

# Analyse Canonique des Correspondances

## Résumé

L'Analyse Canonique des Correspondances (Ter\_Braak, C.J.F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* : 69, 69-77) est la version ACP/AFC d'une méthode très générale de couplages de tableaux qu'on peut mettre en œuvre sur tout type de variables dans le module Projectors. Les illustrations portent sur des relevés avifaunistiques couplés à des relevés d'architecture de la végétation.

## Plan

1 — Analyse des correspondances préliminaire .....	2
1.1 — Données traitées (Cartes AviAnnaba) .....	2
1.2 — Moyennes réciproques .....	4
1.3 — Sélection de données .....	6
1.4 — Réorganisation du tableau.....	7
1.5 — Variables supplémentaires .....	9
2 — Analyse discriminante.....	15
3 — Analyse canonique des correspondances .....	20
3.1 — Mise en œuvre.....	21
3.2 — L'ACC comme AFC sous contrainte .....	25
3.3 — L'ACC comme AFC à coordonnées modélisables.....	29
3.4 — L'ACC comme analyse d'inertie .....	35
4 — Conclusion .....	36
4.1 — Bilan méthodologique .....	36
4.2 — Bilan expérimental .....	37
Références .....	38

D. Chessel, S. Benyacoub & S. Dolédec

# 1 — Analyse des correspondances préliminaire

L'analyse canonique des correspondances est une méthode très liée à l'analyse des correspondances. Elle s'emploie pour étudier les relations entre un tableau X de paramètres mésologiques et un tableau Y de relevés faunistiques ou floristiques. Fondamentalement le tableau Y est considéré comme relevant de l'analyse des correspondances. Les illustrations sont extraites de la thèse d'un des auteurs<sup>1</sup>.

## 1.1 — Données traitées (Cartes AviAnnaba)

220 relevés fournissent deux types de mesures. Le premier concerne l'avifaune, le second l'architecture de la végétation. Ces 220 relevés sont uniformément répartis (20 relevés par strate) en onze strates d'échantillonnage, à savoir :

- a — Pelouses (h = 25 cm)                      b — Maquis bas (h = 60 cm)  
 c — Maquis moyen (h = 1,40 m)            d — Maquis haut (h = 2,40 m)  
 e — Maquis à strate arborée claire (h = 5,50 m et r = 14%)  
 f — Maquis à strate arborée dense (h = 5,40 m et r = 35%)  
 g — Subéraie à sous-bois (h=7,70 m)    h — Subéraie sans sous-bois (h=10,4 m)  
 i — Zénaie (h=17,6 m)                      j — Pinéraie à pin maritime (h=7,60m)  
 k — Eucalyptaie (h=18 m)

Le tableau d'avifaune Dens compte 220 lignes et 48 colonnes-espèces dont le code est contenu dans le tableau 1. Il donne une estimation de l'abondance locale (stationnelle) d'une espèce dans un relevés.

1	COCO	Caille des blés	25	SYAT	Fauvette à tête noire
2	COPA	Pigeon ramier	26	SYHO	Fauvette orphée
3	STTU	Tourterelle des bois	27	SYCO	Fauvette grisette
4	CUCA	Coucou gris	28	SYCN	Fauvette à lunette
5	PIVI	Pic de Levaillant	29	SYUN	Fauvette pitchou
6	PIMA	Pic épeiche	30	PHCO	Pouillot véloce
7	PIMI	Pic épeichette	31	PHBO	Pouillot de Bonelli
8	JYTO	Torcol	32	REIG	Roitelet triple bandeau
9	GACR	Cochevis huppé	33	MUST	Gobe-mouches gris
10	GATH	Cochevis de Thekla	34	FIHY	Gobe-mouches noir
11	ALAR	Alouette des champs	35	PAMA	Mésange charbonnière
12	LUAR	Alouette lulu	36	PACA	Mésange bleue
13	ANCA	Pipit rousseline	37	PAAT	Mésange noire
14	MOFL	Bergeronnette printanière	38	CEBR	Grimpereau des jardins
15	PYBA	Bulbul gris	39	MICA	Bruant proyer
16	TCSE	Téléphone tchagra	40	COCR	Gros-bec
17	LASE	Pie-grièche à tête rousse	41	CACH	Verdier
18	TRTR	Troglodyte	42	CACR	Chardonneret
19	TUVI	Grive draine	43	CACN	Linotte mélodieuse
20	TUME	Merle noir	44	SESE	Serin cini
21	SATO	Traquet pâte	45	FRCO	Pinson des arbres
22	LUME	Rosignol philomèle	46	STUN	Etourneau unicolor
23	ERRU	Rouge-gorge	47	OROR	Loriot d'Europe
24	CIJU	Cisticole des joncs	48	GAGL	Geai des chênes

Tableau 1 : code des taxons du tableau 220 relevés - 48 espèces

Le tableau de milieu Mil compte 220 lignes et 8 colonnes-variables, à savoir :

- 1 - Altitude (m)                                      2 - Hauteur totale de la strate arborée (m)  
 3 - Hauteur de la première branche (m)      4 - Recouvrement de la strate arborée (%)  
 5 - Hauteur de la strate buissonnante (m)    6 - Recouvrement de la strate buissonnante  
 7 - Hauteur de la strate herbacée (m)        8 - Recouvrement de la strate herbacée (%)

On est en présence du problème classique de couplage d'un tableau faunistique et d'un tableau mésologique. Le tableau Dens relève d'une analyse des correspondances<sup>2</sup>. On compare ici trois méthodes d'introduction de l'information sur le milieu dans l'interprétation de la structure faunistique. On rappelle ainsi une des propriétés fondamentale de l'analyse des correspondances en usage en écologie.

Exécuter l'analyse des correspondances du tableau Dens (module COA) :

```

CORrespondence Analysis
Data file [Icon] Dens 220 48

Input file: Dens.fcta
--- Number of rows: 220, columns: 48
-----
Total inertia: 3.91579
Num. Eigenval.  R.Iner.  R.Sum  | Num. Eigenval.  R.Iner.  R.Sum  |
01  +7.9612E-01 +0.2033 +0.2033 | 02  +5.2279E-01 +0.1335 +0.3368 |
03  +3.0008E-01 +0.0766 +0.4135 | 04  +2.2619E-01 +0.0578 +0.4712 |
...

```

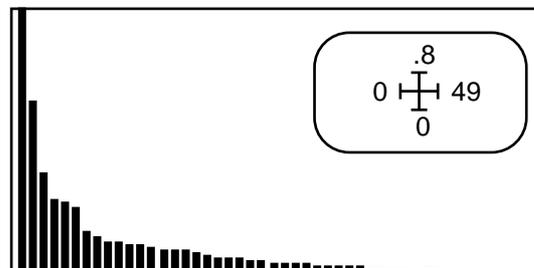


Figure 1 : Valeurs propres de l'analyse des correspondances du tableau Dens (220-48).

Conserver 6 facteurs. Il est clair que deux facteurs sont certainement interprétables (figure 1). Effectuer la carte des relevés en utilisant la partition en strates comme information complémentaire. Créer un fichier Stra avec l'option Creat1Categ de TextToBin (11 classes de 20 points) et utiliser ScatterClass pour obtenir la figure 2 (en haut). Ne pas oublier d'introduire la pondération des lignes propre à l'AFC :

```

Stars
HY coordinates file [Icon] Dens.fcli 220 6
H-axis column number (default = 1) [Icon]
Y-axis column number (default = 2) [Icon]
Categories file (.cat) [Icon] Stra.cat
Variable label file (or #) [Icon]
Row weights file (option) [Icon] Dens.fcpl 220 1
Draw points (no = 2) [Icon]

```

L'axe 1 isole la strate pelouse et l'axe 2 ordonne les autres strates. L'expérience se distingue donc franchement de celle de Blondel et Frochet (Blondel & Coll.<sup>3</sup>), comme celle de Prodon & Lebreton<sup>4</sup> par absence d'effet Guttman discuté dans Lebreton & Coll.<sup>5</sup>).

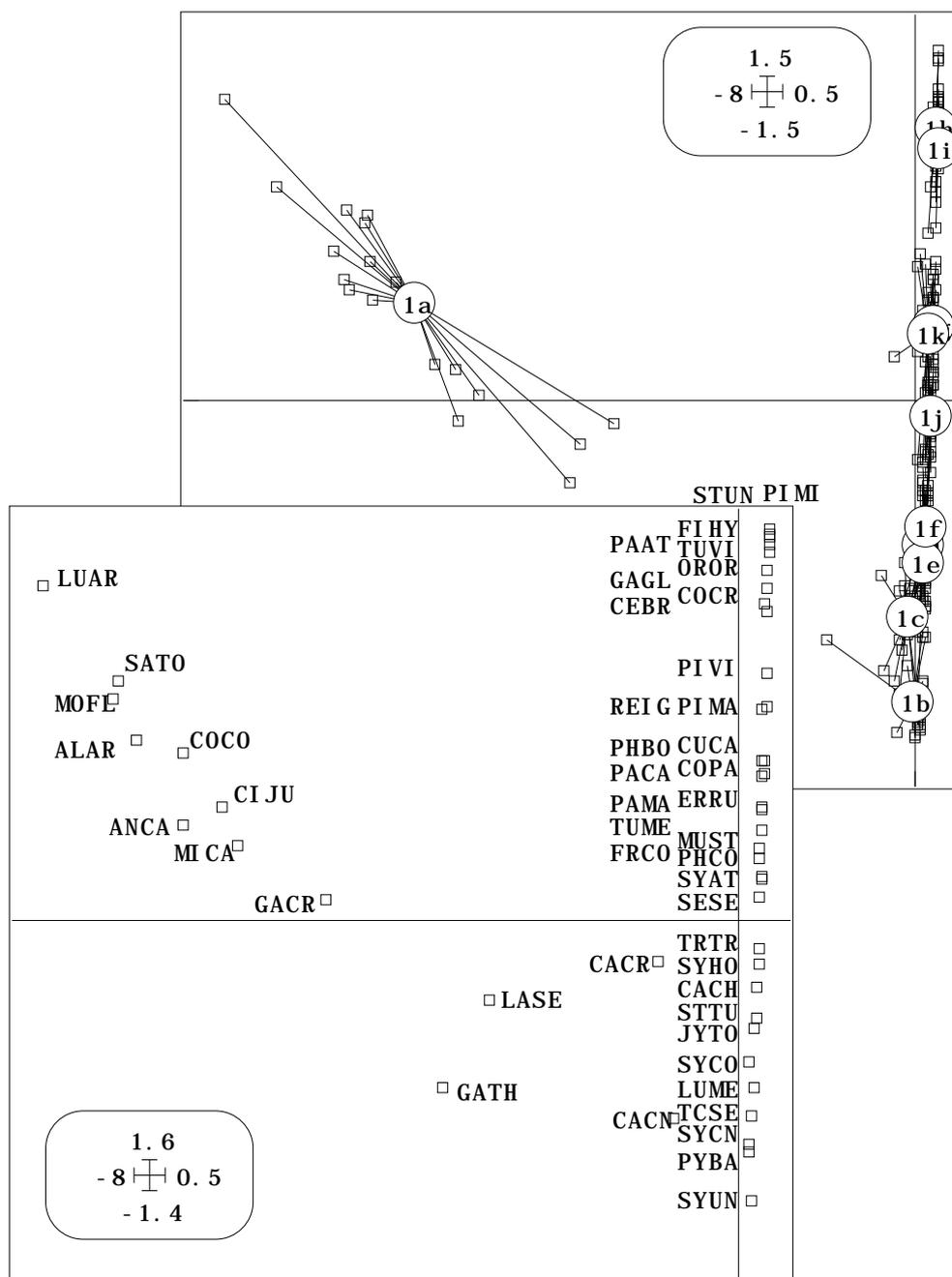


Figure 2 : Plans 1-2 de l'analyse des correspondances du tableau Dens (220-48).

La carte des espèces est obtenue avec Scatter (Figure 2, en bas). On obtient deux structures typiques (partition strate a contre les autres sur l'axe 1, ordination des strates b à j sur l'axe 2) articulées par les espèces de transition entre les strates a et b,c.

## 1.2 — Moyennes réciproques

Une des propriétés de l'analyse des correspondances est de positionner les espèces par des codes numériques pour maximiser la variance des positions des relevés obtenus par averaging, comme de positionner les relevés par des codes numériques pour maximiser la variance des positions des espèces obtenues par averaging<sup>6</sup>. Les deux cartes de la figure 2 sont donc reliées par des relations de moyennes conditionnelles à une dilatation près. Le module CoaBetCol permet d'obtenir la figure 3 pour illustrer le principe.



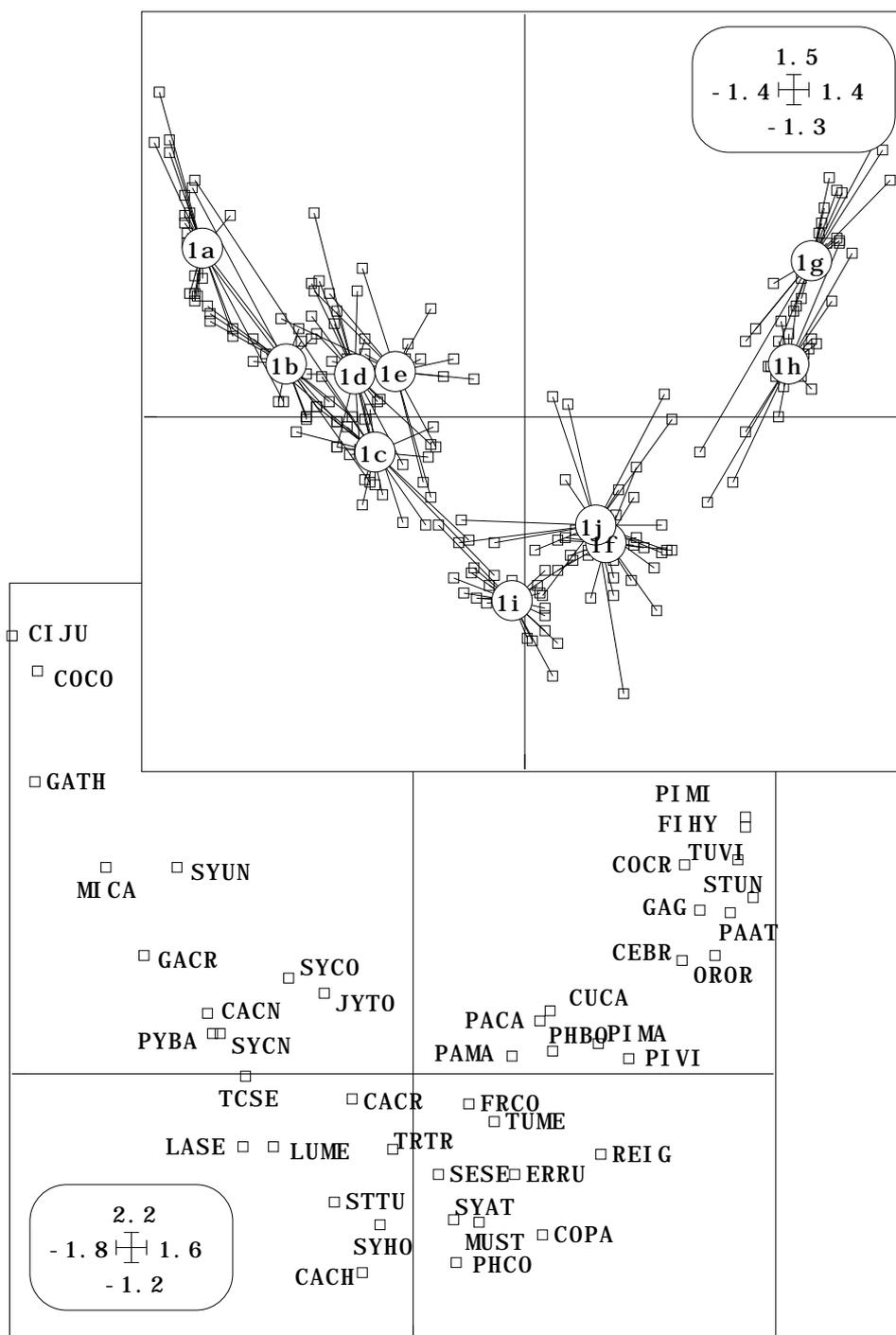


Figure 4 : Plans 1-2 de l'analyse des correspondances du tableau D (220-43).

### 1.3 — Sélection de données

Extraire les lignes 21 à 220 du tableau Dens, avec RowColSelection (dans FilesUtil), pour obtenir D1 (200 lignes et 48 colonnes).

Row-Col Selection		
Input file	<input type="text" value="Dens"/>	220 48
Selection of rows (default = all)	<input type="text" value="21a220"/>	
Selection of columns (default = all)	<input type="text"/>	
Output file	<input type="text" value="D1"/>	

Éliminer les colonnes nulles (taxons strictement limités à la strate a), avec SelectTaxa (présence minimum = 1), pour obtenir D (200 lignes et 43 colonnes). Éliminer avec Excel™ les noms d'espèces n° 11, 12, 13, 14 et 21 pour obtenir Code\_Esp (43 chaînes de 4 caractères). Refaire l'analyse des correspondances de D :

Correspondence Analysis		
Data file	<input type="text" value="D"/>	200 43

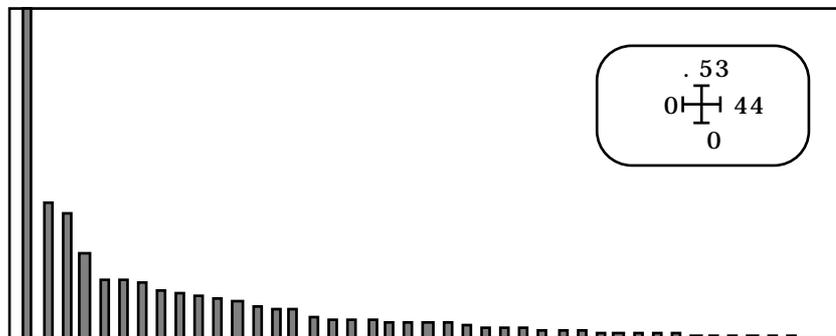


Figure 5 : Valeurs propres de l'analyse des correspondances de D.

Garder 4 facteurs en observant que le premier (qui est le second de la précédente analyse) est prépondérant, d'autant plus que le second semble être un effet Guttman sur le premier (figure 4). Extraire du tableau Mil (220-8) les lignes 21 à 220 pour obtenir M (200-8) :

Row-Col Selection		
Input file	<input type="text" value="Mil"/>	220 8
Selection of rows (default = all)	<input type="text" value="21a220"/>	
Selection of columns (default = all)	<input type="text"/>	
Output file	<input type="text" value="M"/>	

Refaire le fichier Stra (200-1), en utilisant 10 classes de 20 points. On cherche dans M la combinaison de variables qui explique le gradient.

#### 1.4 — Réorganisation du tableau.

L'analyse des correspondances code numériquement les espèces (figure 6A). Ce code permet d'attribuer à chaque relevé une moyenne et une variance (figure 6B : moyenne figurée par un cercle, intervalle d'un écart-type autour de la moyenne figuré par un segment). La variance des moyennes est optimisée. En multipliant chaque moyenne par la racine carrée de la première valeur propre, on obtient un code numérique des relevés (figure 6C) qui permet de positionner les espèces par moyenne conditionnelle (figure

6D). Là encore on maximise la variance des moyennes conditionnelles. En multipliant chaque moyenne par la racine carrée de la première valeur propre, on obtient un code numérique des espèces (figure 6A).

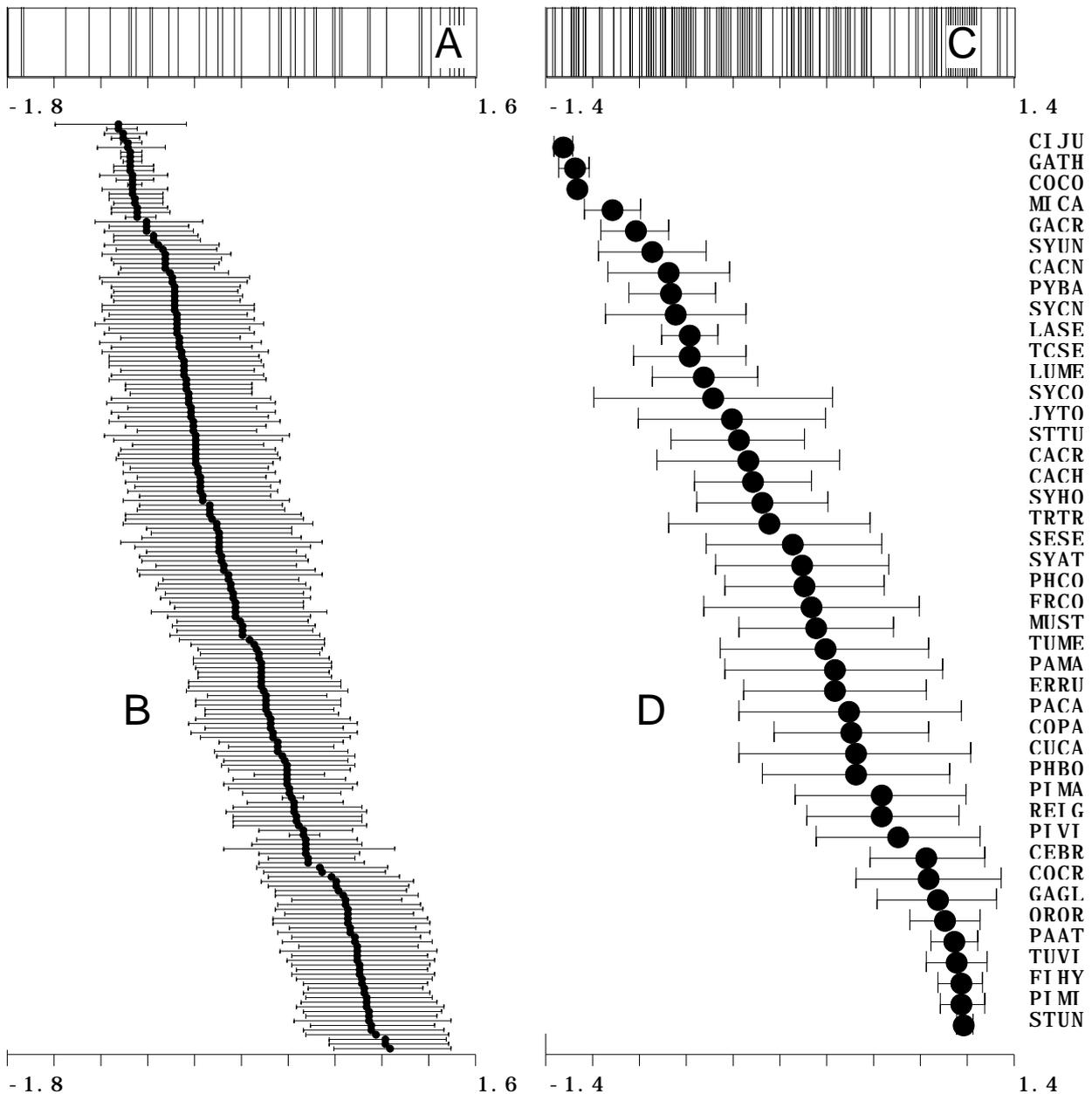


Figure 6: Ordination réciproque des espèces et des relevés par l'analyse des correspondances du tableau D (220-43). La figure est obtenue par deux passages dans **Tab-MeanVar**.

L'AFC optimise la variance des positions des relevés par un code numérique qui dérive du tableau par double averaging. On peut discuter à partir de la figure 6 la diversité en espèces d'un relevé ou l'amplitude d'habitat d'un taxon<sup>7</sup>. Pour faire la figure 6, utiliser le module Tab-MeanVar. Pour la partie gauche utiliser les paramètres :

Tab-MeanVar			
Table file (Bin)	D	200	43
1=Column curve/2=Row curve (default=1)	2		
H file (Bin) (default=1,2, ..., n)	D.fcco	43	4
Selected column (default=1)	1		
Option : curve ordination file (Bin)	D.fcli	200	4
Selected column (default=1)	1		
1 = Curve label (default=No)			
Option : Label curve file (Txt)			

et pour la partie droite utiliser les paramètres :

Tab-MeanVar			
Table file (Bin)	D	200	43
1=Column curve/2=Row curve (default=1)	1		
H file (Bin) (default=1,2, ..., n)	D.fcli	200	4
Selected column (default=1)	1		
Option : curve ordination file (Bin)	D.fcco	43	4
Selected column (default=1)	1		
1 = Curve label (default=No)	1		
Option : Label curve file (Txt)	Code_Esp		

Assembler sous MacDraw™. On retiendra dans la figure 6 la partie droite (C et D). Le code des relevés est issu du tableau traité. On peut imposer qu'il provienne d'une combinaison de variables d'un tableau distinct du tableau traité. On parle d'analyse des correspondances sous contraintes ou d'analyse des correspondances sur variables instrumentales<sup>8</sup> ou d'analyse canonique des correspondances (canonical correspondence analysis, nom choisi par l'inventeur de la méthode<sup>9</sup>).

## 1.5 — Variables supplémentaires

La plus simple des manières de relier la structure d'ordination des relevés (le facteur 1 de l'AFC de D) et les variables de milieu est d'étudier le mode de variation de ces variables en fonction de ce facteur. Il faut d'abord normaliser ces variables. Le code issu de l'AFC étant centré pour la pondération marginale du tableau D, il est indispensable, pour des raisons de cohérence, d'utiliser cette pondération pour réduire le tableau M. Utiliser PCA avec :

Correlation matrix PCA			
Matrix input file	 M	200	8
Row weights (default=1/n)	 3		
Column weights (default=1)			
Option: file for row weighting	 D.fcpl	200	1
Option: file for column weighting			
1 = Save correlation matrix	 1		

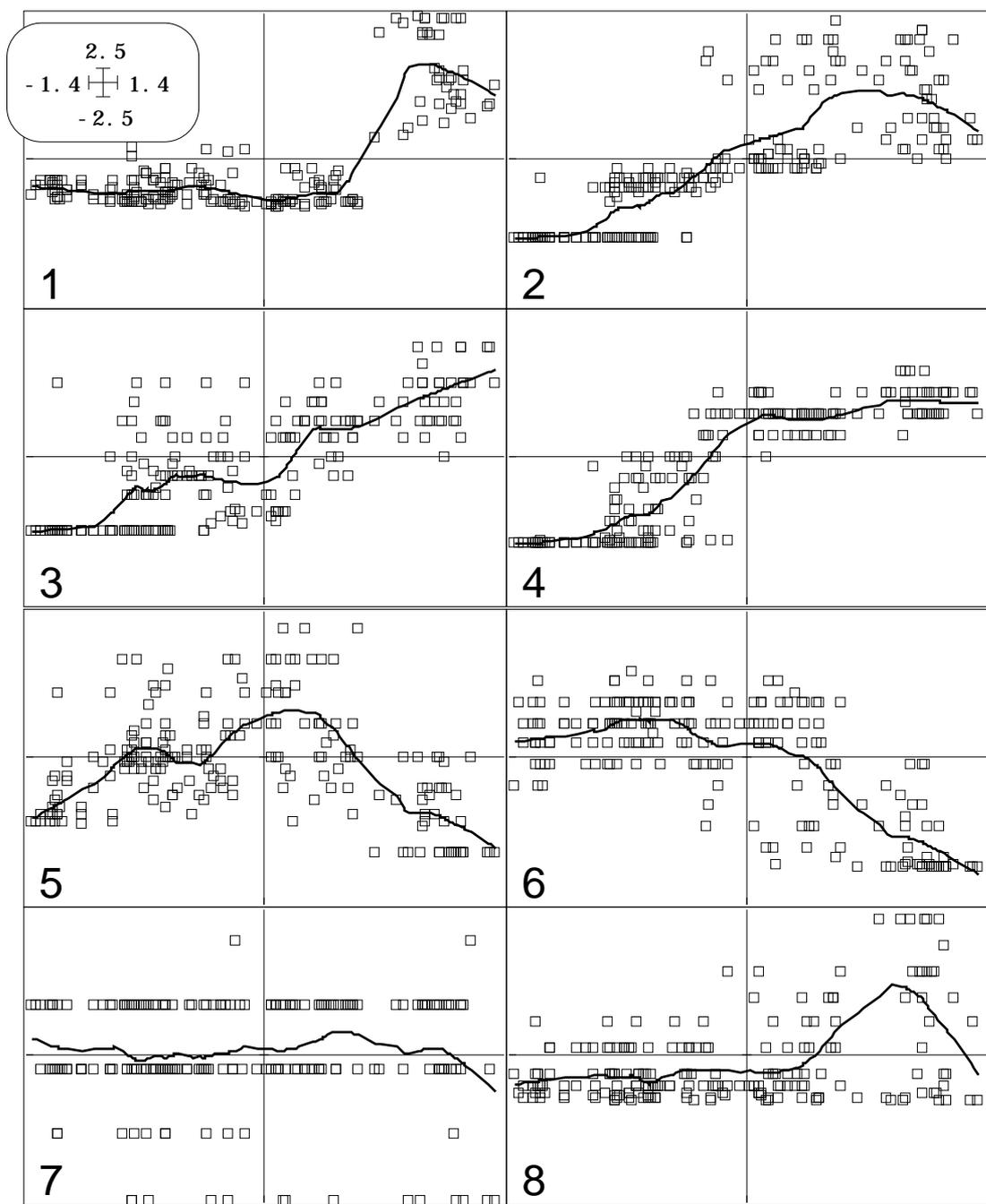


Figure 7 : En abscisse position des 200 relevés sur le facteur 1 de l'AFC de Dens. En ordonnée, valeurs normalisées des 8 variables de milieu. 1 - Altitude ,2 - Hauteur totale de la strate arborée, 3 - Hauteur de la première branche, 4 - Recouvrement de la strate arborée (%), 5 - Hauteur de la strate buissonnante, 6 - Recouvrement de la strate buissonnante, 7 - Hauteur de la strate herbacée, 8 - Recouvrement de la strate herbacée.

```

Classical Principal Component Analysis (Hotteling 1933)
Input file: M
---- Row weights:
File M.cnpl contains the row weights. It has 200 rows and 1 column
Row weights are selected from file D.fcpl (Sum = 1)
---- Column weights:
File M.cnpc contains the column weights. It has 200 rows and 1 column
Each column has unit weight (Sum = 8)
---- Table:
File M.cnta contains the centred and normed table
Zero mean and unit variance for each column. It has 200 rows and 8 columns
---- Info: means and variances
File M.cnma contains the descriptive of the analysis

```

```

It contains successively:
Number of rows: 200
Number of columns: 8
means and variances:
Col.: 1 Mean: 288.118 Variance: 115579
Col.: 2 Mean: 796.709 Variance: 355840
Col.: 3 Mean: 200.713 Variance: 25080.1
Col.: 4 Mean: 39.753 Variance: 752.946
Col.: 5 Mean: 148.47 Variance: 8318.37
Col.: 6 Mean: 53.4323 Variance: 807.397
Col.: 7 Mean: 22.2264 Variance: 81.3167
Col.: 8 Mean: 17.3949 Variance: 517.298

```

```

-----
File M.cn+r contains the Correlation matrix
from statistical triplet M.cnta
It has 8 rows and 8 columns

```

```

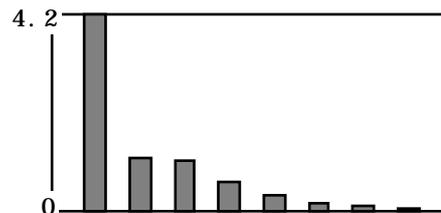
----- Correlation matrix -----
[ 1] 1000
[ 2] 651 1000
[ 3] 614 728 1000
[ 4] 499 792 603 1000
[ 5] -467 -157 -206 -95 1000
[ 6] -628 -628 -561 -548 620 1000
[ 7] 29 209 22 97 -58 -119 1000
[ 8] 734 673 390 404 -262 -452 376 1000

```

```

-----
Total inertia: 8

```



Le graphe des valeurs propres indique la présence d'un facteur essentiel. On a conservé trois facteurs en prévision des analyses suivantes. Obtenir alors la figure 7 avec Lowess (dans CurveModels) et les paramètres :

Lowess			
X file (default = 1, 2, 3, ..., n)	<input type="text" value="D.fcli"/>	200	4
X file column number (default = 1)	<input type="text"/>		
Y file (no default)	<input type="text" value="M.cnta"/>	200	8
Number of points for regression ?	<input type="text" value="40"/>		
Weight file (optional)	<input type="text" value="D.fcpl"/>	200	1
Variable label file (optional)	<input type="text" value="#"/>		

La relation entre les variables de milieu et le code  $F_1$  peut se quantifier de deux manières. La première consiste à calculer chaque corrélation et faire la moyenne des carrés de corrélation. La seconde consiste à faire la régression multiple du code  $F_1$  sur les 8 variables et à retenir le carré de corrélation multiple.

Pour obtenir les coefficients de corrélation simple, utiliser MatAlg :

Diagonal Inner product C=H'DY			
Input file for H matrix		M.cnta	200 8
Option for H matrix (default=none)		2	
Input file for Y matrix		D.fcli	200 4
Option for Y matrix (default=none)			
D inner product (default = 1/n)		2	
Option: weighting file		D.fcpl	200 1

qui affiche :

```

X input file: M.cnta
--- Number of rows: 200, columns: 8
Y input file: D.fcli
--- Number of rows: 200, columns: 4
Diagonal inner product: D.fcpl
XtDY output file: screen
--- Number of rows: 8, columns: 4
Input file: screen
--- Number of rows: 8, columns: 4
-----
[ 1]  7.1916e-01  4.0641e-01  3.5745e-01 -2.9833e-02
[ 2]  8.1337e-01 -4.9424e-02  3.7442e-02 -9.8269e-02
[ 3]  7.5467e-01  2.0188e-01 -5.7553e-02 -8.0774e-02
[ 4]  8.4718e-01 -2.1049e-01 -7.9569e-02 -9.5956e-02
[ 5] -2.1036e-01 -6.6323e-01  1.7338e-01 -1.3265e-01
[ 6] -7.1072e-01 -3.7757e-01  1.8950e-01 -1.9357e-01
[ 7]  9.0509e-03 -4.3928e-02  9.3247e-02  4.2700e-03
[ 8]  4.8698e-01  1.4559e-01  4.1143e-01  6.1083e-04
-----

```

Cette option sert aux calculs des matrices de corrélation, de covariances, de produits scalaires entre fichiers de manière très souple. Il convient de noter ici qu'on utilise un fichier D.fcli centré par colonnes (propriétés des coordonnées) qui est normalisé par l'option standardisation (2) et un fichier M.cnta qu'on ne modifie pas puisqu'il est déjà normalisé. L'usage de poids (2) et du fichier D.fcpl est impératif, puisqu'il définit les moyennes, variances et covariances associées à l'AFC de D. Le résultat est une matrice de corrélation variables (lignes), facteurs (colonnes). Demander une nouvelle exécution en conservant le résultat dans le fichier Cor8-4. Repasser alors les paramètres :

Diagonal Inner product C=H'DY			
Input file for H matrix		Cor8-4	8 4
Option for H matrix (default=none)			
Input file for Y matrix		Cor8-4	8 4
Option for Y matrix (default=none)			
D inner product (default = 1/n)			

On a :

```

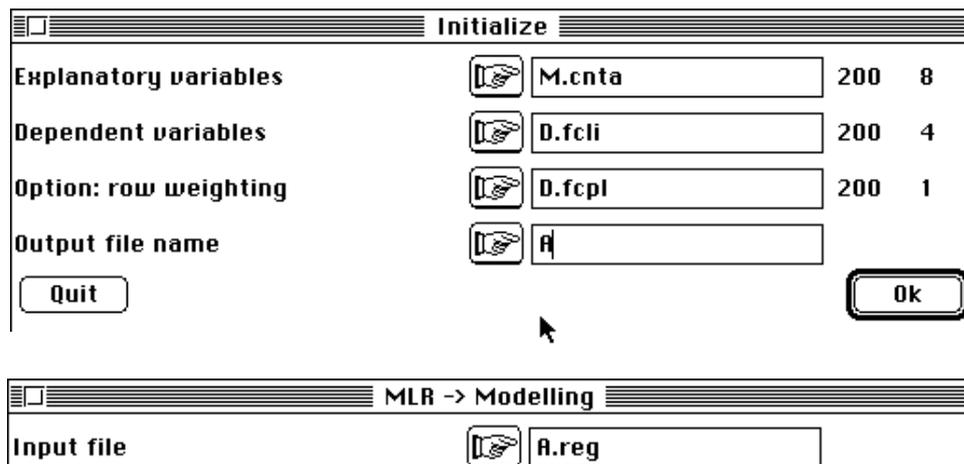
-----
[ 1]  4.0658e-01  8.8059e-02  2.5840e-02 -9.7271e-03
[ 2]  8.8059e-02  1.0728e-01  2.2264e-03  1.9699e-02
[ 3]  2.5840e-02  2.2264e-03  4.7845e-02 -7.6365e-03
[ 4] -9.7271e-03  1.9699e-02 -7.6365e-03  1.0171e-02
-----

```

ce qui indique que 0.4066 est la moyenne des 8 carrés de coefficients de corrélation linéaire variables-facteur. Il suffit de connaître la définition du produit de matrice  $Y^t D_n X$  de terme général :

$$a_{jk} = \sum_{i=1}^n P_i Y_{ij} X_{ik}$$

La liaison des variables de milieu avec le facteur 1 est donc forte et un test d'hypothèse est inutile. Pour obtenir le carré de corrélation multiple ( $R^2$ ), après l'ACP de M, utiliser alors LinearReg avec :



On a exécuté la régression multiple pondérée des facteurs (D.fcli) sur les variables (M.cnta).

#### Multiple Linear Regression

```

-----
Explanatory variable file: M.cnta
It has 200 rows and 8 columns
Var. | Mean      | Variance |
  1 | 4.237e-09 | 1.000e+00 |
  ...
  8 | 7.953e-09 | 1.000e+00 |
-----
Dependent variable file: D.fcli
It has 200 rows and 4 columns
R2 = Squared multiple correlation coefficient
Var. | Mean      | Variance | R2      |
  1 | 1.288e-09 | 7.275e-01 | 9.047e-01 |
  2 | 1.196e-09 | 4.681e-01 | 6.606e-01 |
  3 | -1.881e-09 | 4.510e-01 | 5.589e-01 |
  4 | -1.675e-09 | 3.700e-01 | 1.559e-01 |
-----

```

On vérifie que toutes les variables utilisées sont centrées pour la pondération utilisée. On a donc une très bonne prévisibilité (90% de variance expliquée) du  $F_1$  de l'ACP par les variables de milieu :

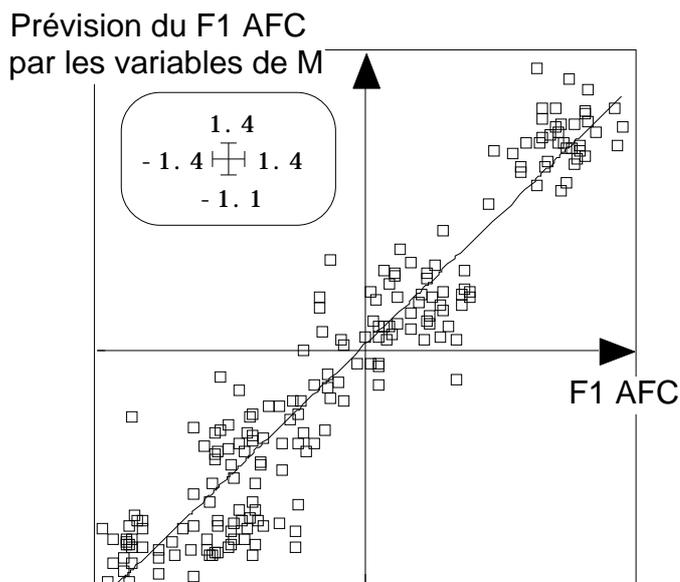
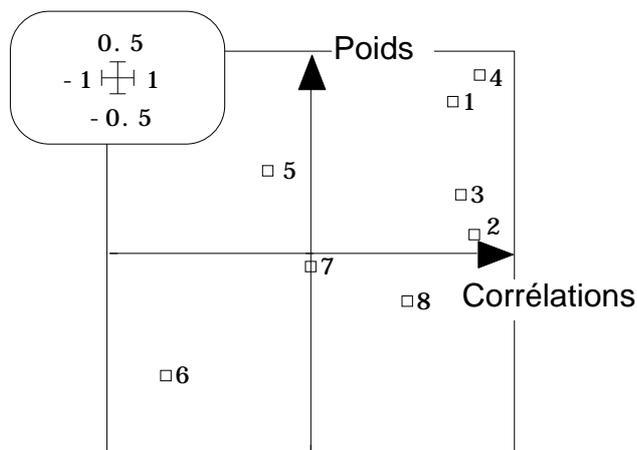


Figure 8 : Liaison entre le facteur 1 de l'AFC de D et sa prévision par les variables de M.

-----  
 File A.MLRw1 has 8 rows and 4 columns  
 It contains regression coefficients  
 Rows : explanatory variables / Columns : dependant variables  
 Models for normalized (mean = 0 / variance =1) variables



La relation entre poids (coefficient de régression) et corrélations ne montre pas d'aberrations et indique une situation numérique stable. On a ici cherché la combinaison linéaire de variable de milieu qui maximise sa corrélation avec le F1 du tableau d'avifaune. C'est donc normalement que cette corrélation (0.9505) est supérieure à celle qui existe entre le F1 du tableau d'avifaune et le F1 du tableau M (0.8608), obtenue comme précédemment (MatAlg) et représentée dans la figure 9.

Ces exercices montrent de multiples manières de relier les deux tableaux : on trouvera une liste étendue de telles techniques dans Mercier 1991<sup>10</sup>. La régression multiple du F1 de l'AFC sur les variables du milieu est la méthode utilisée dans Prodon & Lebreton 1981 (op. cit.). Les méthodes les plus récentes, en optimisant un critère reliant les deux tableaux, l'emportent en principe sur ces dernières, car rien ne garantit que le F1 de l'AFC est le code le plus prévisible par les variables de milieu. Cependant ce F1 est déjà très fortement relié à ces variables, donc vraisemblablement proche de l'optimum.

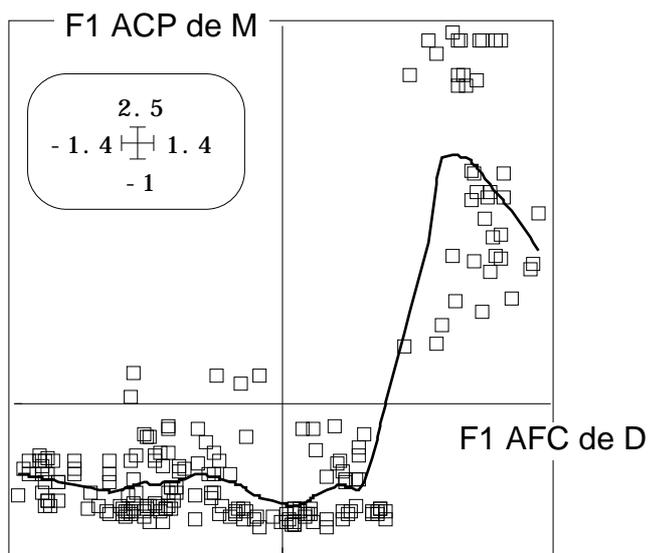


Figure 9 : Relation entre les premières coordonnées factorielles des deux analyses de base.

Dans les remarques qui précèdent rien n'est donc gratuit. Les concepts fondamentaux de l'analyse canonique des correspondances sont en jeu. Nous avons affaire à des gradients sur lesquels se disposent les niches écologiques des espèces, ce qui a été exploré par l'AFC<sup>11</sup>. Le gradient principal est prévisible par régression multiple sur les variables de milieu. Cette combinaison de la régression et de l'analyse des correspondances est caractéristique de l'analyse canonique des correspondances, qui sera notée ACC.

L'ACC entretient également des relations étroites avec l'analyse discriminante dont elle est une généralisation. C'est pourquoi nous discuterons aussi de cette méthode.

## 2 — Analyse discriminante

L'analyse canonique des correspondances est une extension de l'analyse discriminante en ce sens qu'elle fabrique des combinaisons linéaires de variance unité maximisant une certaine quantité. Elle pose alors des problèmes d'interprétation très voisins.

Initialiser l'analyse discriminante inter-strates du tableau M dans Discrimin :

Initialize: LinkPrep			
Statistical triplet	<input type="button" value="👉"/>	M.cnta	200 8
Categories file (.cat)	<input type="button" value="👉"/>	Stra.cat	
Selected variable (default=1)	<input type="button" value="👉"/>		
Output file name	<input type="button" value="👉"/>	AFD	

L'inertie totale de 8 (8 variables normées) présentent 76% de variabilité inter-classes. Le test sur la trace est évidemment extrêmement significatif :

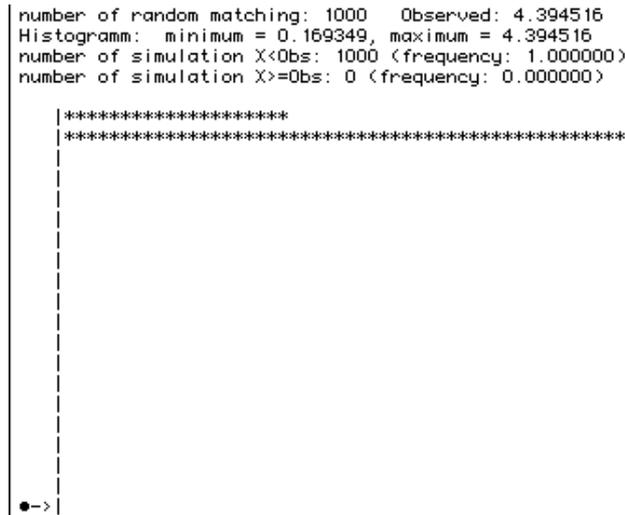


Figure 10 : Distribution de fréquence des valeurs simulés de la trace de l'analyse discriminante sur un échantillon de 1000 permutations des 200 lignes.



Il y a manifestation trois axes discriminants :

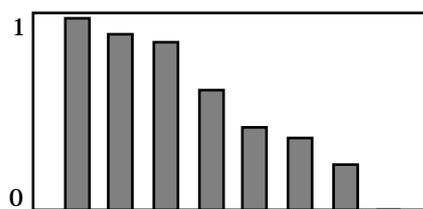


Figure 11 : Valeurs propres (rapports de corrélation) de l'analyse discriminante.

```

Discriminant analysis
Categories defined by column 1 of file Stra
Input statistical triplet: table M.cnta
Number of rows: 200, columns: 8
total inertia (norm C- generalised inverse) = rank of the data matrix:
8.000000

```

between-class inertia (norm C-): 4.394516 (ratio: 0.549314)

Num.	Eigenval.	R. Iner.	R. Sum	Num.	Eigenval.	R. Iner.	R. Sum
01	+9.7686E-01	+0.2223	+0.2223	02	+8.8736E-01	+0.2019	+0.4242
03	+8.5737E-01	+0.1951	+0.6193	04	+6.1439E-01	+0.1398	+0.7591
05	+4.2853E-01	+0.0975	+0.8566	06	+3.7489E-01	+0.0853	+0.9419
07	+2.3929E-01	+0.0545	+0.9964	08	+1.5833E-02	+0.0036	+1.0000

File AFD.divp contains the eigenvalues and relative inertia for each axis

It has 8 rows and 2 columns

File AFD.difa contains coefficient of discriminant scores

It has 8 rows and 3 columns

File :AFD.difa

```

----- Mi ni mum/Maxi mum:
Col.: 1  Mi ni = -0.43038  Maxi = 0.1952
Col.: 2  Mi ni = -0.87052  Maxi = 0.86784
Col.: 3  Mi ni = -0.91253  Maxi = 1.4656

```

File AFD.dili contains canonical row scores with unit norm  
It has 200 rows and 3 columns

File : AFD.dili

----- Minimum/Maximum:  
Col.: 1 Mini = -2.0949 Maxi = 1.3962  
Col.: 2 Mini = -2.2004 Maxi = 1.6722  
Col.: 3 Mini = -3.2069 Maxi = 2.8186

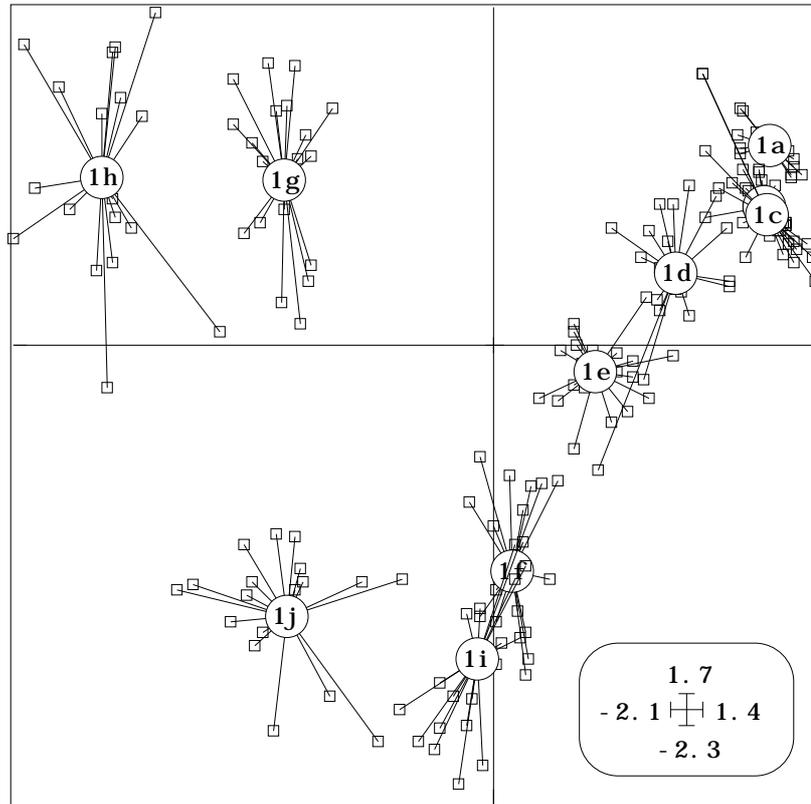


Figure 12 : plan 1-2 de l'analyse discriminante. Les 200 relevés sont positionnés par des codes de variance 1 optimisant la variance inter-groupes. Code des strates : 1a— Maquis bas, 1b— Maquis moyen, 1c— Maquis haut, 1d— Maquis à strate arborée claire, 1e— Maquis à strate arborée dense, 1f— Subéraie à sous-bois, 1g— Subéraie sans sous-bois, 1h— Zénaie, 1i— Pinéraie à pin maritime, 1j— Eucalyptaie

File AFD.di ax contains the principal axes  
It has 8 rows and 3 columns

File : AFD. di ax

----- Minimum/Maximum:  
Col.: 1 Mini = -0.93833 Maxi = 0.75752  
Col.: 2 Mini = -0.57598 Maxi = 0.56103  
Col.: 3 Mini = -0.24801 Maxi = 0.47535

File AFD.dicp contains the correlations between PCA scores  
and DA scores. It has 8 rows and 3 columns

File : AFD. dicp

----- Minimum/Maximum:  
Col.: 1 Mini = -0.95626 Maxi = 0.20706  
Col.: 2 Mini = -0.45436 Maxi = 0.67492  
Col.: 3 Mini = -0.51025 Maxi = 0.30668

L'analyse discriminante donne des combinaisons linéaires de variables de variance unité. Ces combinaisons sont destinées à optimiser la variance inter-classe sous la contrainte que la variance totale soit égale à 1. Ce que signifie ces combinaisons

s'exprime par les cartes factorielles (ScatterClass sur D.d\*li et Stra : figure 12). Le facteur 3 sépare les strates 7 et 8.

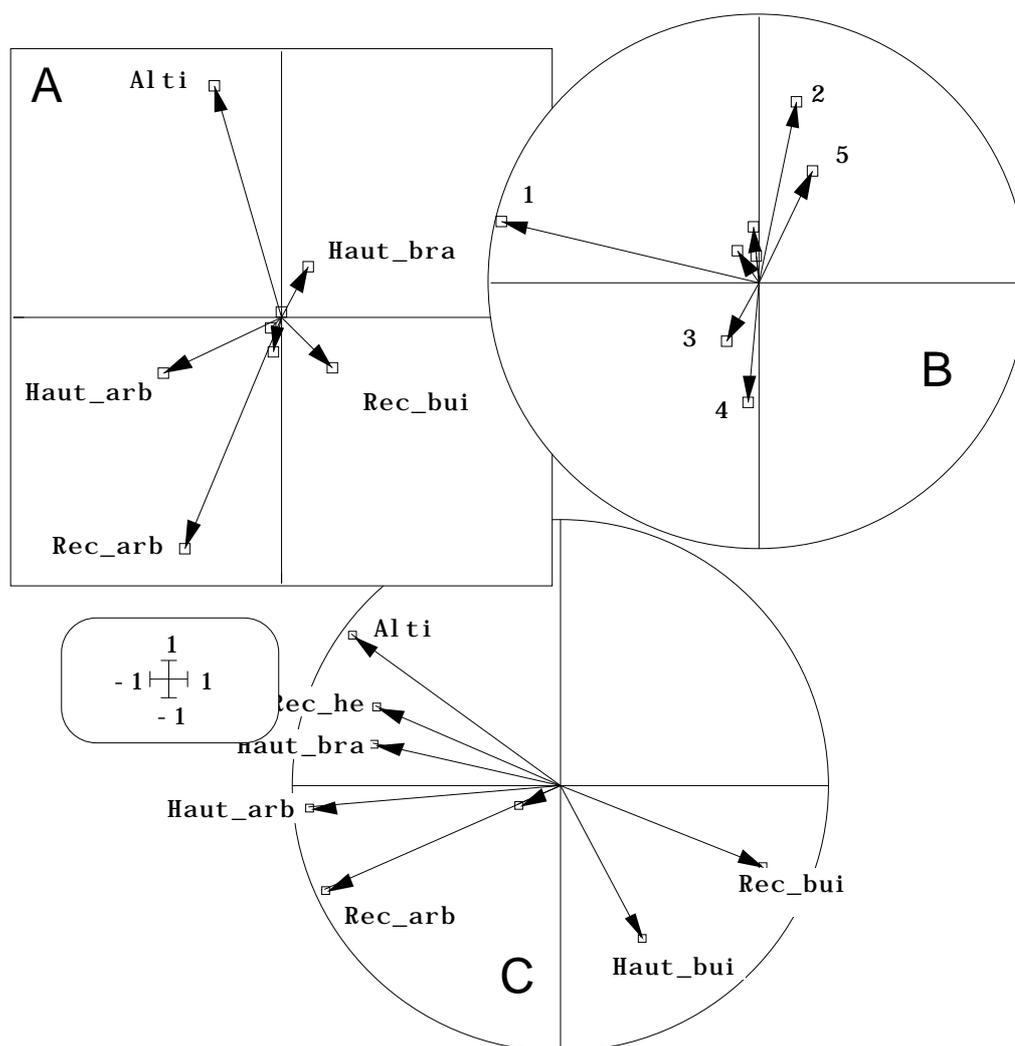


Figure 13 : dépouillement de l'analyse discriminante de D. A — Poids canoniques des variables. B — Corrélations codes canoniques/coordonnées ACP. C — Corrélations codes canoniques/variables.

La lecture d'un plan d'analyse discriminante est donc sans difficulté puisqu'il y apparaît explicitement l'objectif (séparer les groupes d'individus par des combinaisons de variables). Moins évidente est la lecture de la participation des variables à la constitution de ces fonctions discriminantes. Cette lecture fait l'objet de nombreux articles dont celui de Rencher<sup>12</sup>. On peut prendre les coefficients de ces combinaisons (poids canoniques, canonical weights). Dans l'analyse discriminantes ces poids sont dans le fichier AFD.difa.

Représenter les poids par Scatters sur les colonnes 1 et 2 de AFD.difa (figure 13A). On peut prendre les corrélations entre les variables et ces combinaisons (appelées encore codes canoniques) : ces coefficients de corrélation sont dans AFD.diap. Les représenter par Scatters (figure 13C). On peut enfin utiliser les corrélations entre les codes canoniques de l'analyse discriminante et les coordonnées factorielles de l'ACP du tableau (fichier AFD.dicp, figure 13B).

La logique des trois composantes de la figure 13 est disparate. En B on voit que le premier code canonique est très corrélé avec la première coordonnée de l'ACP. Cela est prévisible : dans un plan d'échantillonnage stratifié, on peut s'attendre à ce que

l'essentiel de la variabilité entre points soit induite par la répartition de ces points entre les strates d'échantillonnage. Par contre les différences observée entre A et C est moins simple. Il n'y a pas de coefficients élevés dans la constitution de l'axe 1, alors que les poids de l'altitude (Alti) et du recouvrement des arbres (Rec\_arb) sont grands et de signes opposés.

Il faut alors se souvenir que la seule contrainte qui porte sur ces poids est qu'ils doivent "fabriquer" un code de variance unité. On peut faire une variance 1 avec une combinaison de variables et des petits coefficients si ces variables sont très corrélées, ce qui est le cas :

```

Alti /
+0.65/Haut_Arb/
+0.61/      +0.73/Haut_Bra/
+0.50/      +0.79/      +0.60/Rec_arb/
-0.47/      -0.16/      -0.21/      -0.09/Haut_Bui /
-0.63/      -0.63/      -0.56/      -0.55/      +0.62/Rec_Bui /
+0.03/      +0.21/      +0.02/      +0.10/      -0.06/      -0.12/Haut_her /
+0.73/      +0.67/      +0.39/      +0.40/      -0.26/      -0.45/
+0.38/Rec_Her/

```

Cette matrice de corrélation est obtenue avec MatAlg. Remarquer, en passant, qu'en utilisant AFD.dili, tout en gardant le fichier des poids, on vérifie que les variances des codes canoniques sont unitaires et que les covariances sont nulles.

Les variables Alti, Haut\_arb et Rec\_arb ont des poids négatifs (figure 13A) pour l'axe 1, sont fortement corrélées entre elles et fortement corrélées au code canonique 1 (figure 13C). Il n'y a donc pas de contradictions. La variable Rec\_bui est corrélée positivement aux précédentes, corrélée positivement avec le code 1 et porte un poids positif : aucune difficulté n'intervient.

L'axe 1 est le caractère arboré de la formation et plusieurs types de forêts sont en altitude. toutes ces corrélations conduisent à faire de l'axe 1 de l'ACP l'axe 1 de l'analyse discriminante. Noter alors que le poids positif de la variable Haut-Bra est associé à une corrélation négative avec le code et une corrélation positive avec les variables Alti, Haut\_arb et Rec\_arb. Il y a contradiction entre le poids de cette variable (négatif) et le résultat de sa participation (corrélation positive).

On n'en tiendra donc aucun compte, comme on ignorera la variable Rec—Her qui n'est liée à aucune autre, a des poids nuls et des corrélations négligeables. Pour le code canonique 2, au contraire les poids sont forts pour Alti (positif) et Rec\_arb (négatif). On utilise donc des poids de signes contraires pour deux variables corrélées entre elles et les corrélations obtenues avec le code 2, si elles ne sont pas fortes et ne peuvent l'être, sont cependant du signe des poids.

Le code canonique concerne donc la partition induite par l'altitude dans ce qu'elle est partiellement indépendante de l'architecture de la végétation : il opposerait donc les forêts de montagne et celles de plaines. On le vérifie sans peine sur la figure 14 en utilisant ScatterClass pour les paramètres :

Stars	
XY coordinates file	M 200 8
X-axis column number (default = 1)	4
Y-axis column number (default = 2)	1
Categories file (.cat)	Stra.cat
Variable label file (or #)	
Row weights file (option)	D.fcpl 200 1
Draw points (no = 2)	2

Le plan de l'analyse discriminante (figure 12) est très voisin d'une rotation à 45° de la figure 14, modifiée par l'influence des autres variables. C'est donc bien par l'usage simultané des poids et des corrélations qu'il est possible d'interpréter des codes canoniques. En ce sens l'analyse discriminante est toujours plus difficile à interpréter que l'analyse inter-classes, l'analyse canonique est toujours plus difficile à interpréter que l'analyse inter-batterie, l'analyse canonique des correspondances est toujours plus difficile à interpréter que l'analyse de co-inertie.

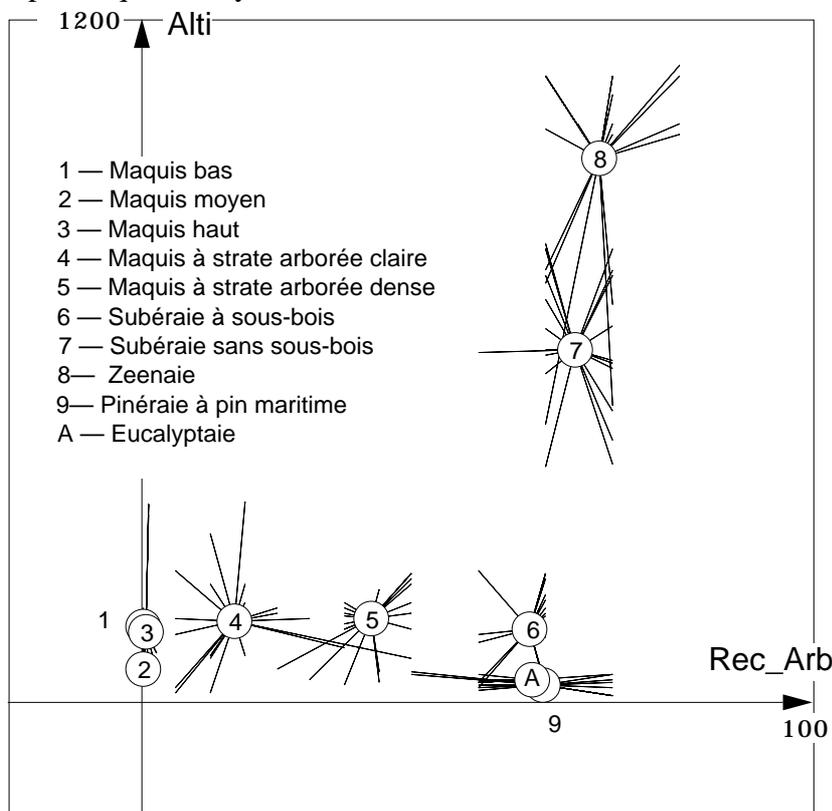


Figure 14 : Données brutes du tableau M pour les variables altitude et recouvrement de la strate arborée.

### 3 — Analyse canonique des correspondances

Dans ADE-4 aucun module, ni aucune option n'est spécifique de l'analyse canonique des correspondances. On considère que celle-ci est un cas particulier, dans la version ACP-AFC d'une méthode universelle de couplages de deux triplets statistiques compatibles, exactement comme ACP, AFC, ACM ... sont des cas particuliers d'une méthode générale d'analyse d'un triplet statistique, exactement comme les analyses inter et intra-classes, les analyses discriminantes et les analyses de co-inertie s'appliquent à une quelconque des approches d'un triplet statistique élémentaire. Ceci permet de générer un très grand nombre de cas particuliers dont certains sont célèbres et d'autres aussi utiles dans certaines circonstances.

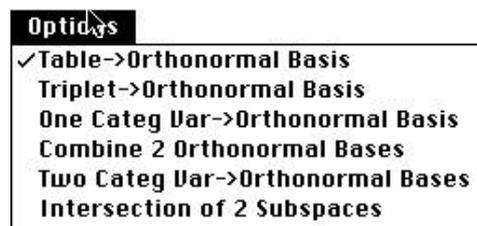
L'ACC a été inventée pour elle-même et fait l'objet d'un logiciel fort connu<sup>13</sup>. L'ACC appartient à la famille des analyses à projecteur<sup>14</sup> sur variables instrumentales et a été présentée comme telle par Sabatier & Coll.<sup>15</sup> et Lebreton & Coll. 1991 (op. cit.). Elle appartient ainsi au logiciel BIOMEKO. L'ACC a donc un fond théorique commun avec l'ACP sur variables instrumentales<sup>16</sup>, l'analyse en composantes explicatives d'Obadia<sup>17</sup>, l'analyse des covariances partielles<sup>18</sup>, l'analyse des redondances de

Wollenberg<sup>1920</sup>, étendue aux variables qualitatives<sup>21</sup>. Les polémiques sur la notion de redondance ne manquent pas<sup>22</sup>.

Le module Projectors est une tentative de regrouper en un minimum d'option un grand nombre de pratiques mettant en œuvre un tableau à analyser et un sous-espace de référence, dans la logique générale de l'ouvrage de Takeuchi & Coll.<sup>23</sup>. Utilisé après une ACP normée et une AFC, il permet l'exécution d'une ACC au sens de Ter Braak auquel on peut préférer la notation d'AFCVI (analyse des correspondances sur variables instrumentales).

### 3.1 — Mise en œuvre

Ouvrir le module Projectors.



La première partie des options définit des sous-espaces par des bases orthonormales. On peut considérer qu'il s'agit d'une transformation initiale des données d'intérêt technique.



```

Explanatory variable file: M.cnta
It has 200 rows and 8 columns
-----
Orthonormal basis: M.@ob
It has 200 rows and 8 columns
Row weighting file: M.@pl (the same as M.cnpl)
Coordinates of the vectors of the orthonormal basis
in the basis of columns of M.cnta in : M.@co
File M.@co has 8 rows and 8 columns
-----

```



```

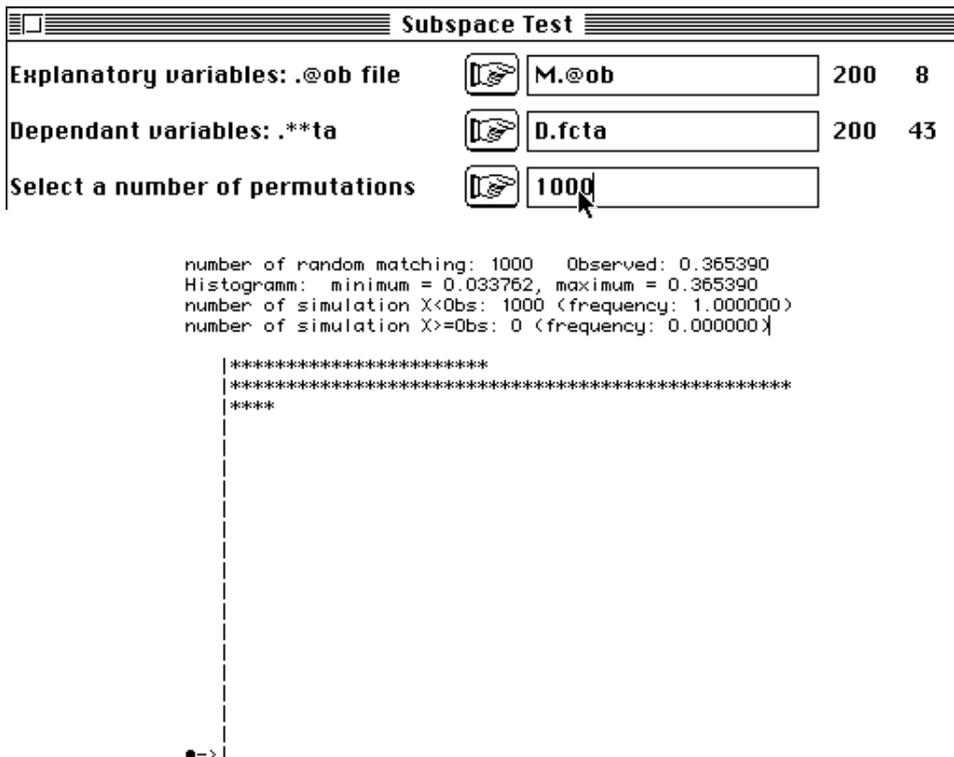
Projected inertia on a subspace
-----
Orthonormal basis: M.@ob
It has 200 rows and 8 columns
Dependant variable file: D.fcta
It has 200 rows and 43 columns
-----

```

Subspace A	A Orthogo	Total	A+%	A-%

1	1. 5161e- 04	1. 1952e- 02	1. 2104e- 02	125	9874
...					
15	8. 1175e- 03	1. 0554e- 02	1. 8672e- 02	4347	5652
...					
18	3. 9765e- 02	4. 1445e- 02	8. 1211e- 02	4896	5103
19	1. 0890e- 03	5. 9652e- 02	6. 0741e- 02	179	9820
20	2. 2740e- 02	3. 1484e- 02	5. 4224e- 02	4193	5806
...					
23	1. 7917e- 01	4. 6029e- 02	2. 2520e- 01	7956	2043
...					
32	9. 1448e- 02	3. 6766e- 02	1. 2821e- 01	7132	2867
...					
43	1. 0122e- 02	6. 2579e- 03	1. 6380e- 02	6179	3820
---	-----	-----	-----	-----	-----
Tot	8. 6120e- 01	1. 4957e+00	2. 3569e+00	3653	6346
---	-----	-----	-----	-----	-----

L'option de décomposition de l'inertie donne pour chaque espèce du tableau faunistique (D.fcta) le pourcentage de variance expliquée par la régression multiple de la colonne correspondante sur l'ensemble des variables de milieu. Ce pourcentage varie de 1% (taxon 1) à 80% (taxon 23). Au total 37% de l'inertie de D.fcta est prévisible par régression multiple et 63% n'est pas modélisable. En fait, ce point de vue, très général, est peu adapté dans la cas présent à cause des courbes de réponses sur le gradient des variables de milieu qui ne peuvent être modélisée par régression linéaire si la courbe est "en cloche" (optimum au centre du gradient).



Le pourcentage d'inertie projetée n'en est pas moins très significatif. L'analyse est justifiée :



Instrumental variables

----- input -----

Orthonormal basis: M.@ob

It has 200 rows and 8 columns

Dependant variable file: D.fcta

It has 200 rows and 43 columns

----- output -----

Projected variable file: A.ivta

It has 200 rows and 43 columns

Inertia: 8.6120e-01

File A.ivpc contains the column weights

It has 43 rows and 1 column

File A.ivpl contains the row weights

It has 200 rows and 1 column

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+4.8126E-01	+0.5588	+0.5588	02	+1.5151E-01	+0.1759	+0.7348
03	+1.2151E-01	+0.1411	+0.8759	04	+3.9462E-02	+0.0458	+0.9217
05	+2.7873E-02	+0.0324	+0.9540	06	+2.2392E-02	+0.0260	+0.9800
07	+1.2586E-02	+0.0146	+0.9947	08	+4.6002E-03	+0.0053	+1.0000

File A.ivvp contains the eigenvalues and relative inertia for each axis

It has 43 rows and 2 columns

File A.ivcl contains column scores (dependant variables) with unit norm

It has 43 rows and 3 columns

File :A.ivcl

----- Minimum/Maximum:  
Col.: 1 Mini = -2.1184 Maxi = 1.8966  
Col.: 2 Mini = -3.4183 Maxi = 1.8989  
Col.: 3 Mini = -3.1928 Maxi = 5.2881

File A.ivco contains column scores (dependant variables) with lambda norm

It has 43 rows and 3 columns

File :A.ivco

----- Minimum/Maximum:  
Col.: 1 Mini = -1.4696 Maxi = 1.3158  
Col.: 2 Mini = -1.3306 Maxi = 0.73914  
Col.: 3 Mini = -1.113 Maxi = 1.8433

File A.ivls contains row scores (dependant variables before projection)

It has 200 rows and 3 columns

File :A.ivls

----- Minimum/Maximum:  
Col.: 1 Mini = -1.3785 Maxi = 1.2462  
Col.: 2 Mini = -1.1542 Maxi = 1.0902  
Col.: 3 Mini = -1.4607 Maxi = 0.97773

File A.ivli contains predicted row scores (dependant variables after projection)

It has 200 rows and 3 columns

File :A.ivli

----- Minimum/Maximum:  
Col.: 1 Mini = -1.3298 Maxi = 1.0658  
Col.: 2 Mini = -0.69953 Maxi = 1.0014  
Col.: 3 Mini = -0.82816 Maxi = 1.0226

File A.ivl1 contains canonical row scores

linear combination of explanatory variables with unit norm  
It has 200 rows and 3 columns

File : A.ivl1

```
----- Minimum/Maximum:
Col.:  1 Mini = -1.917 Maxi = 1.5363
Col.:  2 Mini = -1.7971 Maxi = 2.5727
Col.:  3 Mini = -2.3758 Maxi = 2.9336
```

File A.ivfa contains weights  
coefficients of linear combination of explanatory variables with  
unit norm  
It has 8 rows and 3 columns

File : A.ivfa

```
----- Minimum/Maximum:
Col.:  1 Mini = -0.44835 Maxi = 0.32258
Col.:  2 Mini = -0.46585 Maxi = 0.63876
Col.:  3 Mini = -0.32184 Maxi = 1.0929
```

Il y a trois méthodes d'interprétation des résultats obtenus. L'essentiel du listing est concentré dans le petit tableau :

Num	Variance	R2	Ei g. Val.
1	0.5271	0.9131	0.4813
2	0.2131	0.7111	0.1515
3	0.1948	0.6238	0.1215

### 3.2 — L'ACC comme AFC sous contrainte

Pour lire efficacement le petit tableau qui précède, il convient de se reporter un instant au listing de l'AFC initiale :

```
-----
Total inertia:  2.35693
-----
Num. Eigenval.  R.Iner.  R.Sum  | Num. Eigenval.  R.Iner.  R.Sum  |
01  +5.2929E-01 +0.2246 +0.2246 | 02  +2.1915E-01 +0.0930 +0.3175 |
03  +2.0337E-01 +0.0863 +0.4038 | 04  +1.3688E-01 +0.0581 +0.4619 |
05  +9.8964E-02 +0.0420 +0.5039 | 06  +9.6921E-02 +0.0411 +0.5450 |
```

L'inertie de l'ACC (0.8612) est la part expliquée (36.53%) de l'inertie de l'AFC (2.357). L'AFC décompose cette inertie initiale en valeurs propres qui sont des variances de moyennes conditionnelles par colonnes-espèces sur des codes normalisés de relevés. La première valeur propre (0.5293) est l'optimum possible.

Nous avons vu que l'analyse discriminante cherche des combinaisons de variables de variance unité qui maximise la variance entre les moyennes par strates. **L'ACC partage avec l'analyse discriminante la recherche de combinaisons de variance unité. Elle partage avec l'AFC le but de cette recherche : maximiser la variance entre les positions moyennes des espèces sur ce code.**

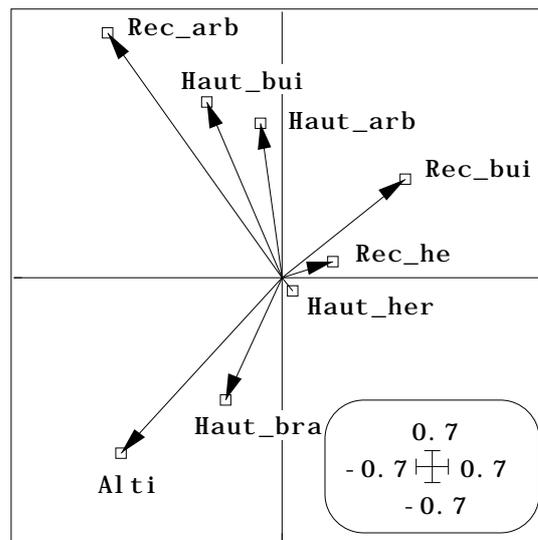
Dans l'AFC ce code est aussi de variance unité pour la pondération des lignes, mais il n'est soumis qu'à cette contrainte. En ACC le code est une combinaison des variables de milieu. Le meilleur code proposé par l'AFC (coordonnée factorielle de variance 1) est donc forcément meilleur que le meilleur code de l'ACC (code canonique) puisque ce dernier subit pour le même objectif une contrainte complémentaire.

Pour une variance de 1, l'AFC propose une variance entre moyennes par espèce de 0.5293. Pour une variance de 1 et une combinaison de variables de milieu l'ACC propose une variance entre moyennes par espèce de 0.4813 (*Eigenvalue* ci-dessus). Pour l'axe 2, 0.2192 est supérieur à 0.1515, et pour l'axe 3, 0.2034 est supérieur à 0.1215. Les valeurs propres d'une ACC ont donc la même signification que les valeurs propres d'une AFC.

La première méthode de dépouillement est donc naturelle : coefficients de variables de milieu appelés encore poids des variables (fichier A.ivfa), valeurs de ces combinaisons qui font une ordination des relevés avec des codes centrés de variance unité (fichier A.iv11), positions moyennes des espèces sur ces codes (fichier A.ivco). Utilisons les deux premières valeurs propres pour mettre en œuvre ce principe.

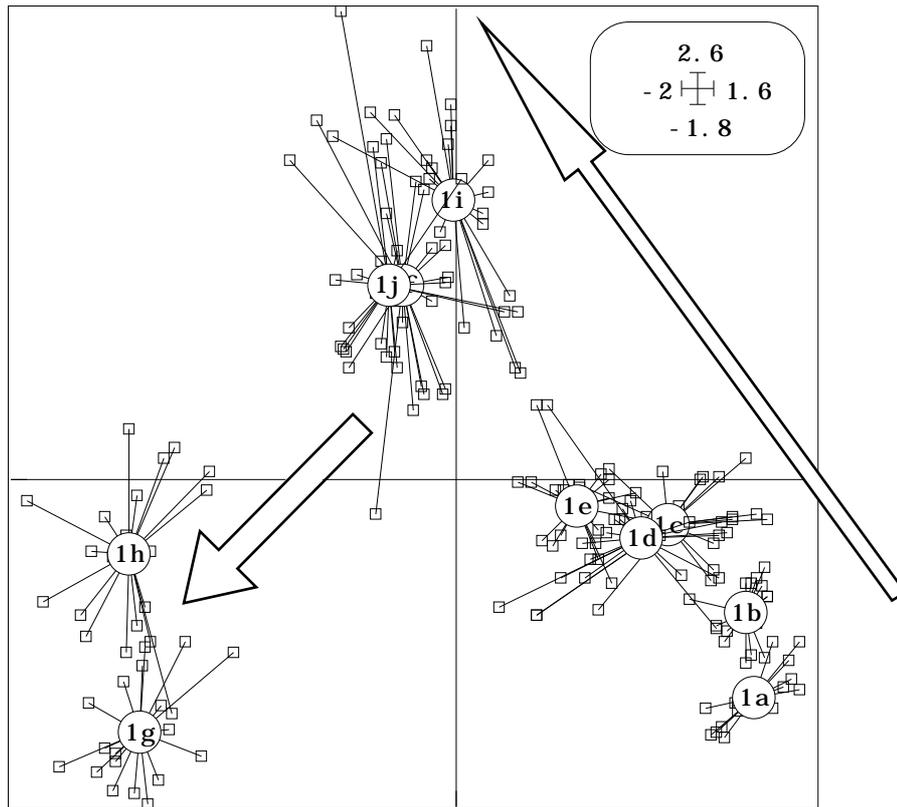
Pour les poids, utiliser Scatters sur A.ivfa :

Labels			
XY coordinates file	<input type="button" value="..."/>	A.ivfa	8 3
X-axis column number (default = 1)	<input type="button" value="..."/>		
Y-axis column number (default = 2)	<input type="button" value="..."/>		
Label file (or # for item numbers)	<input type="button" value="..."/>	Lab_Var	
Draw vectors from origin (yes = 1)	<input type="button" value="..."/>	1	



Pour les codes, on peut utiliser ScatterClass sur A.iv11, afin de repérer comment ils ont pris en compte la répartition en strates :

Stars			
XY coordinates file	<input type="button" value="..."/>	A.iv11	200 3
X-axis column number (default = 1)	<input type="button" value="..."/>		
Y-axis column number (default = 2)	<input type="button" value="..."/>		
Categories file (.cat)	<input type="button" value="..."/>	Stra.cat	
Variable label file (or #)	<input type="button" value="..."/>		
Row weights file (option)	<input type="button" value="..."/>	A.ivpl	200 1



Enfin les espèces prennent position par averaging sur ces codes. Utiliser Scatters sur A.ivl1 et Scatters sur A.ivco, garder les mêmes échelles et superposer (figure 15).

Si la structure générale des nuages est conservée, on note des différences sensibles dans la forme des sous-nuages par classe de milieu. L'AFC (figure 4) peut encore ordonner les relevés dans les strates, tandis que l'ACC ne trouve plus cette possibilité dans le tableau de variables de milieu (il faudrait sans doute posséder des variables qui n'ont un sens que par type de milieu, ce qui est hors des possibilités de ces méthodes).

A droite, on trouve les formations ouvertes (recouvrement des arbres nul), au centre les formations buissonnantes et moyennement arborées, à gauche les forêts. Le groupe des forêts est redivisé par l'altitude (6-9-10/7-8). On a deux gradients orthogonaux, respectivement un gradient d'architecture de la végétation (ouvert-fermé) et un gradient d'altitude. L'axe 1 les met bout-à-bout pour optimiser la variance des moyennes conditionnelles, comme toutes les analyses précitées. L'axe 2 souligne la *supercherie*. L'avifaune restituée presque intégralement la stratification des observations (figure 15).

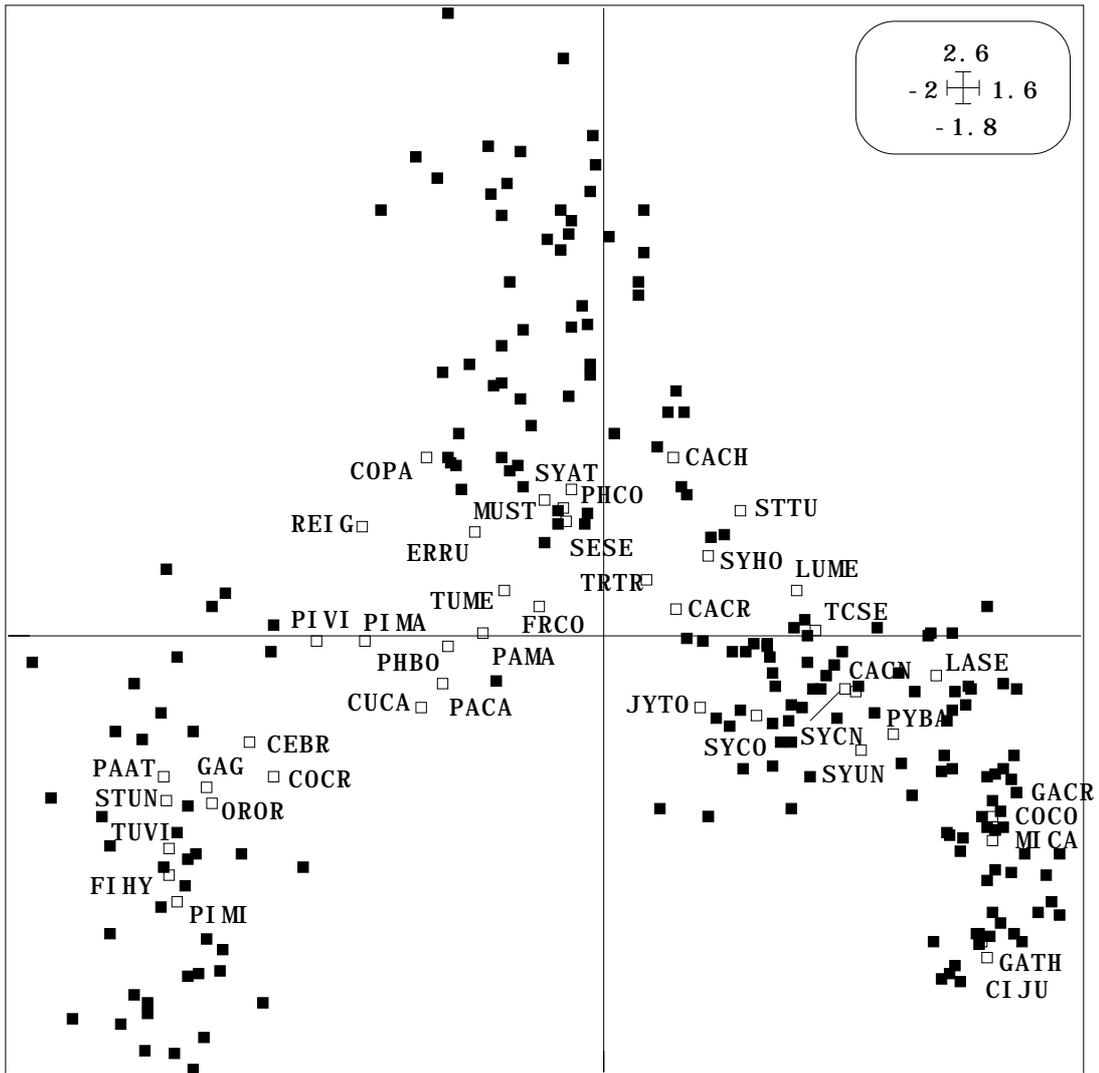
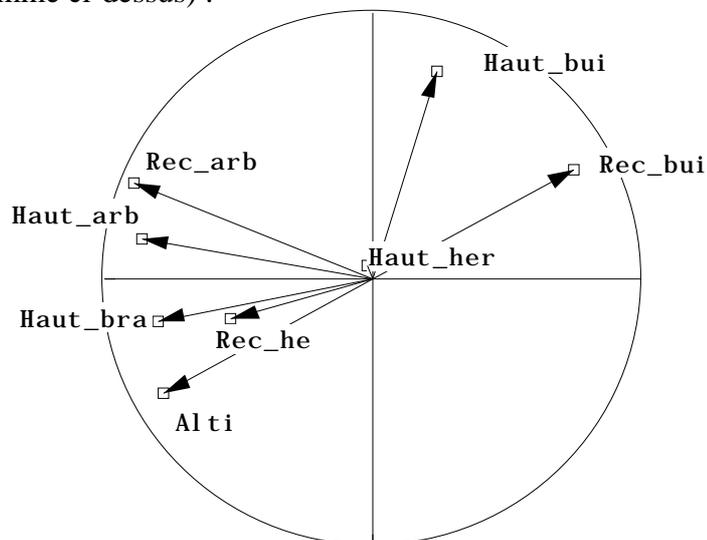


Figure 15 : Plan factoriel 1-2 de l'analyse canonique des correspondances. Carrés noirs : relevés positionnés par des combinaisons linéaires de variables de milieu de variances unité. Carrés blancs : taxons placés par la moyenne de leur distribution sur les 200 relevés.

Pour interpréter le plan du point de vue des variables on peut rajouter les corrélations entre codes canoniques (A.iv11) et variables initiales (utiliser MatAlg, option Diagonal Inner Product comme ci-dessus) :



On remarquera la ressemblance (au signe près, ce qui est sans signification) des structures de corrélation codes/variables entre ACC et AD (analyse discriminante, figure 13). On remarquera aussi que les poids canoniques présentent des nuances sensibles : le cas illustre la possibilité de construire des codes canoniques très voisins avec des poids non identiques. Cette instabilité numérique est caractéristique des méthodes de régression et discrimination. La figure 15 est de toute manière explicite. Les strates sont rangées dans un ordre précis 1/2/3-4-5/6-7-8-9-10/ le long de la première bissectrice du plan.

On retiendra que plus le nombre des variables explicatives (variables de milieu) est grand, plus difficiles d'interprétation seront les codes canoniques (Cf. la fiche régression multiple). **Mais en ACC on ne s'apercevra de rien, car l'ACC tendra dans ce cas à n'être rien d'autre que l'AFC du tableau floro-faunistique.**

### 3.3 — L'ACC comme AFC à coordonnées modélisables

L'interprétation précédente est celle de Ter Braak (1986, 1987, op. cit.). Il est possible de la compléter en utilisant l'averaging double de l'AFC dans l'autre sens.

En effet, nous venons de voir que AFC et ACC partagent le même point de vue de dispersion des moyennes des espèces sur des gradients. L'AFC applique deux fois ce principe puisque, à partir des codes des taxons, on peut placer les relevés par leur moyenne pour optimiser une nouvelle variance conditionnelle.

Ce n'est plus vrai en ACC mais la même opération d'averaging espèces -> relevés garde son sens. Nous avons utilisé les fichiers A.ivl1 (combinaisons de variables de variance unité), A.ivfa (coefficients de ces combinaisons dits poids canoniques) et A.ivco (scores des espèces par leurs moyennes sur les codes canoniques).

Ces scores des espèces sont centrés pour la pondération marginale de l'AFC (D.fcpc). Calculons leurs variances :

Diagonal Inner product C=H'DY			
Input file for X matrix		A.ivco	43 3
Option for X matrix (default=none)			
Input file for Y matrix		A.ivco	43 3
Option for Y matrix (default=none)			
D inner product (default = 1/n)		2	
Option: weighting file		D.fcpc	43 1

ce qui affiche :

```

X input file: A.ivco
--- Number of rows: 43, columns: 3
Y input file: A.ivco
--- Number of rows: 43, columns: 3
Diagonal inner product: D.fcpc
XtDY output file: screen
--- Number of rows: 3, columns: 3
Input file: screen
--- Number of rows: 3, columns: 3
-----
[ 1]  4.8126e-01 -7.6831e-10 -6.4729e-10
[ 2] -7.6831e-10  1.5151e-01  1.1994e-09
[ 3] -6.4729e-10  1.1994e-09  1.2151e-01
-----

```

On vérifie donc que les coordonnées des espèces ont pour variances les valeurs propres et des covariances nulles. On peut normaliser (pour la même pondération) ces

scores des espèces et obtenir ainsi des codes numériques des espèces de moyenne nulle et variance unité. Utiliser pour ce faire Centring dans BibToBin avec :

Centring			
Input file		A.ivco	43 3
Option: file for row weighting		D.fcpc	43 1
Option for H matrix (no default)		3	
Output file		Aux1	

Le fichier Aux1 contient alors des scores des espèces sur lesquels on peut calculer la position moyenne de chaque relevé (averaging espèces -> relevés). Ces scores des relevés seront de moyenne nulle. Leur variance indiquera quelle est la valeur typologique de ces scores.

Between Row Delta dispersion index of Ter BRAAK (1987, equation 3, p. 71)

A est un tableau de nombres positifs (après **fc/COA** ou **fc/ModifCA**)  
 B est un tableau de variables quantitatives  
 R est le tableau des moyennes par lignes de A et par colonne de B  
 Exemple : les lignes de A sont des espèces, les colonnes de B des variables (les colonnes de A et les lignes de B sont des stations).  
 Le programme donne l'indice de dispersion par variables (variance des positions moyennes sur variance initiale)  
 Cet indice est maximisé en analyse canonique des correspondances.  
 Test de permutation sur chacune des variables de B. Cf. théorie des *gradient analysis*

La première variance ne peut dépasser la première valeur propre de l'AFC, qui est la variance maximum qu'on peut produire avec cette opération et une variance de départ unitaire. Consulter la carte de documentation de TestRDelta (ci-dessus, module Basic) et lancer le programme avec les paramètres :

Delta Index			
--.fcta type file (Bin)	D.fcta	200	43
Quantitative variable file (Bin)	aux1	43	3
Output file (creation)	aux1		
Random matching number (0 = no test)	0		

Le tableau central est D (200 relevés-43 espèces). Les codes des espèces sont dans aux1 (43-3). Le fichier créé (Aux1) contient les moyennes par relevé de ces codes espèces et indique que le rapport variance des moyennes / variance initiale vaut respectivement 0.5271, 0.2131, 0.1948. La première valeur est très proche de l'optimum de l'AFC (0.5293). Remarquons alors la modification introduite par rapport à l'AFC.

En AFC, on part d'un code des relevés (coordonnée de variance 1). Par averaging on obtient un code des espèces de variance maximale (0.5293). On peut normaliser le résultat et prendre la coordonnée des espèces de variance 1. Par averaging on obtient un code des relevés de variance maximale (0.5293). La boucle est bouclée si on normalise le résultat et qu'on recommence. On retombe sur les mêmes codes (vecteur propre).

En ACC, on part d'un code des relevés (combinaison de variables de variance 1). Par averaging on obtient un code des espèces de variance maximale sous cette contrainte ( $0.4813 < 0.5293$ ). On peut normaliser le résultat. Par averaging on obtient un code des relevés de variance 0.5271 ( $0.5271 < 0.5293$ ). On n'obtient pas le même résultat.

**La propriété nouvelle est que le code obtenu par averaging n'est pas de variance maximale mais de variance expliquée maximale, variance expliquée par la régression multiple sur les variables de milieu. Pour faire de la variance expliquée, il faut d'abord faire de la variance (0.5271) puis la conserver dans la régression, c'est-à-dire optimiser simultanément le carré de corrélation multiple.**

On peut donc, pour simplifier le discussion en ne s'occupant que du  $R^2$ , normaliser les codes colonnes de auxi1 par Centring :

Centring			
Input file		auxi1	200 3
Option: file for row weighting		D.fcpl	200 1
Option for X matrix (no default)		3	
Output file		auxi2	

Le fichier auxi2 contient des codes de variance unité obtenu par averaging sur des codes des espèces et normalisation. Faisons alors la régression des codes colonnes de auxi2 sur les variables normalisée de M.cnta.

Utiliser LinearReg avec :

Initialize			
Explanatory variables		M.cnta	200 8
Dependent variables		auxi2	200 3
Option: row weighting		D.fcpl	200 1
Output file name		W	

MLR -> Modelling		
Input file		W.reg

```

Multiple Linear Regression
-----
Explanatory variable file: M.cnta
It has 200 rows and 8 columns
Var. | Mean      | Variance |
  1 | 4.237e-09 | 1.000e+00 |
  ...
  8 | 7.953e-09 | 1.000e+00 |
-----
Dependent variable file: auxi2
It has 200 rows and 3 columns
R2 = Squared multiple correlation coefficient
Var. | Mean      | Variance | R2      |
  1 | 6.335e-10 | 1.000e+00 | 9.131e-01 |
  2 | 1.238e-10 | 1.000e+00 | 7.111e-01 |
  3 | 5.191e-11 | 1.000e+00 | 6.238e-01 |
-----
File W.MLRmod has 200 rows and 3 columns

```

```

It contains linear models
-----
File W.MLRres has 200 rows and 3 columns
It contains (data - model) matrix
-----
File W.MLRw1 has 8 rows and 3 columns
It contains regression coefficients

```

Donc les carrés de corrélation multiple ( $R^2$ ) des codes de auxi2, dans la régression sur les variables de milieu, sont respectivement de 0.9131 (excellent), 0.7111 et 0.6238. Ces manipulations sont proposées à titre d'exercices. Les résultats apparaissent directement dans le listing déjà cité :

Num	Variance	R2	Ei g. Val .
1	0.5271	0.9131	0.4813
2	0.2131	0.7111	0.1515
3	0.1948	0.6238	0.1215

En partant d'une combinaison de variable de variance 1, le facteur 1 de l'ACC donne une variance des moyennes par espèces de 0.4813. Cette valeur est forcément inférieure à 0.5293 maximum possible atteint par l'AFC. En partant d'un code des espèces de variance 1, le facteur 1 de l'AFC donne une variance des moyennes par relevé de 0.5293, maximum possible. Le facteur 1 de l'ACC donne une variance expliquée maximale des moyennes par relevés de 0.4813 (valeur propre) qui se décompose en variance simple (0.5271) et pourcentage d'explication (0.9131). La relation qui décompose la valeur propre de l'ACC est donc très simple :

Variance expliquée (0.4113) = Variance (0.5271) x Pourcentage d'explication (0.9131)

L'essentiel des remarques qui précèdent tiennent dans le schéma de la procédure utilisée par Ter Braak dans le logiciel CANOCO (op. cit.). L'algorithme (figure 16) se compose de :

0) *Coder numériquement les échantillons (relevés) avec un code arbitraire non constant centré et réduit.*

1) *Calculer la moyenne conditionnelle par espèce.*

2) *Calculer la moyenne conditionnelle par relevé de ce code espèce*

3) *Faire la régression multiple pondérée (poids marginaux du tableau relevés-espèces) de cette nouvelle variable sur les p variables environnementales. Calculer les valeurs prédites par cette régression.*

4) *Centrer et réduire les valeurs obtenues.*

5) *Réutiliser ce code en 1) et arrêter la boucle lorsque deux itérations successives donnent un résultat assez voisin.*

Le facteur 1 de l'ACC est le code qui redonne lui-même dans un tour complet (vecteur propre).

On peut donc interpréter l'ACC en partant des codes numériques de variance unité des taxons, qui sont dans A.ivc1 :

Labels	
XY coordinates file	A.ivc1 43 3
X-axis column number (default = 1)	<input type="text"/>
Y-axis column number (default = 2)	<input type="text"/>
Label file (or # for item numbers)	Code_Esp



Les positions moyennes des relevés (averaging espèces -> relevés) sont dans A.ivls et les positions prédites par les variables de milieu sont dans A.ivli. Ces deux fichiers sont appariés :

Match two scatters			
HY coordinates file		A.ivls	200 3
H-axis column number (default = 1)			
Y-axis column number (default = 2)			
Second HY coordinates file		A.ivli	200 3

Les trois représentations sont superposées dans la figure 17.

Ceci permet d'affirmer que dans l'optimisation mixte AFC et Régression multiple, la méthode affiche clairement ses limites. Dès qu'elle pourra prévoir une variable quelconque (quand les explicatives seront nombreuses), cette analyse sera simplement une analyse des correspondances. Ce n'est pas le cas ici, avec 8 explicatives pour 200 relevés.

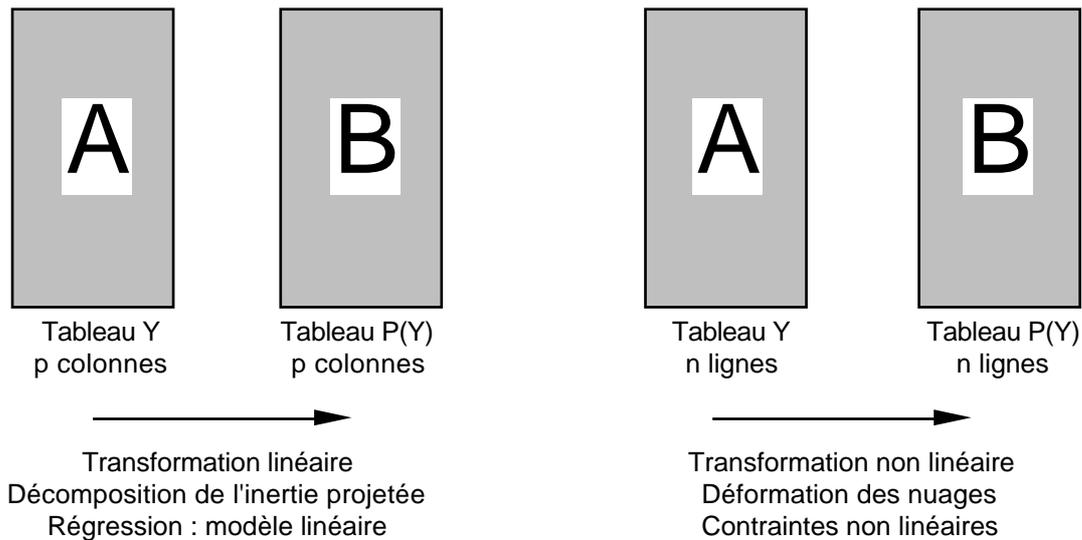
On a vu que deux méthodes de dépouillement sont possibles. La première met en jeu coefficients des variables de milieu (A.ivfa), combinaison de variables de variance 1 (A.ivl1), averaging espèces->relevés (A.ivco). La seconde utilise des codes de variance unité des espèces (A.ivc1), averaging espèces -> relevés (A.ivls), prédiction linéaire par les variables de milieu (A.ivli). La troisième relie les deux première.

### 3.4 — L'ACC comme analyse d'inertie

Les fichiers A.ivli et A.ivco sont les coordonnées lignes et colonnes d'une même analyse d'inertie dont :

- les lignes sont les relevés (200),
- les poids des lignes sont les poids des relevés de l'AFC (A.ivpl est identique à D.fcpl),
- les colonnes sont les espèces (43),
- les poids des colonnes sont les poids des espèces de l'AFC (A.ivpc est identique à D.fcpc),
- l'inertie est l'inertie expliquée par la régression de D.fcta sur les colonnes de M.cnta (0.8612),
- le tableau (A.ivta) est le tableau D.fcta projeté (régression linéaire) colonne par colonne sur le sous-espace défini par les colonnes de M.cnta.

De ce point de vue, l'approche géométrique de l'opération doit être souligné. Le schéma le plus général tient dans la figure :



Y est un tableau d'un schéma de dualité (par exemple D.fcta). Il a n (200) lignes et p (43) colonnes. P est un projecteur de  $\mathbf{R}^n$ . On peut projeter chaque colonne de Y et on obtient un tableau n-p dont les colonnes sont les projetées des colonnes de départ. Sur chacune de ces projections on a un théorème de Pythagore qui décompose l'inertie de départ en inertie projetée et résiduelle. C'est de la modélisation simple, du type régression multiple. En faisant l'analyse de P(Y), comme nuage de colonnes, on cherche des composantes principales de  $\mathbf{R}^n$  sur lesquelles on peut projeter les colonnes de P(Y), ou les colonnes de Y. Cela donne exactement le même résultat (théorème des trois perpendiculaires).

Le même tableau Y est un nuage de lignes, donc de n points de  $\mathbf{R}^p$ . Le même tableau P(Y) est encore un nuage de n points de  $\mathbf{R}^p$ . Mais la fonction qui associe à une ligne de Y la ligne correspondante de P(Y) n'est pas linéaire. C'est la transformation qui en analyse inter-classes envoie un point sur le centre de gravité de sa classe, ce qu'on exprime clairement avec des graphes en étoile. Le nuage déformé des lignes de P(Y) donne des axes principaux sur lesquels on projette les lignes de P(Y) (par exemple les centres des classes) et les lignes de Y (les extrémités des étoiles). On a donc deux systèmes de points projetés. Les lignes de P(Y) sont dans des fichiers ---.ivli (car ce sont les lignes du tableau qui sert à l'analyse) et les lignes de Y sont dans des fichiers ---.ivls (car ce sont des lignes supplémentaires). Travailler avec les lignes projetées permet de comprendre la contrainte qui est introduite par le projecteur. Les graphiques obtenus sont en général simples. Travailler avec les lignes de Y permet de comprendre si cette contrainte simplifie abusivement une réalité plus complexe.

La difficulté et la richesse du système tient en ce que, comme d'habitude, l'analyse de P(Y) vu comme p colonnes et l'analyse de P(Y) vu comme n lignes sont organiquement liées. Pour faire une analyse sur nuage projeté on a deux cas. Le sous-espace de projection est défini par une base orthonormée (fichiers ---@ob) et on utilise l'option **PCA on Instrumental Variables** qui diagonalise dans la plus petite des deux dimensions (nombre de colonnes ou dimension du sous-espace). Le sous-espace est l'orthogonal d'un sous-espace défini par un fichier ---@ob alors on utilise l'option **Orthogonal PCAVI** qui diagonalisera dans la plus petite des dimensions (nombre de lignes ou nombre de colonnes). Ces deux possibilités assure dans la majorité des cas les meilleures conditions numériques.

## 4 — Conclusion

### 4.1 — Bilan méthodologique

Les remarques faites ici n'épuisent pas les propriétés de l'ACC. On a privilégié l'aspect averaging (relevés -> espèces et espèces -> relevés) qui a conduit Ter Braak à cette méthode (1986, 1987, *op.cit.*). Les approches de plusieurs points de vue sont disponibles dans Chessel & Coll. 1987 *op.cit.* et Lebreton & Coll. 1991 *op.cit.*. On retiendra cependant l'essentiel pour l'utilisateur dans le schéma suivant.

Les 3 méthodes donnent des codes numériques des espèces, de variance unité pour les poids de la marge espèces de l'AFC, afin de :

**AFC** : maximiser la variance  $Var$  des moyennes par relevé ;

**ACC** : maximiser la variance expliquée  $Var \cdot R^2$  des moyennes par relevé ;

**Co-inertie** : maximiser la somme des carrés des covariances des moyennes par relevé avec les variables de milieu soit  $Var \cdot \text{somme des carrés des corrélations avec les variables}$ .

L'ACC est donc partiellement une AFC, partiellement une régression multiple et l'analyse de co-inertie est partiellement une AFC, partiellement une ACP.

Les 3 méthodes donnent des codes numériques des relevés afin de maximiser la variance des moyennes par espèce, sous les contraintes :

**AFC** : la variance des codes des relevés vaut 1 pour les poids de la marge relevés de l'AFC ;

**ACC** : la variance des codes des relevés vaut 1 pour les poids de la marge relevés de l'AFC et c'est une combinaison des variables du milieu ;

**Co-inertie** : le code est une combinaison des variables du milieu dont la somme des carrés des coefficients égale 1 (on maximise en partie la variance totale pour maximiser la variance des moyennes).

L'ACC est donc une AFC sous contrainte ou une analyse discriminante généralisée et l'analyse de co-inertie est partiellement une AFC, partiellement une ACP. Cette communauté dans les principes de fonctionnement explique, en partie, la ressemblance entre les résultats obtenus. Il existe des configurations dans lesquelles on pourra noter des différences sérieuses, mais ce n'est pas le cas ici.

### 4.2 — Bilan expérimental

La synthèse des analyses s'exprime simplement. Le milieu intervient sur l'avifaune essentiellement à travers la stratification préalable. L'AFC préliminaire isole brutalement la strate a (pelouses). Les gradients reconstruits ensuite sont formés de cinq groupes (b, maquis bas / c-d-e-f, maquis hauts et maquis à strate arborée / g-j-k, forêts de plaine, h-Subéraie d'altitude, i-Zénaie d'altitude). On recode donc cette nouvelle partition et on édite les moyennes par variables et par espèces (page 36).

			a	b	cdef	gjk	h	i
11	ALAR	Alouette des champs	.86	.	.	.	.	.
12	LUAR	Alouette lulu	.15	.	.	.	.	.
13	ANCA	Pipit rousseline	.40	.	.	.	.	.
14	MOFL	Bergeronnette printanière	1.44	.	.	.	.	.
21	SATO	Traquet pâte	1.06	.	.	.	.	.
1	COCO	Caille des blés	.28	.	.00	.	.	.
9	GACR	Cochevis huppé	1.50	.07	.13	.	.	.
39	MICA	Bruant proyer	3.25	.	.07	.	.	.
24	CIJU	Cisticole des joncs	.5	.1	.	.	.	.
10	GATH	Cochevis de Thekla	.57	.33	.	.	.	.
17	LASE	Pie-grièche à tête rousse	.18	.	.04	.	.	.
15	PYBA	Bulbul gris	.	.12	.64	.	.	.
29	SYUN	Fauvette pitchou	.	4.39	4.09	.	.	.
43	CACN	Linotte mélodieuse	.75	.19	.89	.15	.	.
28	SYCN	Fauvette à lunette	.	20.61	20.16	6.73	.	.
16	TCSE	Téléphone tchagra	.	.08	.49	.13	.	.
22	LUME	Rossignol philomèle	.	.	.83	.43	.	.
27	SYCO	Fauvette grisette	.	.	.25	.	.	.25
3	STTU	Tourterelle des bois	.	.	.30	.37	.	.
26	SYHO	Fauvette orphée	.	.	.23	.21	.	.06
8	JYTO	Torcol	.	.	.14	.	.13	.
42	CACR	Chardonneret	1.36	.13	.75	.76	.53	.
41	CACH	Verdier	.	.	.98	2.21	.38	.
18	TRTR	Troglodyte	.	1.19	4.89	4.97	.	7.60
25	SYAT	Fauvette à tête noire	.	.	1.82	4.94	.	4.55
44	SESE	Serin cini	.	.	1.05	3.19	2.91	.
30	PHCO	Pouillot véloce	.	.	.40	.96	.	1.49
33	MUST	Gobe-mouches gris	.	.	.44	2.25	1.75	.
45	FRCO	Pinson des arbres	.	.16	1.79	3.24	4.70	2.33
20	TUME	Merle noir	.	.	.83	1.61	1.17	2.86
35	PAMA	Mésange charbonnière	.	.13	1.80	2.98	5.46	3.73
23	ERRU	Rouge-gorge	.	.	1.53	4.8	.	10.4
31	PHBO	Pouillot de Bonelli	.	.	.55	1.58	4.25	.5
36	PACA	Mésange bleue	.	.	2.25	2.48	6.94	8.31
2	COPA	Pigeon ramier	.	.	.	.13	.04	.12
4	CUCA	Coucou gris	.	.02	.06	.07	.25	.29
6	PIMA	Pic épeiche	.	.	.12	.72	.85	1.97
32	REIG	Roitelet triple bandeau	.	.	.	2.22	1.90	4.75
5	PIVI	Pic de Levillant	.	.	.01	.08	.1	.48
40	COCR	Gros-bec	.	.	.05	.3	2.6	.5
38	CEBR	Grimpereau des jardins	.	.	.06	.77	4.56	5.25
47	OROR	Loriot d'Europe	.	.	.	.	.33	.43
48	GAGL	Geai des chênes	.	.	.03	.	.50	1.27
7	PIMI	Pic épeichette	.	.	.	.	.47	.42
19	TUVI	Grive draine	.	.	.	.	.68	.60
34	FIHY	Gobe-mouches noir	.	.	.	.	6.26	4.47
46	STUN	Etourneau unicolor	.	.	.	.	.	.62
37	PAAT	Mésange noire	.	.	.	.	2.62	10.71
			a	b	cdef	gjk	h	i
		Al ti	170	132	116	69	637	958
		Haut_arb	0	0	327	1113	1035	1760
		Haut_bra	0	0	121	202	418	363
		Rec_arb	0	0	12	59	65	68
		Haut_bui	24	60	158	205	5	101
		Rec_bui	1	65	70	54	0	22
		Haut_her	24	23	22	24	17	27

## Références

- <sup>1</sup> Benyacoub (S.) 1993 *Ecologie de l'avifaune forestière nicheuse de la région d'El Kala (Nord-Est algérien)*. Thèse de Doctorat, Université de Bourgogne, 250 p., 1993.
- <sup>2</sup> Lebreton, J.D. & Yoccoz, N. (1987) Multivariate analysis of bird count data. *Acta Œcologica, Œcologia Generalis* : 8, 2, 125-144.
- <sup>3</sup> 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.
- <sup>4</sup> Prodon, R. & Lebreton, J.D. (1981) Breeding avifauna of a Mediterranean succession : the holm oak and cork oak series in the eastern Pyrénées. 1 : Analysis and modelling of the structure gradient. *Oikos* : 37, 21-38.
- <sup>5</sup> Lebreton, J.D., Chessel, D., Prodon, R. & Yoccoz, N. (1988) L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Œcologica, Œcologia Generalis* : 9, 1, 53-67.
- <sup>6</sup> Hill, M.O. (1973) Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology* : 61, 237-249.  
  
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.  
  
Chessel, D., Lebreton, J.D. & Prodon, R. (1982) Mesures symétriques d'amplitude d'habitat et de diversité intra-échantillon dans un tableau espèces-relevés: cas d'un gradient simple. *Compte rendu hebdomadaire des séances de l'Académie des sciences . Paris, D* : III, 295, 83-88.
- <sup>7</sup> Thioulouse, J. & Chessel, D. (1992) A method for reciprocal scaling of species tolerance and sample diversity. *Ecology* : 73, 670-680.
- <sup>8</sup> 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.
- <sup>9</sup> Ter Braak, C.J.F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* : 69, 69-77.  
  
Ter Braak, C.J.F. (1987) The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* : 69, 69-77.
- <sup>10</sup> Mercier, P. (1991) *Analyses des relations espèces-environnement et étude de la co-structure d'un couple de tableaux*. Thèse de doctorat, Université Lyon 1. 1-168.
- <sup>11</sup> Ter Braak, C.J.F. (1985) Correspondence analysis of incidence and abundance data : properties in terms of a unimodal response model. *Biometrics* : 41, 859-873.
- <sup>12</sup> Rencher, A.C. (1988) On the use of correlations to interpret canonical functions. *Biometrika* : 75, 363-365.

<sup>13</sup> 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.

Ter Braak, C.J.F. (1987) *Unimodal models to relate species to environment*. Agricultural Mathematics Group, Box 100, NL-6700, AC Wageningen, The Netherlands. 1-152.

<sup>14</sup> Chessel, D., Lebreton, J.D. & Yoccoz, N. (1987) Propriétés de l'analyse canonique des correspondances. Une utilisation en hydrobiologie. *Revue de Statistique Appliquée* : 35, 4, 55-72.

<sup>15</sup> 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.

<sup>16</sup> Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya, A* : 26, 329-359.

<sup>17</sup> Obadia, J. (1978) L'analyse en composantes explicatives. *Revue de Statistique Appliquée* : 24, 4, 5-28.

<sup>18</sup> Lebart, L., Morineau, A. & Fenelon, J.P. (1982) *Traitement des données statistiques. Méthodes et Programmes*. Dunod, 2<sup>o</sup> édition, Paris. 1-518.

<sup>19</sup> Wollenberg, A.L. (1977) Redundancy analysis, an alternative for canonical analysis. *Psychometrika* : 42, 2, 207-219.

<sup>20</sup> Johansson, J.K. (1981) An extension of Wollenberg's redundancy analysis. *Psychometrika* : 46, 93-103.

<sup>21</sup> Israels, A.Z. (1984) Redundancy analysis for qualitative variables. *Psychometrika* : 49, 661-346.

<sup>22</sup> Stewart, D.K & Love, W.A. (1968) A General Canonical Correlation Index. *Psychological Bulletin* : 70, 160-163. Gleason, T.C. (1976) On redundancy in canonical analysis. *Psychological Bulletin* : 83, 1004-1006. Rohlf, F.J. (1977) A note of the measurement of redundancy. *Vegetatio* : 34, 63-64. Muller, K.E. (1981) Relationships between redundancy analysis, canonical correlation, and multivariate regression. *Psychometrika* : 46, 139-142.

<sup>23</sup> Takeuchi, K., Yanai, H. & Mukherjee, B.N. (1982) *The foundations of multivariate analysis. A unified approach by means of projection onto linear subspaces*. John Wiley and Sons, New York. 1-458.