02 | +0.0104 | +1.0000 | | | | |

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 = 3.2291e+02 ddl =187 proba = 2.8101e-09

```

k= 1 Khi2 = 2.1588e+02  ddl =160  proba = 2.1531e-03
k= 2 Khi2 = 1.5735e+02  ddl =135  proba = 9.1452e-02
k= 3 Khi2 = 1.1516e+02  ddl =112  proba = 3.9977e-01
k= 4 Khi2 = 8.2337e+01  ddl = 91   proba = 7.3047e-01
...

```

```

Canonical correlation coefficients
k= 1 rk = 9.9988e-01  rk2 = 9.9976e-01
k= 2 rk = 9.9379e-01  rk2 = 9.8762e-01
k= 3 rk = 9.7363e-01  rk2 = 9.4795e-01
k= 4 rk = 9.3871e-01  rk2 = 8.8118e-01
...

```

Squared canonical correlation coefficients are in the file MilPoiRed.canvp

Il existe une combinaison de variables de milieu et une combinaison de variables faunistiques qui présentent une corrélation de 0.9998. Un miracle ? Non, plus prosaïquement un sous produit (hautement significatif si on accepte les hypothèses gaussiennes, ce qui est exclu) de la géométrie des sous-espaces vectoriels.

File MilPoiRed.cav1 contains correlations between variables (X) and canonical scores

Rows = 11 Col = 2

File :MilPoiRed.cav1

| Col. | Mini | Maxi |
|------|------------|-----------|
| 1 | -9.052e-01 | 7.524e-01 |
| 2 | -7.749e-01 | 6.658e-01 |

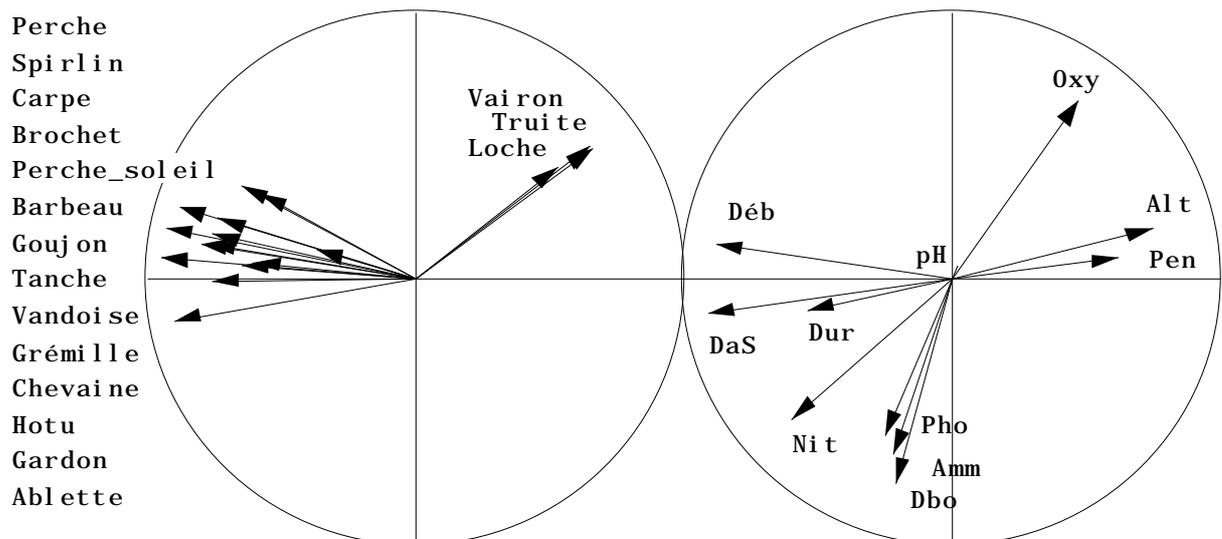
File MilPoiRed.cav2 contains correlations between variables (Y) and canonical scores

Rows = 17 Col = 2

File :MilPoiRed.cav2

| Col. | Mini | Maxi |
|------|------------|-----------|
| 1 | -9.450e-01 | 6.691e-01 |
| 2 | -1.611e-01 | 4.908e-01 |

Sur le plan défini par les scores canoniques on peut projeter les deux paquets de variables :



La figure est obtenue par ;

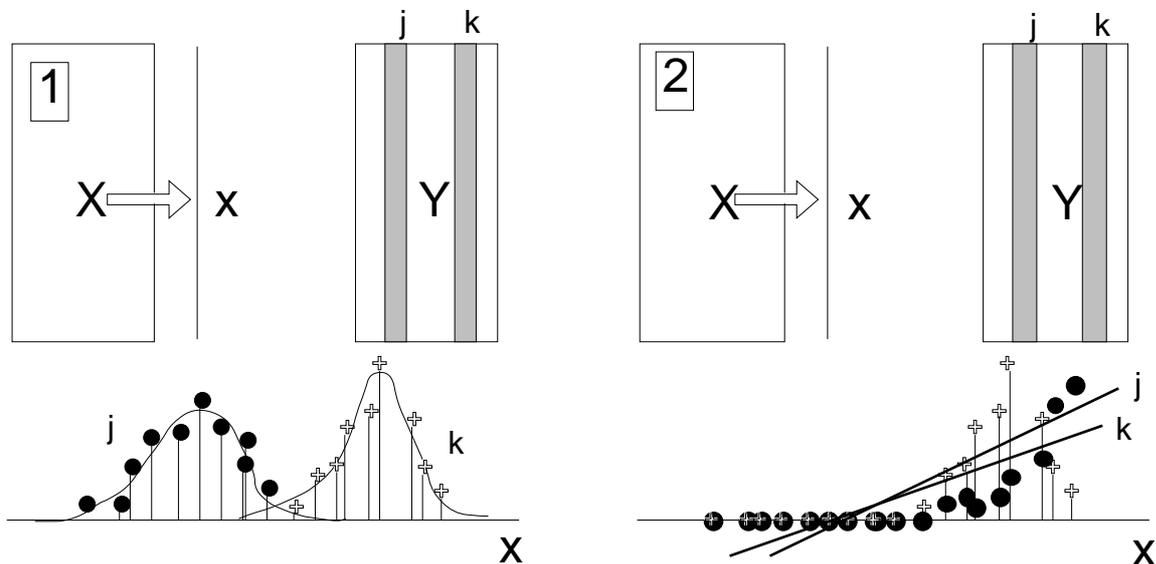
| Correlation circle | |
|------------------------------------|----------------|
| XY coordinates file | Mj PoiRed.cau1 |
| X-axis column number (default = 1) | |
| Y-axis column number (default = 2) | |
| Label file (or #) for items | UarMil_Label |

| Correlation circle | |
|------------------------------------|-----------------|
| XY coordinates file | Mj PoiRed.cau2 |
| X-axis column number (default = 1) | |
| Y-axis column number (default = 2) | |
| Label file (or #) for items | DouPoiRed_Label |

On récupère l'information qui indique le groupe Truite-Vairon-Loche en amont et le reste des espèces en aval. L'exclusion de 13 espèces sur 30 rend de toute manière cette analyse très contestable. Ce n'est pas une découverte puisque les commentaires de Gauch et Wentworth ⁶ étaient déjà sans appel (CCA désigne l'analyse canonique classique ici étudiée) :

Canonical correlation analysis seeks linear axes that reveal the joint structure of two matrices. Potentially, CCA could be a valuable technique for ordination and analysis of dual matrices of community and environmental measurements. Performance of CCA was tested with simulated and real vegetational data. CCA was found to have stringent requirements for linearity, and consequently to have little value for ordination. Indirect ordination of community data by reciprocal averaging, followed by interpretation of environmental relationships of the axes, should generally be more effective.

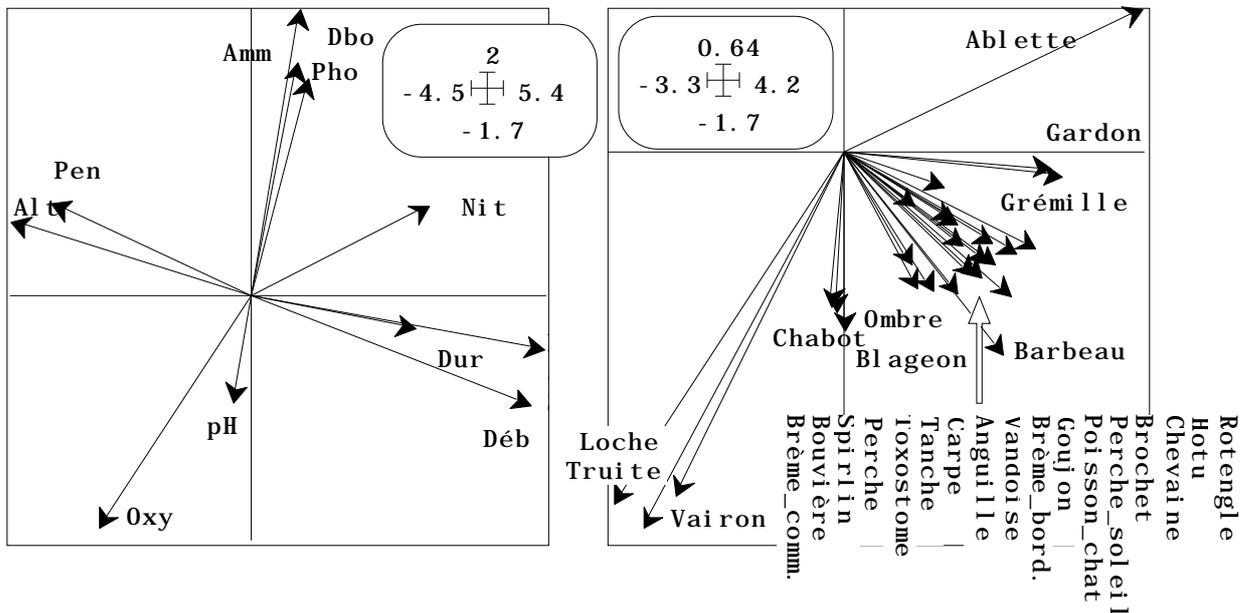
Dans ce sommaire, on remarque la présence de résultats expérimentaux —*Performance of CCA was tested*— et d'une raison donnée —*stringent requirements for linearity*. La réputation de l'analyse canonique est devenue rapidement après les premiers essais très défavorable et Gittins a été un des rares à défendre le concept. La raison globalement évoquée de l'échec est l'omniprésence des réponses curvilinéaires des espèces sur les gradients environnementaux. Austin dans ses synthèses qui font autorité ^{7 8} a fortement défendu cette position qui est maintenant acceptée par tous.



S'appuyer sur la corrélation (2) c'est mettre en avant une liaison linéaire avec une variable environnementale. Reconnaître l'omniprésence des courbes en cloche (1) oblige à remettre en cause toute analyse qui s'intéressera aux liaisons linéaires. La porte est ouverte à la confusion entre le lien linéaire d'une espèce et d'une variable et l'algèbre linéaire base des méthodes à vecteurs propres.

Dans l'exemple précédent, il y a très peu de courbes de réponse en cloche. Les espèces de Poissons sont essentiellement calées en amont (corrélation négative avec la distance à la source) ou en aval (corrélation positive avec la distance à la source). Le

second gradient est la pollution qui exerce une influence négative sur la plupart des taxons. L'analyse canonique est pourtant bien inférieure à l'analyse de co-inertie qui se résume à :



Une partie des problèmes de l'analyse canonique vient de l'usage de la corrélation. Ces problèmes sont partagés par l'analyse de co-inertie ci-dessus (ACP normée du milieu contre ACP centrée de la faune). Une autre partie, sans doute plus importante est liée aux problèmes de dimensions des sous-espaces. Ils sont partagés par toutes les méthodes qui évoque des régressions multiples, implicites ou explicites.

Il y a 11 variables de milieu pour 30 relevés, c'est beaucoup. Il y a 29 espèces, c'est beaucoup trop. Les sous-espaces se chevauchent, ou après une réduction drastique sont très proches. La dissymétrisation du problème s'impose pour des questions de dimensions.

Il y a des espèces qui occupent des positions intermédiaires sur le gradient. Leur perception par la corrélation est très défavorable. L'approche symétrique des variables mésologiques et faunistiques n'est pas une bonne chose. La dissymétrisation du problème s'impose pour des questions de modèles biologiques.

L'analyse canonique des correspondances va introduire deux dissymétries. On a rarement discutée de l'importance relative de chacune d'entre elles. L'analyse de co-inertie va en introduire une seule. On se demandera alors en quoi l'ACC est restée une analyse canonique puisque deux de ses propriétés ont été abandonnées.

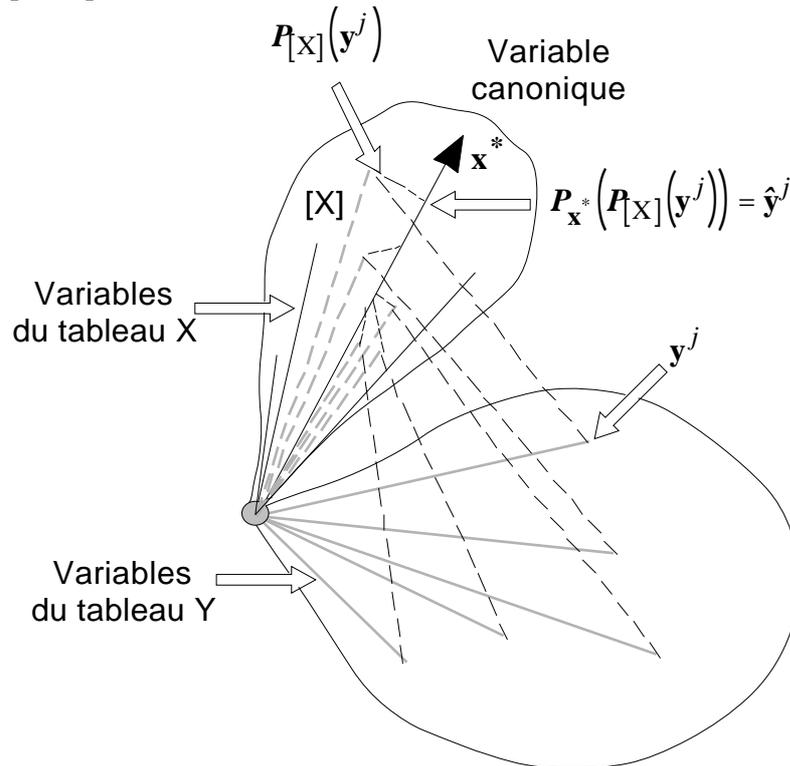
3 — Stratégies de couplages dissymétriques

Écologiquement le milieu influence l'abondance des taxons. Donc les variables mésologiques sont explicatives, les variables faunistiques sont dépendantes.

3.1 — Dissymétrie de prédiction

En ne changeant pas d'espace, les statisticiens ont modifié l'analyse canonique :

utilisant le principe suivant :



Chacune des variables du tableau \mathbf{Y} est projeté sur le sous-espace $[X]$. La nouvelle variable $P_{[X]}(y^j)$ est l'estimation de la variable y^j par régression multiple sur les variables de \mathbf{X} dites explicatives. L'ensemble des variables $P_{[X]}(y^j)$ définissent par ACP leur composante principale x^* . Les variables sont à nouveau projetées sur cette composante principale en

$$P_{x^*}(P_{[X]}(y^j)) = \hat{y}^j$$

On trouve ainsi une combinaison de variables de \mathbf{X} qui sert de prédicteur unique pour toutes les variables de \mathbf{Y} . Cette combinaison de régression sur variables instrumentales et d'analyse en composantes principales porte bien son nom d'ACP sur variables instrumentales⁹. On dit aussi analyse des redondances¹⁰. La proximité conceptuelle de ces approches géométriques dans l'espace des variables a généré des aides à l'interprétation¹¹ des polémiques¹² et des extensions¹³. Obadia a proposé le terme fort judicieux de composantes explicatives¹⁴. Ter Braak l'a introduite en écologie¹⁵.

On pratique cette procédure avec le module Projectors (voir fiche 3.5). On définit une base orthonormale du sous-espace $[X]$:



Chacune des variables fournit dans la projection un taux de variance expliquée :

| Triplet Inertia Decomposition | | | | | | |
|----------------------------------|------------|------------|-------------|------|------|----|
| Explanatory variables: .@ob file | | | DouMil.@ob | | 30 | 11 |
| Dependent variables: .**ta | | | DouPoi.cpta | | 30 | 27 |
| | Subspace A | A Orthogo | Total | A+ | A- | |
| 1 | 6.3789e-01 | 2.1211e-01 | 8.5000e-01 | 7504 | 2495 | |
| 2 | 3.1389e+00 | 8.8443e-01 | 4.0233e+00 | 7801 | 2198 | |
| ... | | | | | | |
| 25 | 2.6473e+00 | 1.8494e-01 | 2.8322e+00 | 9347 | 652 | |
| 26 | 4.4110e+00 | 4.7896e-01 | 4.8900e+00 | 9020 | 979 | |
| 27 | 1.9536e+00 | 6.9691e-02 | 2.0233e+00 | 9655 | 344 | |
| Tot | 5.0258e+01 | 1.5820e+01 | 6.6078e+01 | 7605 | 2394 | |

Au total 76% de la variabilité faunistique est prise en compte dans ces régressions séparées. Cette quantité se teste par un test de Monte-Carlo :

| Subspace Test | | | |
|----------------------------------|-------------|----|----|
| Explanatory variables: .@ob file | DouMil.@ob | 30 | 11 |
| Dependent variables: .**ta | DouPoi.cpta | 30 | 27 |
| Select a number of permutations | 1000 | | |

number of random permutations: 1000 Observed: 7.606e-01
Histogram: minimum = 1.872e-01, maximum = 7.606e-01
number of simulations X<Obs: 1000 (frequency: 1.000e+00)
number of simulations X>=Obs: 0 (frequency: 0.000e+00)

```

**
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
*****
**
**

```

•-->

La synthèse est légitime :

| PCA on Instrumental Variables | | | |
|----------------------------------|-------------|----|----|
| Explanatory variables: .@ob file | DouMil.@ob | 30 | 11 |
| Dependent variables: .**ta | DouPoi.cpta | 30 | 27 |
| Output file name | Z | | |

```

Instrumental variables
----- input -----
Orthonormal basis: DouMil.@ob
It has 30 rows and 11 columns
Dependent variable file: DouPoi.cpta
It has 30 rows and 27 columns
----- output -----
Projected variable file: Z.ivta

```

It has 30 rows and 27 columns
 Inertia: 5.0258e+01
 File Z.ivpc contains the column weight
 It has 27 rows and 1 column
 File Z.ivpl contains the row weight
 It has 30 rows and 1 column

| Num. | Eigenval. | R.Iner. | R.Sum | Num. | Eigenval. | R.Iner. | R.Sum |
|------|-------------|---------|---------|------|-------------|---------|---------|
| 01 | +3.8418E+01 | +0.7644 | +0.7644 | 02 | +5.9540E+00 | +0.1185 | +0.8829 |
| 03 | +2.4162E+00 | +0.0481 | +0.9310 | 04 | +1.3387E+00 | +0.0266 | +0.9576 |
| 05 | +7.4306E-01 | +0.0148 | +0.9724 | 06 | +5.9064E-01 | +0.0118 | +0.9841 |
| 07 | +3.5800E-01 | +0.0071 | +0.9912 | 08 | +1.7964E-01 | +0.0036 | +0.9948 |
| 09 | +1.2978E-01 | +0.0026 | +0.9974 | 10 | +8.0960E-02 | +0.0016 | +0.9990 |
| 11 | +4.9398E-02 | +0.0010 | +1.0000 | | | | |

Deux composantes explicatives prennent en compte 88% de la variance expliquée, elle-même représentant 76 % de la variance de départ.

| Num | Variance | R2 | EigenVal. |
|-----|----------|--------|-----------|
| 1 | 42.59 | 0.9019 | 38.42 |
| 2 | 7.76 | 0.7673 | 5.954 |

La valeur propre est une variance expliquée, produit d'une variance (propriété d'ACP) et d'un taux d'explication (propriété de régression multiple). Les questions de dimensionalité ont été résolus.

```

files Z.ivfa
      Z.ivll
      Z.ivco
allow a convenient interpretation
-----

```

On pourra interpréter graphiquement en enchaînant le poids des variables explicatives (*loadings*) constituant la variable canonique (combinaison linéaire de variance unité) sur laquelle sont projetées les variables expliquées (espèces). La projection des variables des deux blocs est encore possible sur les variables canoniques (après normalisation si on veut ne voir que les angles).

Diagonal Inner product C=H'DY

Input file for H matrix:

Option for H matrix (default=none):

Input file for Y matrix:

Option for Y matrix (default=none):

D inner product (default = 1/n):

Option: weight file:

Output file (default = Screen):

Correlation circle

H-Y coordinates file:

H-axis column number (default = 1):

Y-axis column number (default = 2):

Label file (or #) for items:

Diagonal Inner product C=H'DY

Input file for H matrix:

Option for H matrix (default=none):

Input file for Y matrix:

Option for Y matrix (default=none):

D inner product (default = 1/n):

Option: weight file:

Output file (default = Screen):

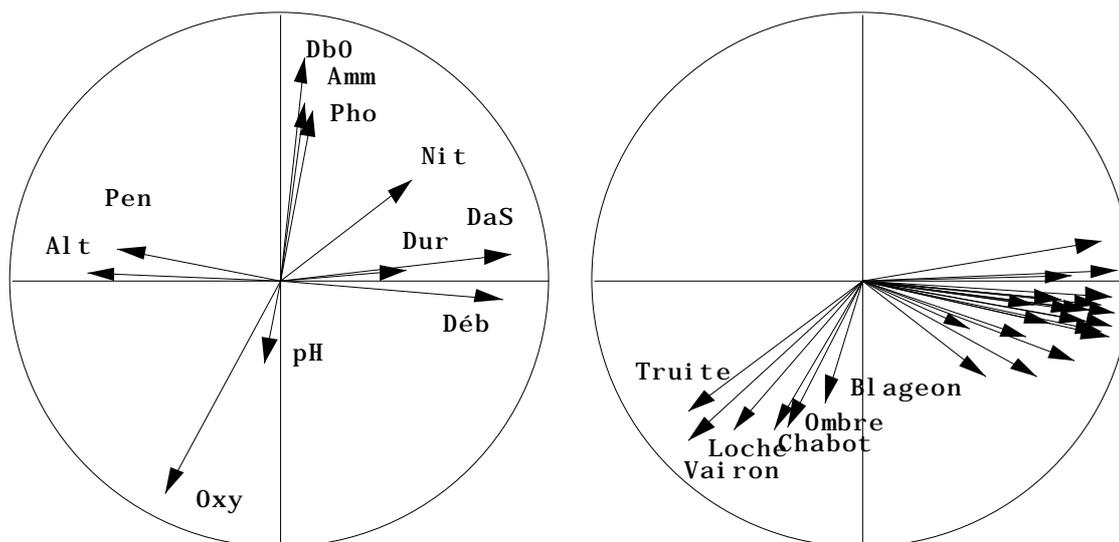
Correlation circle

H-Y coordinates file:

H-axis column number (default = 1):

Y-axis column number (default = 2):

Label file (or #) for items:



La difficulté liées aux courbes de réponse n'est pas résolue.

3.2 — Dissymétrie de centrage

Une seconde modification est possible. Au lieu de considérer que les variables de [Y] sont des vecteurs colonnes obtenues par centrage ou normalisation (le tableau Y est celui d'une ACP centrée ou d'une ACP normée) on considère qu'elles dérivent d'une AFC. Ceci oblige en outre à changer de métrique et à passer de la métrique des poids uniformes dans une ACPVI à la métrique des poids issus de l'AFC dans une AFCVI, nom donnée à l'analyse canonique des correspondances dans ¹⁶, étudiée dans ¹⁷ et explicitée en détail dans ¹⁸. Initialement Ter Braak n'a pas ce point de vue ¹⁹ :

Originally, CCA was derived as an approximation to maximum likelihood Gaussian ordination with linear external constraints (ter Braak 1986²⁰, 1988²¹).

Fondamentalement, Ter Braak est parti de l'AFC vue comme procédure convergente dans laquelle peuvent être introduites des opérations supplémentaires comme la régression ou le *detrending* ou les deux (*detrended CCA*). Il utilise en introduction l'article de Gauch & Wentworth déjà cité comme condamnation de l'analyse canonique des corrélations et la modification de la procédure de l'analyse des correspondances de Hill et Gauch ²² qui l'éloigne de la notion d'analyse canonique reconnue par Gittins. On comprend difficilement le choix du terme analyse canonique des correspondances puisque la méthode est d'abord une alternative à l'analyse canonique classique et un enrichissement de l'analyse des correspondances détendancée qui n'est plus une analyse canonique. Comme AFCVI, la CCA est une analyse dissymétrique doublement, donc ne peut s'admettre comme analyse canonique. Ce qui ne change rien à son utilité, qui est incontestable.

L'exécution dans ADE-4 rend transparente le point de vue dissymétrique. L'AFC du tableau faunistique est lancée :



Son interprétation n'est pas indispensable, mais son exécution l'est, puisqu'elle donne à chaque relevé un poids proportionnel à l'abondance des Poissons qu'on y trouve. Le poids du relevé 8 est nul :

| DouPoi.fcpl | |
|-------------|--------|
| 6 | 0.0209 |
| 7 | 0.0159 |
| 8 | 0.0000 |
| 9 | 0.0139 |
| 10 | 0.0139 |
| 11 | 0.0110 |
| 12 | 0.0179 |
| 13 | 0.0189 |
| 14 | 0.0270 |

L'ACP normée du tableau de milieu est entreprise avec cette pondération :

| Correlation matrix PCA | |
|-----------------------------|------------------|
| Matrix input file | DouMil 30 11 |
| Row weights (default=1/n) | 3 |
| Column weights (default=1) | |
| Option: file for row weight | DouPoi.fcpl 30 1 |

L'interprétation n'est pas indispensable puisque les variables sont destinées à définir un sous-espace de projections :

| Triplet->Orthonormal Basis | |
|----------------------------|-------------------|
| Explanatory variables | DouPoi.cnta 30 27 |
| Option: output file name | |

```
Orthonormal basis: DouMil.@ob
It has 30 rows and 11 columns
Row weight file: DouMil.@pl
(the same as DouMil.cnpl)
Coordinates of the vectors of the orthonormal basis
in the basis of columns of DouMil.cnta in : DouMil.@co
File DouMil.@co has 11 rows and 11 columns
```

L'inertie du tableau faunistique est décomposée :

| Triplet Inertia Decomposition | |
|----------------------------------|-------------------|
| Explanatory variables: .@ob file | DouMil.@ob 30 11 |
| Dependent variables: .**ta | DouPoi.fcta 30 27 |

Projected inertia on a subspace

```
-----
Orthonormal basis: DouMil.@ob
It has 30 rows and 11 columns
Dependent variable file: DouPoi.fcta
It has 30 rows and 27 columns
-----
```

| | Subspace A | A Orthogo | Total | A+ | A- |
|-----|------------|------------|------------|------|------|
| 1 | 5.2686e-02 | 1.6664e-02 | 6.9350e-02 | 7597 | 2402 |
| 2 | 1.5411e-01 | 4.9463e-02 | 2.0357e-01 | 7570 | 2429 |
| ... | | | | | |
| 23 | 2.5610e-02 | 3.2852e-03 | 2.8895e-02 | 8863 | 1136 |
| 24 | 2.1951e-02 | 1.3382e-02 | 3.5333e-02 | 6212 | 3787 |
| 25 | 2.2318e-02 | 1.5199e-03 | 2.3838e-02 | 9362 | 637 |
| 26 | 5.5846e-02 | 9.6113e-03 | 6.5458e-02 | 8531 | 1468 |
| 27 | 1.6130e-02 | 1.0157e-03 | 1.7146e-02 | 9407 | 592 |
| Tot | 8.3687e-01 | 3.3005e-01 | 1.1669e+00 | 7171 | 2828 |

File Provi.<>pc contains the weight of each column
 It has 11 rows and 1 column

File Provi.<>ma contains the name of the input statistical triplets
 1 ---> DouMil.cnta
 2 ---> DouPoi.fcta

total inertia: 2.590287



Le résultat est encore plus significatif :

Warning: non uniform weight. The results from permutations
 are valid only if the row weights come from the fixed table.

 The fixed table is table 2: DouPoi.fcta
 number of random permutations: 1000 Observed: 2.590e+00
 Histogram: minimum = 1.377e-01, maximum = 2.590e+00
 number of simulations X<Obs: 1000 (frequency: 1.000e+00)
 number of simulations X>=Obs: 0 (frequency: 0.000e+00)

```

*****
*****
*****
*****
*****
**
*

```

•->

L'analyse elle-même est exécutée :



Instrumental variables
 ----- input -----
 Orthonormal basis: DouMil.@ob
 It has 30 rows and 11 columns
 Dependent variable file: DouPoi.fcta
 It has 30 rows and 27 columns
 ----- output -----
 Projected variable file: Y.ivta
 It has 30 rows and 27 columns
 Inertia: 8.3687e-01
 File Y.ivpc contains the column weight
 It has 27 rows and 1 column
 File Y.ivpl contains the row weight

It has 30 rows and 1 column

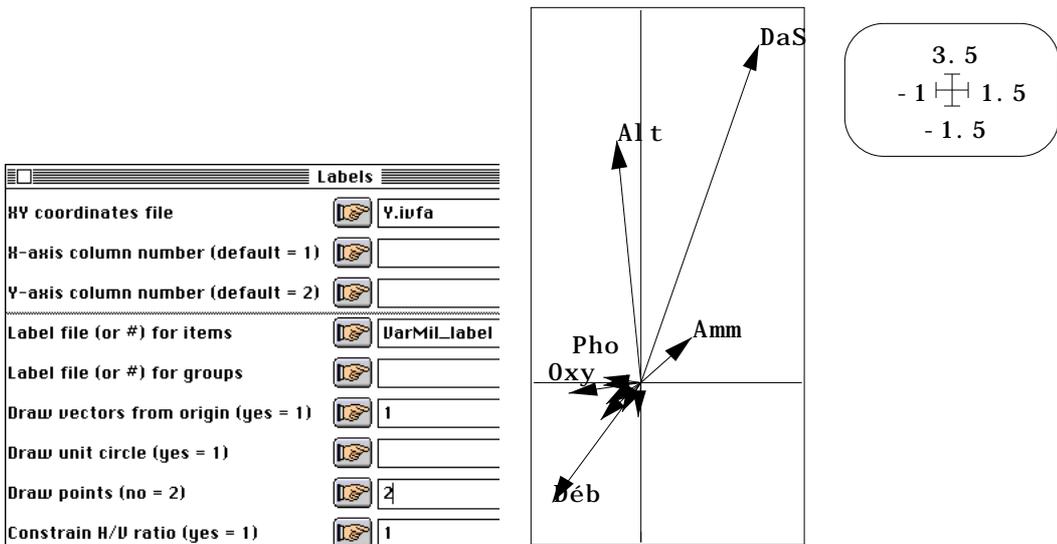
| Num. | Eigenval. | R.Iner. | R.Sum |
|------|-------------|---------|---------|
| 01 | +5.3452E-01 | +0.6387 | +0.6387 |
| 02 | +1.2184E-01 | +0.1456 | +0.7843 |
| 03 | +6.8703E-02 | +0.0821 | +0.8664 |
| 04 | +4.9168E-02 | +0.0588 | +0.9252 |
| 05 | +2.7090E-02 | +0.0324 | +0.9575 |
| 06 | +1.2941E-02 | +0.0155 | +0.9730 |
| 07 | +9.8670E-03 | +0.0118 | +0.9848 |
| 08 | +5.4252E-03 | +0.0065 | +0.9913 |
| 09 | +3.5336E-03 | +0.0042 | +0.9955 |
| 10 | +2.1655E-03 | +0.0026 | +0.9981 |
| 11 | +1.6117E-03 | +0.0019 | +1.0000 |

```

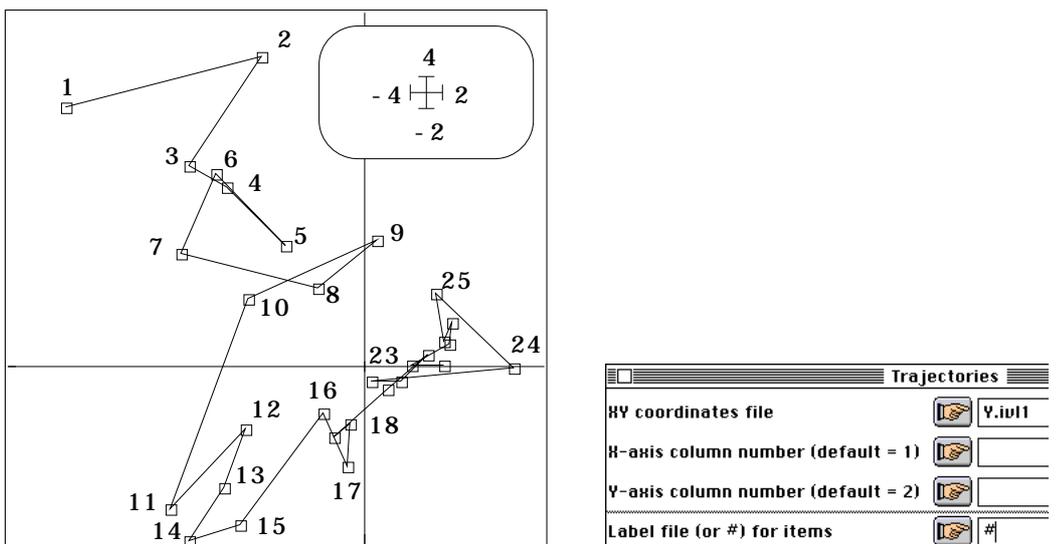
files Y.ivfa
      Y.ivl1
      Y.ivco
allow a convenient interpretation
-----

```

Un poids est donné à chacune des variables :

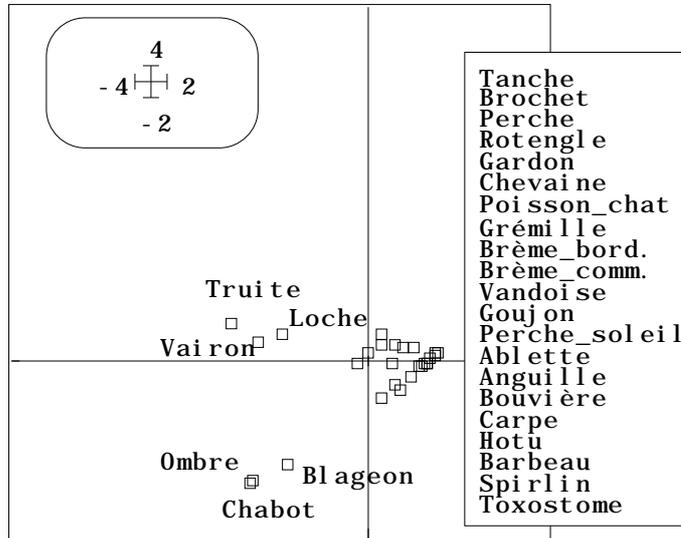


Deux variables canoniques normalisées sont calculées :



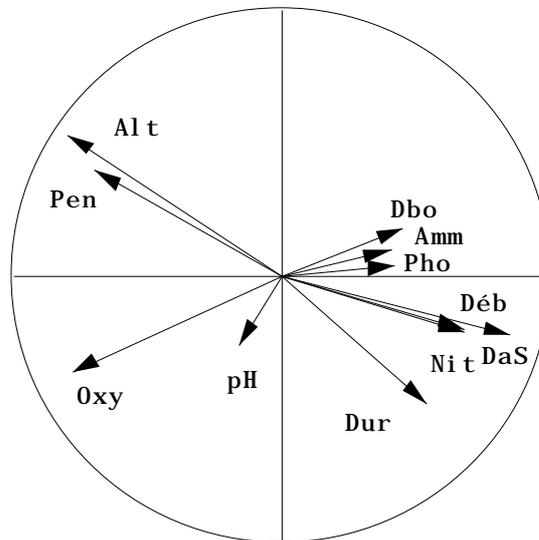
La variance des positions moyennes est alors optimales (séparation des niches) :

| Labels | |
|------------------------------------|-----------|
| HV coordinates file | Y.ivco |
| H-axis column number (default = 1) | |
| Y-axis column number (default = 2) | |
| Label file (or #) for items | Poi_Label |



Nulle trace du rôle de la pollution n'apparaît sur ce graphe, ce qui est une conséquence logique du choix de l'analyse. La séparation des zones à Truite, à Ombre et à Barbeau est par contre optimale. Seule question épineuse : faut-il interpréter par les poids canoniques ou les corrélations variables/variables canoniques ? Cette question a rarement une réponse claire :

| Diagonal Inner product C=H'DY | |
|------------------------------------|-------------|
| Input file for H matrix | DouMil.cnta |
| Option for H matrix (default=none) | |
| Input file for Y matrix | Y.iv11 |
| Option for Y matrix (default=none) | |
| D inner product (default = 1/n) | 2 |
| Option: weight file | DouPoi.fcpl |
| Output file (default = Screen) | R3 |



Doublement dissymétrique, contenant une régression et un centrage différent pour la faune et le milieu, en quoi l'ACC est-elle une analyse canonique ?

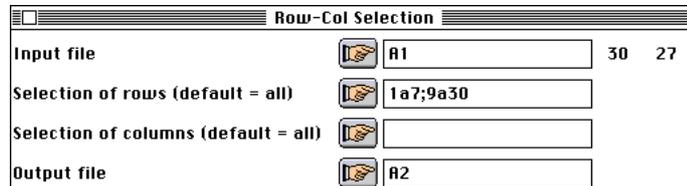
4 — Analyse canonique et liste d'occurrences

4.1 — Présentation des données

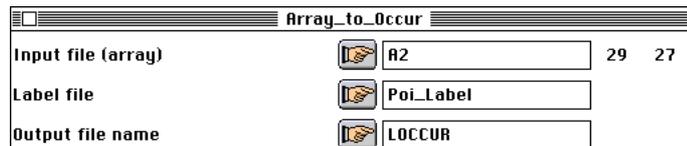
La question est abordée d'abord par le biais de la pratique. Pour simplifier la discussion le tableau faunistique est transformé en présence-absence :



Le relevé 8 qui est vide est exclu :



Le résultat est transformé en deux listes (OccurData: Array_to_Occur) :



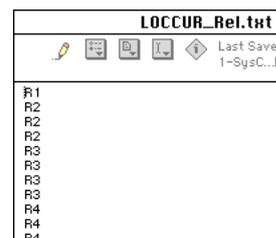
```
Input file :A2
Rows-samples: 29   Col-species: 27
Label input file :Poi_Label
Rows-labels: 27
```

On trouvera en première lecture le résultat bizarre. En effet toutes les cases du tableau sont énumérée et, chacune d'entre elles étant l'occurrence d'une espèce dans un relevé, on écrit dans un fichier le numéro de l'espèce et dans l'autre le numéro du relevé. On se retrouve avec deux variables qualitatives :

```
Label output file (species names):LOCCUR_Spe.txt
Rows-occurrences: 375
```



Dans un fichier texte, chaque ligne est une occurrence et on écrit le nom de l'espèce associée. On fait la même chose avec le nom du relevé ;



Les deux fichiers sont lus comme des variables qualitatives :

```
Label output file (sample names):LOCCUR_Rel.txt
Rows-occurrences: 375
```

One categorical variable file: LOCCUR_Spe
Number of rows: 375, variable: 1, categories: 27

Description of categories:

| Num | Label | Occur | F*10000 |
|-----|--------------|-------|---------|
| 1 | Truite | 17 | 453 |
| 2 | Vairon | 20 | 533 |
| 3 | Loche | 24 | 640 |
| 4 | Brochet | 18 | 480 |
| 5 | Chevaine | 25 | 667 |
| 6 | Goujon | 20 | 533 |
| ... | | | |
| 21 | Bouvière | 11 | 293 |
| 22 | Ablette | 14 | 373 |
| 23 | Anguille | 11 | 293 |
| 24 | Grémille | 12 | 320 |
| 25 | Brème_comm. | 9 | 240 |
| 26 | Brème_bord. | 10 | 267 |
| 27 | Poisson_chat | 7 | 187 |

Auxiliary ASCII output file LOCCUR_Spe.123: labels for 27 modalities

One categorical variable file: LOCCUR_Rel
Number of rows: 375, variable: 1, categories: 29

Description of categories:

| Num | Label | Occur | F*10000 |
|-----|-------|-------|---------|
| 1 | R1 | 1 | 27 |
| 2 | R2 | 3 | 80 |
| 3 | R3 | 4 | 107 |
| 4 | R4 | 8 | 213 |
| ... | | | |
| 26 | R26 | 22 | 587 |
| 27 | R27 | 22 | 587 |
| 28 | R28 | 26 | 693 |
| 29 | R29 | 21 | 560 |

Auxiliary ASCII output file LOCCUR_Rel.123: labels for 29 modalities

La fréquence de chaque espèce et son poids, la richesse de chaque relevé et son poids sont édités. Pour achever cette curieuse opération chaque variable qualitative est transformée en tableau disjonctif complet :



File LOCCUR_Rel01 contains dummy coded variable file
It has 375 rows and 29 columns (categories)

File LOCCUR_Rel01.num contains the number of modalities for each variable
It has 1 rows (variables) and 1 column



File LOCCUR_Spe01 contains dummy coded variable file
It has 375 rows and 27 columns (categories)

File LOCCUR_Spe01.num contains the number of modalities for each variable
It has 1 rows (variables) and 1 column

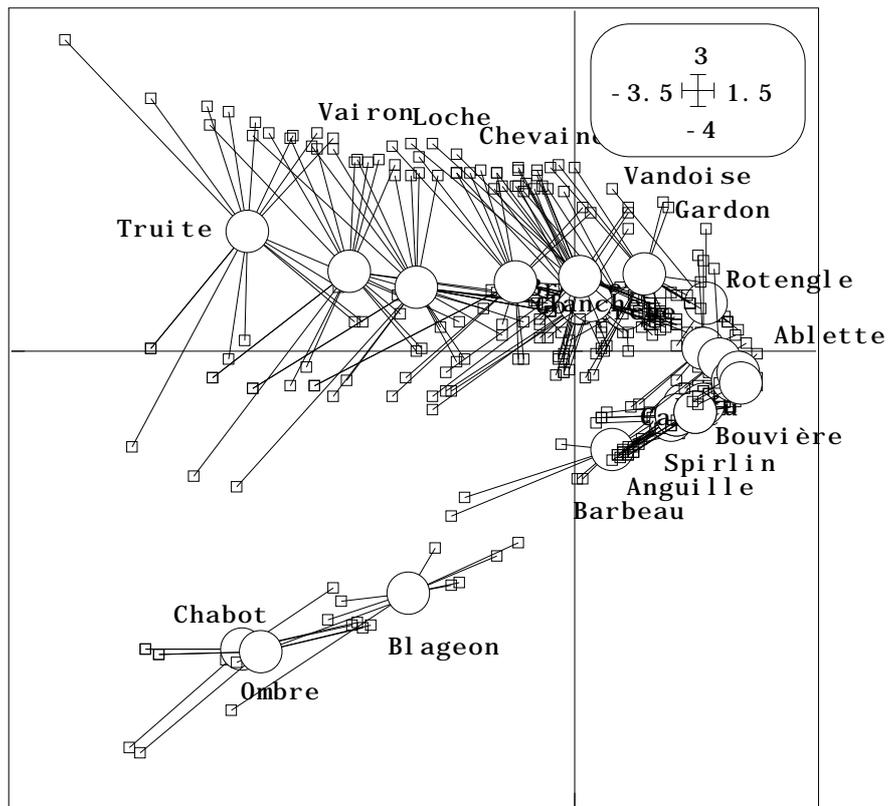
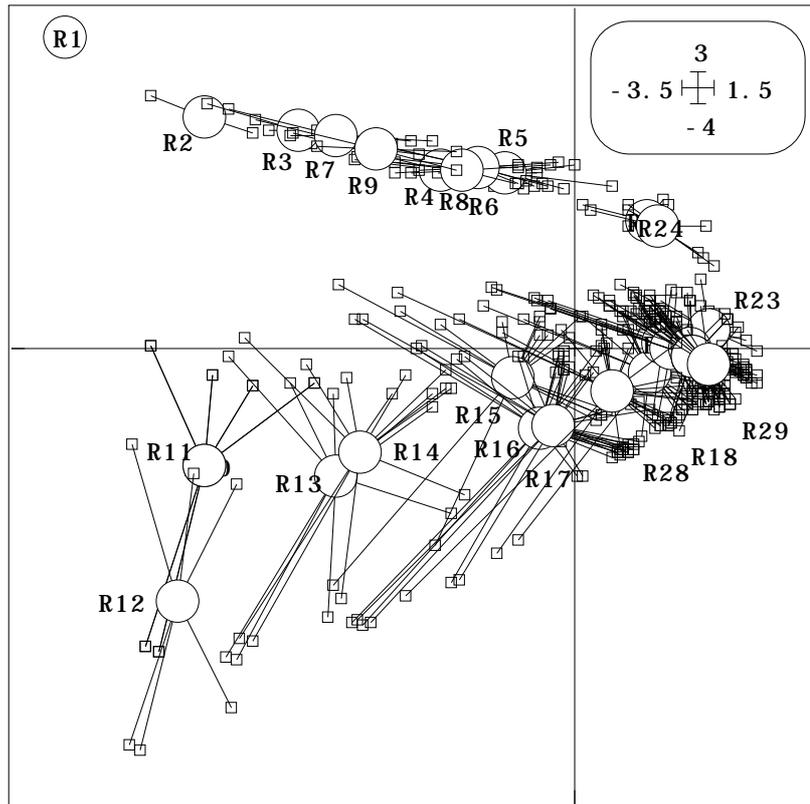
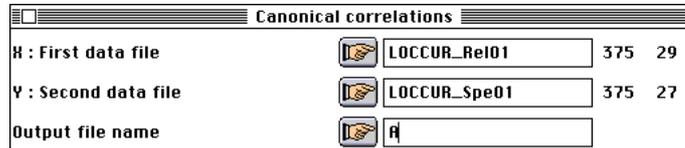


Figure 1 : *Reciprocal scaling* diversité-amplitude en AFC comme analyse canonique.

Nous voilà donc avec deux tableaux 375-29 et 375-27 au lieu d'un tableau 29-27. Il faut bien avoir une raison pour faire une chose aussi bizarre !

4.2 — L'AFC est une analyse canonique

Cette raison réside dans l'analyse canonique de ces deux tableaux :



First input file (X): LOCCUR_Spe01
Rows: 375 Columns: 27

Second input file (Y): LOCCUR_Rel01
Rows: 375 Columns: 29

Generic output file name: A

| Num. | Eigenval. | R.Iner. | R.Sum | Num. | Eigenval. | R.Iner. | R.Sum |
|------|-------------|---------|---------|------|-------------|---------|---------|
| 01 | +4.0887E-01 | +0.4304 | +0.4304 | 02 | +1.4459E-01 | +0.1522 | +0.5825 |
| 03 | +9.1223E-02 | +0.0960 | +0.6786 | 04 | +6.4034E-02 | +0.0674 | +0.7460 |
| 05 | +4.7416E-02 | +0.0499 | +0.7959 | 06 | +3.7991E-02 | +0.0400 | +0.8359 |
| 07 | +2.9445E-02 | +0.0310 | +0.8668 | 08 | +2.8554E-02 | +0.0301 | +0.8969 |
| 09 | +2.4527E-02 | +0.0258 | +0.9227 | 10 | +1.8208E-02 | +0.0192 | +0.9419 |
| 11 | +1.3691E-02 | +0.0144 | +0.9563 | 12 | +1.0692E-02 | +0.0113 | +0.9675 |
| 13 | +7.1907E-03 | +0.0076 | +0.9751 | 14 | +6.4162E-03 | +0.0068 | +0.9819 |
| 15 | +3.9444E-03 | +0.0042 | +0.9860 | 16 | +3.7951E-03 | +0.0040 | +0.9900 |
| 17 | +3.0736E-03 | +0.0032 | +0.9932 | 18 | +1.9007E-03 | +0.0020 | +0.9952 |
| 19 | +1.5744E-03 | +0.0017 | +0.9969 | 20 | +1.3062E-03 | +0.0014 | +0.9983 |
| 21 | +9.6483E-04 | +0.0010 | +0.9993 | 22 | +6.6953E-04 | +0.0007 | +1.0000 |

Vérifier que ce sont exactement les valeurs propres de l'AFC du tableau en présence-absence. L'analyse des correspondances est basée sur l'analyse canonique des deux tableaux disjonctifs complets. Vérifier enfin que le fichier A2_score issu de COA: Reciprocal scaling est exactement le même que celui des scores canoniques A.can11 de l'analyse canonique.

On est passé d'une vision des espèces et des relevés comme objets à une vision des occurrences comme données de base (Figure 1). La manipulation directe des occurrences qui peut paraître un exercice d'école devient incontournable dès que les données sont effectivement des listes d'occurrence.

Il est donc vraisemblable que cette propriété s'étend en ACC. L'exercice qui consiste à écrire effectivement les tableaux disjonctifs complet est évidemment pédagogique. En partant de deux variables qualitatives sous forme de fichiers de variables qualitatives sont directement obtenus par l'option OccurData: CANOCO-Classes :



4.3 — L'ACC est une analyse discriminante

En face de la liste des occurrences d'espèces, on peut mettre la liste des conditions de milieu. Ici chaque occurrence est codée comme le relevé dans laquelle on la trouve. Utiliser FilesUtil: Row_Duplicate :

| Row_Duplicate | |
|-----------------------------|----------------|
| Input file | Mil 29 11 |
| .cat file | LOCCUR_Rel.cat |
| Column number (default = 1) | |
| Output file | LMIL |

On se retrouve avec un fichier 375 occurrences - 11 variables de milieu.

| Canonical correlations | |
|------------------------|---------------------|
| X : First data file | LMIL 375 11 |
| Y : Second data file | LOCCUR_Spe01 375 27 |
| Output file name | c |

Faire en parallèle l'ACC du couple DouMil (30-11) et DouPoi01 (30-27)

| CCA | |
|----------------------------------|---------------------|
| Explanatory variables: .@ob file | DouMil.@ob 30 11 |
| Dependant variables: .**ta | DouPoi01.fcta 30 27 |
| Output file name | CCA |

et l'analyse discriminante de LMIL sur LOCCUR_Spe :

| Correlation matrix PCA | |
|------------------------|-------------|
| Matrix input file | LMIL 375 11 |

Observer, au début des deux procédures que la matrice des corrélations de DouMil calculée avec la pondération de DouPoi.fcpl est rigoureusement la même que celle de LMIL calculée avec la pondération uniforme :

```
File DouMil.cn+r contains the Correlation matrix
from statistical triplet DouMil.cnta
It has 11 rows and 11 columns
----- Correlation matrix -----
[ 1] 1000
[ 2] -911 1000
[ 3] -726 736 1000
[ 4] 930 -788 -697 1000
[ 5] -28 6 -242 -44 1000
[ 6] 745 -712 -623 805 21 1000
[ 7] 363 -325 -375 279 -81 308 1000
[ 8] 596 -678 -561 383 -11 385 691 1000
[ 9] 267 -273 -299 152 -122 211 929 751 1000
[10] -565 384 380 -407 203 -337 -715 -588 -658 1000
[11] 316 -241 -239 187 -201 235 869 547 831 -838 1000
-----

File LMIL.cn+r contains the Correlation matrix
from statistical triplet LMIL.cnta
It has 11 rows and 11 columns
----- Correlation matrix -----
[ 1] 1000
[ 2] -911 1000
[ 3] -726 736 1000
[ 4] 930 -788 -697 1000
[ 5] -28 6 -242 -44 1000
[ 6] 745 -712 -623 805 21 1000
[ 7] 363 -325 -375 279 -81 308 1000
[ 8] 596 -678 -561 383 -11 385 691 1000
[ 9] 267 -273 -299 152 -122 211 929 751 1000
[10] -565 384 380 -407 203 -337 -715 -588 -658 1000
[11] 316 -241 -239 187 -201 235 869 547 831 -838 1000
```

Initialize: LinkPrep

Statistical triplet: 375 11

Categories file (.cat):

Selected variable (default=1):

Output file name:

Discriminant analysis: Run

--.dis input file:

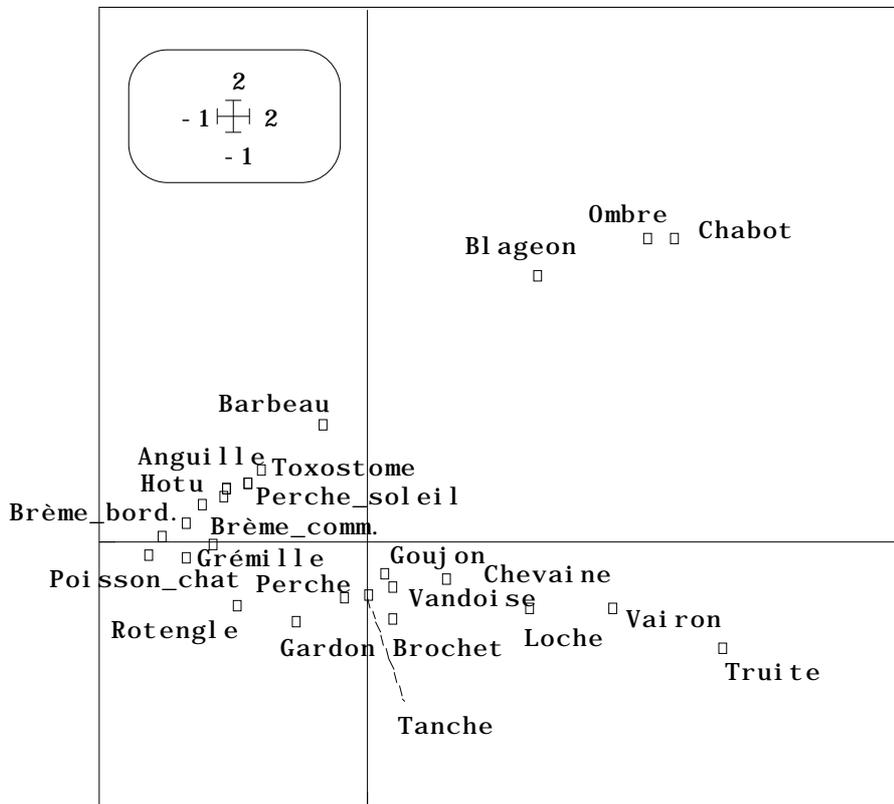
Observer alors que les valeurs propres de l'analyse discriminante et les valeurs propres de l'analyse canonique sont identiques :

| DISCRI.diup | | | CCA.ivup | | |
|-------------|--------|--------|----------|--------|--------|
| | 1 | 2 | | 1 | 2 |
| 1 | 0.3480 | 0.5613 | 1 | 0.3480 | 0.5613 |
| 2 | 0.1247 | 0.2011 | 2 | 0.1247 | 0.2011 |
| 3 | 0.0758 | 0.1222 | 3 | 0.0758 | 0.1222 |
| 4 | 0.0274 | 0.0442 | 4 | 0.0274 | 0.0442 |
| 5 | 0.0145 | 0.0233 | 5 | 0.0145 | 0.0233 |
| 6 | 0.0107 | 0.0173 | 6 | 0.0107 | 0.0173 |
| 7 | 0.0067 | 0.0109 | 7 | 0.0067 | 0.0109 |
| 8 | 0.0046 | 0.0074 | 8 | 0.0046 | 0.0074 |
| 9 | 0.0035 | 0.0056 | 9 | 0.0035 | 0.0056 |
| 10 | 0.0024 | 0.0038 | 10 | 0.0024 | 0.0038 |
| 11 | 0.0018 | 0.0029 | 11 | 0.0018 | 0.0029 |

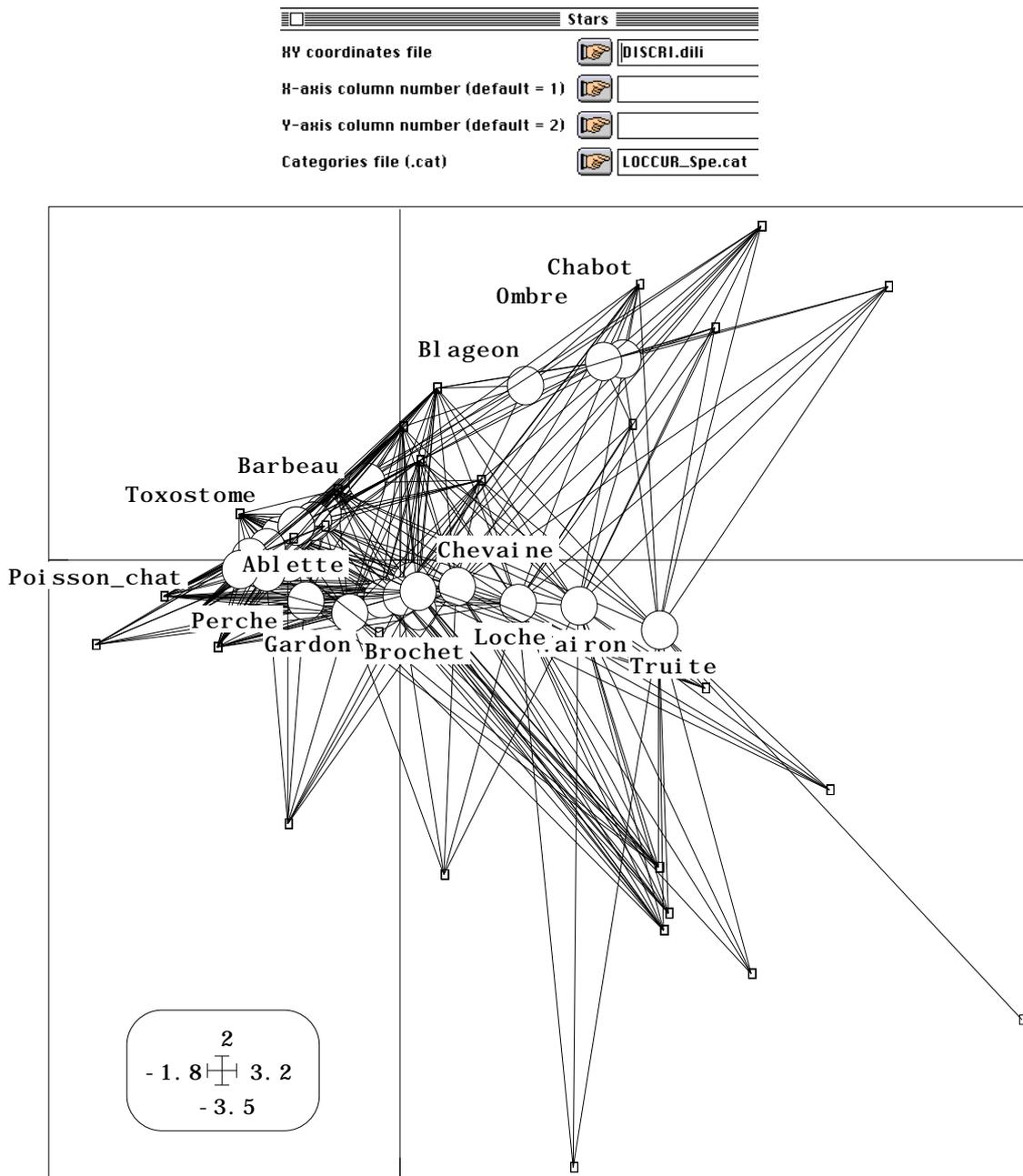
Observer encore l'identité de la construction des variables canoniques :

| CCA.ivfa | | | | DISCRI.difa | | | |
|----------|---------|---------|---------|-------------|---------|---------|---------|
| | 1 | 2 | 3 | | 1 | 2 | 3 |
| 1 | -2.0553 | -2.5965 | -0.6542 | 1 | -2.0553 | -2.5965 | -0.6542 |
| 2 | -0.1570 | -1.8841 | 0.0766 | 2 | -0.1569 | -1.8841 | -0.0766 |
| 3 | 0.4967 | 0.0505 | 0.1439 | 3 | 0.4967 | 0.0505 | -0.1439 |
| 4 | 1.3572 | 1.0778 | 1.0181 | 4 | 1.3572 | 1.0778 | -1.0181 |
| 5 | 0.2109 | 0.2076 | -0.0422 | 5 | 0.2109 | 0.2076 | 0.0422 |
| 6 | 0.1103 | 0.2667 | -0.6003 | 6 | 0.1103 | 0.2667 | -0.6003 |
| 7 | 0.2390 | 0.0321 | 0.0684 | 7 | 0.2390 | 0.0321 | -0.0684 |
| 8 | 0.3417 | 0.2902 | 0.7639 | 8 | 0.3417 | 0.2902 | -0.7639 |
| 9 | -0.4906 | -0.6734 | -0.9376 | 9 | -0.4906 | -0.6734 | 0.9376 |
| 10 | 0.3264 | 0.2841 | -0.6141 | 10 | 0.3264 | 0.2841 | 0.6141 |
| 11 | 0.3767 | 0.4456 | -0.8042 | 11 | 0.3767 | 0.4456 | 0.8042 |

L'ACC donnera la position des espèces (CCA.ivco):



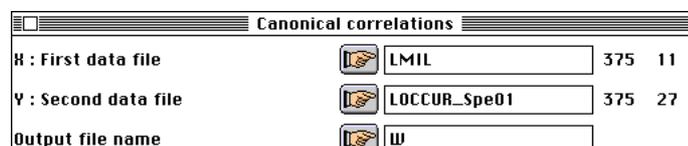
L'analyse discriminante fera de même en indiquant la taille de la niche :



On retrouve la position des relevés par des combinaisons de variables de milieu de variance unité maximisant la variance de la position moyenne des taxons (chaque espèce est liée à chacun des relevés qui la contient).

4.4 — L'ACC est une analyse canonique

Reste alors une opération qui ne manque pas de sel, à savoir l'analyse canonique pure et simple de LMIL et LOCCUR_Spe01 :



First input file (X): LMIL
 Rows: 375 Columns: 11

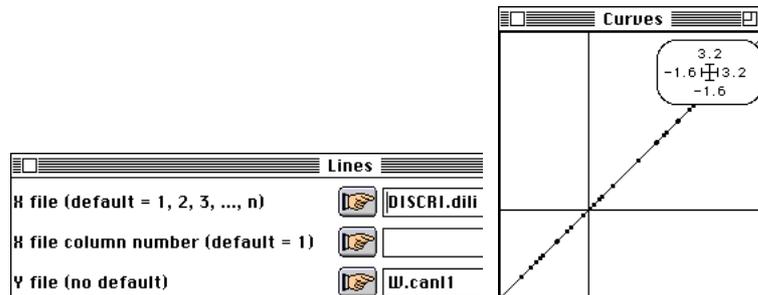
Second input file (Y): LOCCUR_Spe01
 Rows: 375 Columns: 27

| Num. | Eigenval. | R.Iner. | R.Sum | Num. | Eigenval. | R.Iner. | R.Sum |
|------|-------------|---------|---------|------|-------------|---------|---------|
| 01 | +3.4796E-01 | +0.5613 | +0.5613 | 02 | +1.2470E-01 | +0.2011 | +0.7624 |
| 03 | +7.5777E-02 | +0.1222 | +0.8846 | 04 | +2.7416E-02 | +0.0442 | +0.9288 |
| 05 | +1.4472E-02 | +0.0233 | +0.9522 | 06 | +1.0695E-02 | +0.0173 | +0.9694 |
| 07 | +6.7500E-03 | +0.0109 | +0.9803 | 08 | +4.5739E-03 | +0.0074 | +0.9877 |
| 09 | +3.4550E-03 | +0.0056 | +0.9933 | 10 | +2.3696E-03 | +0.0038 | +0.9971 |
| 11 | +1.7958E-03 | +0.0029 | +1.0000 | | | | |

Projected variable file: CCA.ivta
 It has 30 rows and 27 columns
 Inertia: 6.1997e-01
 File CCA.ivpc contains the column weights
 It has 27 rows and 1 column
 File CCA.ivpl contains the row weights
 It has 30 rows and 1 column

| Num. | Eigenval. | R.Iner. | R.Sum | Num. | Eigenval. | R.Iner. | R.Sum |
|------|-------------|---------|---------|------|-------------|---------|---------|
| 01 | +3.4796E-01 | +0.5613 | +0.5613 | 02 | +1.2470E-01 | +0.2011 | +0.7624 |
| 03 | +7.5777E-02 | +0.1222 | +0.8846 | 04 | +2.7416E-02 | +0.0442 | +0.9288 |
| 05 | +1.4472E-02 | +0.0233 | +0.9522 | 06 | +1.0695E-02 | +0.0173 | +0.9694 |
| 07 | +6.7500E-03 | +0.0109 | +0.9803 | 08 | +4.5739E-03 | +0.0074 | +0.9877 |
| 09 | +3.4550E-03 | +0.0056 | +0.9933 | 10 | +2.3696E-03 | +0.0038 | +0.9971 |
| 11 | +1.7958E-03 | +0.0029 | +1.0000 | | | | |

Les valeurs propres des analyses sont encore strictement identiques.



Les variables canoniques de l'ACC et de l'AD sont les mêmes. Ils sont constants par relevés et sont exactement ceux de l'ACC (édités par relevés) :

File W.can11 contains canonical variates (X)
 Rows = 375 Col = 3

File :W.can11

| Col. | Mini | Maxi |
|------|------------|-----------|
| 1 | -1.556e+00 | 3.198e+00 |
| 2 | -3.454e+00 | 1.897e+00 |
| 3 | -2.099e+00 | 4.939e+00 |

File CCA.iv11 contains canonical row scores
 linear combination of explanatory variables with unit norm
 It has 30 rows and 3 columns

File :CCA.iv11

| Col. | Mini | Maxi |
|------|------------|-----------|
| 1 | -1.556e+00 | 3.198e+00 |
| 2 | -3.454e+00 | 1.897e+00 |
| 3 | -4.939e+00 | 2.099e+00 |

On arrêtera ici l'expérience. L'ACC est une analyse canonique sur les indicatrices des espèces qui est une analyse discriminante inter-espèces.

5 — Conclusion

5.1 — Quand les données sont des relevés

Utiliser une stratégie d'analyse des correspondances pour traiter un tableau florofaunistique, c'est décider que seules les occurrences des espèces observées sont en cause. L'AFC du tableau est l'analyse canonique entre les indicatrices des espèces et les indicatrices des relevés, c'est donc une double analyse discriminante sur un paquet d'indicatrices les groupes étant définis par l'autre.

Il s'agit d'une stratégie très particulière qui n'est pas utilisée comme telle ²³. La métrique du Khi2 très largement évoquée comme fondement de l'utilisation en écologie est un argument fallacieux. Elle se justifie pleinement si la notion de relevé comme individu statistique est en cause. C'est le cas par exemple des mélanges de relevés de taille variable (impossibilité de contrôler l'espace représenté), d'auteurs et de conditions diverses (pêche électrique et pêche au filet, bonnes et mauvaises conditions météorologiques, ...). La seule nécessité porte sur la réalité de chaque occurrence qui est l'individu statistique réel.

Quand on rajoute un tableau de variables mésologiques sur les mêmes relevés, la stratégie d'ACC étend la discussion. On doit l'employer ou s'en méfier pour les mêmes raisons. C'est une analyse canonique entre les indicatrices des espèces et les variables de milieu dupliquées sur toutes les occurrences d'un relevé. C'est donc une analyse discriminante décrite très clairement par son inventeur ²⁴. La perturber par un *detrending*, c'est nier toute valeur à la précision du support mathématique. Y voir une ACPVI ²⁵ n'est sans doute pas très utile en écologie. En tant qu'analyse canonique, elle n'est pas utilisée comme telle. Mais la maîtrise de ce point de vue est prometteur. En effet l'intégration des niveaux taxonomiques de détermination peut se faire par le biais de sous-espaces emboîtés des indicatrices des espèces de manière symétrique à la définition par les variables de milieu d'un sous-espace des indicatrices des relevés.

Globalement, l'ACC est une analyse canonique qui mérite d'être repérée comme telle. Elle impose des contraintes très fortes qui sont souvent inutiles ou défavorables. Elle propose des développements qu'il conviendra d'explorer.

5.2 — Quand les données sont des occurrences

Quand il n'y a plus de relevés, mais des listes d'occurrences référencées à un milieu, les individus statistiques sont clairement et uniquement ces occurrences. Le programme d'analyse canonique des correspondances n'est plus utilisable, mais le concept est alors parfaitement valide. Définir une méthode par un programme *vire* ici à la caricature.

L'analyse canonique des correspondances, comme analyse canonique entre les indicatrices des espèces et les variables de milieu portant sur les occurrences devient optimale tant conceptuellement que numériquement. On peut faire de l'analyse canonique des correspondances sans relevés et ... s'attendre à quelques discussions pigmentées !

Références

- ¹ Chessel, D. & Mercier, P. (1993) Couplage de triplets statistiques et liaisons espèces-environnement. In : *Biométrie et Environnement*. Lebreton, J.D. & Asselain, B. (Eds.) Masson, Paris. 15-44.
- ² Gittins, R. (1985) *Canonical analysis, a review with applications in ecology*. Springer-Verlag, Berlin. 1-351.
- ³ Tucker, L.R. . (1958) An inter-battery method of factor analysis. *Psychometrika* : 23, 2, 111-136.
- ⁴ Browne, M.W. (1979) The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology* : 32, 75-86.
- ⁵ Escoufier, Y. (1987) The duality diagramm : a means of better practical applications. In : *Development in numerical ecology*. Legendre, P. & Legendre, L. (Eds.) NATO advanced Institute , Serie G .Springer Verlag, Berlin. 139-156.
- ⁶ Gauch, H.G. Jr. & Wentworth, T.R. (1976) Canonical correlation analysis as an ordination technique. *Vegetatio* : 33, 1, 17-22.
- ⁷ Austin, M.P. (1985) Continuum concept, ordination methods and niche theory. *Annual Review of Ecology and Systematics* : 16, 39-61.
- ⁸ Austin, M.P. (1987) Models for the analysis of species' response to environmental gradients. *Vegetatio* : 69, 35-45.
- ⁹ Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya, A* : 26, 329-359.
- ¹⁰ Wollenberg, A.L. (1977) Redundancy analysis, an alternative for canonical analysis. *Psychometrika* : 42, 2, 207-219.
- ¹¹ Muller, K.E. (1981) Relationships between redundancy analysis, canonical correlation, and multivariate regression. *Psychometrika* : 46, 139-142.
- ¹² Gleason, T.C. (1976) On redundancy in canonical analysis. *Psychological Bulletin* : 83, 1004-1006.
- ¹³ Israels, A.Z. (1984) Redundancy analysis for qualitative variables. *Psychometrika* : 49, 661-346.
- ¹⁴ Obadia, J. (1978) L'analyse en composantes explicatives. *Revue de Statistique Appliquée* : 24, 4, 5-28.
- ¹⁵ Ter Braak, C.J.F. (1987) CANOCO, a FORTRAN program for Canonical community ordination by [partial][detrended][canonical] correspondence analysis and redundancy analysis. Software documentation. Version 2.1, TNO Institute of Applied Computer Science, Wageningen..
- ¹⁶ 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.

- ¹⁷ Sabatier, R., Lebreton, J.D. & Chessel, J.D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In : *Multivariate data analysis*. Coppi, R. & Bolasco, S. (Eds.) Elsevier Science Publishers B.V., North-Holland. 341-352.
- ¹⁸ 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.
- ¹⁹ 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.
- ²⁰ Ter Braak, C.J.F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* : 67, 1167-1179.
- ²¹ Ter Braak, C.J.F. (1988) Partial Canonical Analysis. In : *Classification and related methods of data analysis*. Bock, H.H. (Ed.) North Holland. 551-558.
- ²² Hill, M.O. & Gauch, H.G. (1980) Detrended correspondence analysis: an improved ordination technique. *Vegetatio* : 42, 47-58.
- ²³ Thioulouse, J. & Chessel, D. (1992) A method for reciprocal scaling of species tolerance and sample diversity. *Ecology* : 73, 670-680.
- ²⁴ Ter Braak, C.J.F. (1987) The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* : 69, 69-77.
- ²⁵ Sabatier, R., Lebreton, J.D. & Chessel, J.D. (1989) Principal component analysis with instrumental variables as a tool for modelling composition data. In : *Multivariate data analysis*. Coppi, R. & Bolasco, S. (Eds.) Elsevier Science Publishers B.V., North-Holland. 341-352.