

Analyse Discriminante des Correspondances

Résumé

L'analyse discriminante des correspondances est une extension de l'analyse discriminante aux tableaux d'analyse des correspondances qui possède une propriété de double averaging optimal. On l'utilise pour des tableaux faunistiques à deux niveaux de relevés. On pourrait l'employer pour des tableaux faunistiques à deux niveaux de détermination.

Plan

1 — Tableau de moyennes et tableau projeté	2
1.1 — Cadre général	2
1.2 — Le sous-espace associé à une variable qualitative.....	3
1.3 — Le tableau des moyennes.....	4
1.4 — Discrimination et inter-classes	5
2 — Discrimination et analyse des correspondances	7
2.1 — Les triplets de l'AFC.....	7
2.2 — Propriété principale de l'ADC.....	8
2.3 — Mise en œuvre	10
3 — Précautions d'utilisation	17
Références	19

D. Chessel et J. Thioulouse

1 — Tableau de moyennes et tableau projeté

1.1 — Cadre général

L'Analyse Canonique des Correspondances¹ (ACC) ou Analyse Factorielle des Correspondances sur Variables Instrumentales² (AFCVI) est la version ACP/AFC d'une méthode de couplage de deux tableaux dont la version ACP/ACP est l'ACP sur variables Instrumentales³ (ACPVI) ou encore Analyse en Composantes Explicatives⁴ (ACE) ou encore Analyse des Redondances⁵ (AR).

L'ensemble de ces méthodes et nombre d'autres se résument à une famille de schémas de dualité simple. Soit un tableau \mathbf{X} à n lignes et p colonnes. \mathbf{X} subit une transformation préalable et reçoit deux pondérations respectivement \mathbf{D}_n des lignes et \mathbf{D}_p des colonnes. Le tableau transformé \mathbf{Y} et ces deux pondérations donnent un triplet statistique $(\mathbf{Y}, \mathbf{D}_p, \mathbf{D}_n)$. L'analyse d'un tel triplet est une ordination linéaire (ACP, AFC, ACM, ...) ⁶

Les colonnes de \mathbf{Y} sont des vecteurs de l'espace euclidien \mathbb{R}^n muni du produit scalaire de matrice \mathbf{D}_n dans la base canonique. Dans cet espace on peut considérer un sous-espace A défini indépendamment du triplet $(\mathbf{Y}, \mathbf{D}_p, \mathbf{D}_n)$. Le sous-espace définit de manière univoque le \mathbf{D}_n -projecteur orthogonal sur A , noté \mathbf{P}_A , connu en général par sa matrice dans la base canonique. $\mathbf{P}_A \mathbf{Y}$ est un nouveau tableau à n lignes et p colonnes, dont les colonnes sont les projetés des colonnes de \mathbf{Y} . Une ACPVI est une analyse du type $(\mathbf{P}_A \mathbf{Y}, \mathbf{D}_p, \mathbf{D}_n)$. L'intérêt fondamental de cette opération est que simultanément :

dans \mathbb{R}^n , la transformation qui à une colonne de \mathbf{Y} associe la colonne de même rang de $\mathbf{P}_A \mathbf{Y}$ est une opération linéaire, donc simple à maîtriser ;

dans \mathbb{R}^p , la transformation qui à une ligne de \mathbf{Y} associe la ligne de même rang de $\mathbf{P}_A \mathbf{Y}$ est une opération non linéaire, donc riche en possibilité.

La projection peut, en outre, s'accompagner du changement de la norme dans \mathbb{R}^p . On obtient alors un schéma du type $(\mathbf{P}_A \mathbf{Y}, \mathbf{Q}_p, \mathbf{D}_n)$ où \mathbf{Q}_p est une matrice symétrique, positive et non dégénérée, au moins sur le sous-espace de \mathbb{R}^p engendré par les lignes de $\mathbf{P}_A \mathbf{Y}$.

Quand on passe de \mathbf{X} , tableau de départ, à $(\mathbf{P}_A \mathbf{Y}, \mathbf{Q}_p, \mathbf{D}_n)$ on a donc le choix de la transformation initiale (analyse de base), de la pondération \mathbf{D}_n (imposée ou issue des données), du sous-espace de projection A et de la métrique de \mathbb{R}^p .

Il n'est donc pas étonnant que l'on croit avoir un grand nombre de méthodes alors qu'on a une paramétrisation d'une famille de méthodes. Les analyses inter et intra-classes, les analyses discriminantes et anti-discriminantes, les ACP et les AFC sous contraintes, les analyses canoniques appartiennent à cette famille.

Nous étudions ici le sous-espace défini par une variable qualitative.

1.2 — Le sous-espace associé à une variable qualitative

Prenons l'exemple le plus simple. A est défini par une variable qualitative, c'est-à-dire une partition de l'ensemble des n dimensions de \mathbb{R}^n en g groupes d'effectifs respectifs, n_1, n_2, \dots, n_g . On a évidemment :

$$n = n_1 + n_2 + \dots + n_g$$

On ne change pas la généralité du raisonnement en supposant que les groupes sont formés d'indices successifs (ce qui n'est pas une contrainte réelle dans les programmes). Pour écrire la partition, on utilise les indicatrices des classes, c'est à dire les variables qui prennent la valeur 1 si la ligne i correspond au groupe et 0 sinon :

1	1	0	0	1/9	7/9	-4/9	-3/9
1	1	0	0	1/9	7/9	-4/9	-3/9
2	0	1	0	1/9	-2/9	5/9	-3/9
2	0	1	0	1/9	-2/9	5/9	-3/9
2	0	1	0	1/9	-2/9	5/9	-3/9
2	0	1	0	1/9	-2/9	5/9	-3/9
3	0	0	1	1/9	-2/9	-4/9	6/9
3	0	0	1	1/9	-2/9	-4/9	6/9
3	0	0	1	1/9	-2/9	-4/9	6/9

On peut centrer les indicatrices, et on voit immédiatement que le sous-espace engendré les indicatrices centrées est de dimension $g-1$ (degré de liberté de l'effet inter-classe en analyse de variance).

On notera A_0 le sous-espace engendré par les indicatrices centrées, c'est-à-dire l'ensemble des variables centrées et constantes par classes.

Comme $\mathbf{1}_n$, le vecteur dont toutes les composantes valent 1, est dans le sous-espace A , le projecteur \mathbf{P}_{A_0} s'écrit :

$$\mathbf{P}_{A_0} = \mathbf{P}_A - \mathbf{P}_{\mathbf{1}_n} = \mathbf{A} \left(\mathbf{A}^t \mathbf{D}_n \mathbf{A} \right)^{-1} \mathbf{A}^t \mathbf{D}_n - \mathbf{U}_{nn} \mathbf{D}_n$$

où \mathbf{A} est le tableau des indicatrices des classes et \mathbf{U}_{nn} est la matrice à n lignes et n colonnes dont toutes les valeurs valent 1.

$\mathbf{A}^t \mathbf{D}_n \mathbf{A}$ est une matrice à g lignes et g colonnes, diagonale, dont les éléments non nuls sont les sommes par classes des poids des lignes. Notons $\mathbf{D}_g = \mathbf{A}^t \mathbf{D}_n \mathbf{A}$ (diagonale des poids des classes). On voit immédiatement que le projeté sur A_0 d'un vecteur \mathbf{z} de \mathbb{R}^n dont la composante de rang i est z_i est un vecteur de \mathbb{R}^n dont la composante i est $m\left(\mathbf{z}_{/Cl(i)}\right) - m(\mathbf{z})$, où le premier terme est la moyenne des éléments de la classe dans laquelle on trouve i et le second est la moyenne générale.

Un vecteur projeté sur A_0 est un vecteur centré (moyenne nulle) constant par classe. Par exemple :

$$\begin{array}{cccccc}
& 1 & 0 & 0 & 0.5 & 0 \\
& 1 & 0 & 0 & 1.5 & 0 \\
& 0 & 1 & 0 & 2 & 1.5 \\
& 0 & 1 & 0 & 2 & 1.5 \\
\mathbf{A} = & 0 & 1 & 0 & \mathbf{z} = 3 & m(\mathbf{z}) = 1 \quad \mathbf{P}_{A_0}(\mathbf{z}) = 1.5 \\
& 0 & 1 & 0 & 3 & 1.5 \\
& 0 & 0 & 1 & -2 & -2 \\
& 0 & 0 & 1 & 0 & -2 \\
& 0 & 0 & 1 & -1 & -2
\end{array}$$

Ceci est la principale difficulté pour comprendre pourquoi la projection sur A_0 représente l'opération de moyennage par classe. Il est plus naturel, en effet de manipuler le vecteur dont les composantes sont les moyennes par classes, moyenne associée au poids des classes plutôt que le vecteur associé aux valeurs constantes par classes associées au poids des lignes :

$$\begin{array}{cccc}
& 0 & 1/9 & \\
& 0 & 1/9 & \\
& 1.5 & 1/9 & \\
& 1.5 & 1/9 & 0 \quad 2/9 \\
\mathbf{P}_{A_0}(\mathbf{z}) = & 1.5 & 1/9 & \mathbf{z}_g = 1.5 \quad 4/9 \\
& 1.5 & 1/9 & -2 \quad 3/9 \\
& -2 & 1/9 & \\
& -2 & 1/9 & \\
& -2 & 1/9 &
\end{array}$$

1.3 — Le tableau des moyennes

On note alors que ce dernier vecteur est simplement obtenu par :

$$\mathbf{z}_g = \mathbf{D}_g^{-1} \mathbf{A}^t \mathbf{D}_n \mathbf{P}_{A_0}(\mathbf{z})$$

On passe donc d'un vecteur \mathbf{z} de \mathbb{R}^n à un vecteur \mathbf{z}_g de \mathbb{R}^g par centrage puis réduction aux moyennes par classes. On peut appliquer cette opération à toutes les colonnes du tableau \mathbf{Y} et on obtient :

$$\mathbf{Y}_g = \mathbf{D}_g^{-1} \mathbf{A}^t \mathbf{D}_n \mathbf{P}_{A_0}(\mathbf{Y})$$

\mathbf{Y}_g est le tableau dont les lignes sont les classes, les colonnes celles du tableau de départ et les valeurs les moyennes par classes.

Les lignes du tableau des moyennes par classes sont dans le même espace que les lignes du tableau de départ, soit \mathbb{R}^p . Les colonnes du tableau des moyennes par classes sont dans \mathbb{R}^g , alors que les colonnes du tableau projeté sont dans \mathbb{R}^n . \mathbb{R}^g est muni du produit scalaire $\mathbf{D}_g = \mathbf{A}^t \mathbf{D}_n \mathbf{A}$ alors que \mathbb{R}^n est muni du produit scalaire \mathbf{D}_n . En

choisissant une norme commune dans \mathbb{R}^p , on trouve donc deux triplets, respectivement celui du tableau projeté $(\mathbf{P}_{A_0}\mathbf{Y}, \mathbf{Q}_p, \mathbf{D}_n)$ et celui du tableau des moyennes $(\mathbf{Y}_g, \mathbf{Q}_p, \mathbf{D}_g)$.

Dans le premier, pour trouver les axes principaux, on diagonalise $\mathbf{Y}^t(\mathbf{P}_{A_0})^t\mathbf{D}_n\mathbf{P}_{A_0}\mathbf{Y}\mathbf{Q}_p$ et dans le second pour trouver les axes principaux on diagonalise $\mathbf{Y}_g^t\mathbf{D}_g\mathbf{Y}_g\mathbf{Q}_p$. Pour vérifier qu'il s'agit de la même chose, il suffit de noter que :

$$\mathbf{Y}_g^t\mathbf{D}_g\mathbf{Y}_g = \mathbf{Y}^t(\mathbf{P}_{A_0})^t\mathbf{D}_n\mathbf{A}\mathbf{D}_g^{-1}\mathbf{D}_g\mathbf{D}_g^{-1}\mathbf{A}^t\mathbf{D}_n\mathbf{P}_{A_0}\mathbf{Y} = \mathbf{Y}^t(\mathbf{P}_{A_0})^t\mathbf{D}_n\mathbf{P}_{A_0}\mathbf{P}_{A_0}\mathbf{Y}$$

Donc :

$$\mathbf{Y}_g^t\mathbf{D}_g\mathbf{Y}_g = \mathbf{Y}^t(\mathbf{P}_{A_0})^t\mathbf{D}_n\mathbf{P}_{A_0}\mathbf{Y}$$

car \mathbf{P}_{A_0} est un projecteur.

Et comme c'est un projecteur orthogonal on peut rajouter (pour simplifier les programmes) que :

$$\mathbf{Y}_g^t\mathbf{D}_g\mathbf{Y}_g = \mathbf{Y}^t\mathbf{D}_n\mathbf{P}_{A_0}\mathbf{P}_{A_0}\mathbf{Y} = \mathbf{Y}^t\mathbf{D}_n\mathbf{P}_{A_0}\mathbf{Y}$$

On retiendra donc que l'analyse du tableau de moyennes par classes centrées $(\mathbf{Y}_g, \mathbf{Q}_p, \mathbf{D}_g)$ et celle du tableau projeté $(\mathbf{P}_{A_0}\mathbf{Y}, \mathbf{Q}_p, \mathbf{D}_n)$ ont les mêmes axes principaux et mêmes valeurs propres.

On notera, de plus, que si les colonnes de \mathbf{Y} sont centrées par la transformation initiale on a plus simplement $\mathbf{P}_{A_0}\mathbf{Y} = \mathbf{P}_A\mathbf{Y} = \mathbf{A}(\mathbf{A}^t\mathbf{D}_n\mathbf{A})^{-1}\mathbf{A}^t\mathbf{D}_n\mathbf{Y} = \mathbf{A}\mathbf{D}_g^{-1}\mathbf{A}^t\mathbf{D}_n\mathbf{Y}$

C'est pourquoi nous avons introduit à la pratique des variables instrumentales par le biais des inter et intra-classes ⁷.

1.4 — Discrimination et inter-classes

Le choix du sous-espace associé à une variable qualitative définit l'objectif : maximiser la variance inter-classes. Reste deux paramètres libres, le tableau de départ et le produit scalaire dans \mathbb{R}^p .

Si le tableau de départ est formé de variables quantitatives et qu'on en fait une ACP centrée ou normée, il reste encore à choisir le produit scalaire. Dans le triplet de départ $(\mathbf{Y}, \mathbf{D}_p, \mathbf{D}_n)$, \mathbf{Y} est un tableau centré à n lignes (individus) et p colonnes (variables). La diagonale des poids des lignes est généralement $\mathbf{D}_n = \frac{1}{n}\mathbf{I}_n$ où \mathbf{I}_n est la matrice identité mais l'usage d'une pondération quelconque ne complique pas les résultats. La diagonale des poids des colonnes est toujours $\mathbf{D}_p = \mathbf{I}_p$.

La matrice des covariances totales (corrélations si les variables sont normées) est notée :

$$\mathbf{T} = \mathbf{Y}^t \mathbf{D}_n \mathbf{Y}$$

Elle est de même rang que \mathbf{Y} et en général de rang plein car il y a plus de lignes que de colonnes. Elle est donc inversible et $\mathbf{T}^{-1} = (\mathbf{Y}^t \mathbf{D}_n \mathbf{Y})^{-1}$ est un produit scalaire de \mathbb{R}^p . Certains auteurs l'appellent norme de Mahalanobis (8 p. 487, 9 p.258) bien que initialement cette dernière est définie à l'origine de manière différente. $\mathbf{P}_A \mathbf{Y}$ est un tableau n lignes (individus) et p colonnes (variables). Il a une matrice de covariances :

$$\mathbf{B} = (\mathbf{P}_A \mathbf{Y})^t \mathbf{D}_n \mathbf{P}_A \mathbf{Y} = \mathbf{Y}^t \mathbf{D}_n \mathbf{P}_A \mathbf{Y}$$

\mathbf{B} (\mathbf{B} pour between) désigne les covariances inter-classes calculées sur les variables des moyennes par classe. On lui associe immédiatement la matrice des covariances des variables recentrées par classes :

$$\mathbf{W} = (\mathbf{Y} - \mathbf{P}_A \mathbf{Y})^t \mathbf{D}_n (\mathbf{Y} - \mathbf{P}_A \mathbf{Y}) = \mathbf{T} - \mathbf{B}$$

\mathbf{W} (\mathbf{W} pour within) désigne les covariances intra-classes. A l'origine ¹⁰, la métrique de Mahalanobis est définie par \mathbf{W}^{-1} . On peut donc considérer les triplets $(\mathbf{P}_A \mathbf{Y}, \mathbf{I}_p, \mathbf{D}_n)$, $(\mathbf{P}_A \mathbf{Y}, \mathbf{T}^{-1}, \mathbf{D}_n)$ et $(\mathbf{P}_A \mathbf{Y}, \mathbf{W}^{-1}, \mathbf{D}_n)$. Le premier définit ses axes principaux comme vecteurs propres \mathbf{I}_p -normés de la matrice \mathbf{B} . Le second définit ses axes principaux comme vecteurs propres \mathbf{T}^{-1} -normés de la matrice $\mathbf{B}\mathbf{T}^{-1}$. Le troisième définit ses axes principaux comme vecteurs propres \mathbf{W}^{-1} -normés de la matrice $\mathbf{B}\mathbf{W}^{-1}$.

En fait, les deux derniers ont les mêmes vecteurs propres. En effet :

$$\mathbf{B}\mathbf{T}^{-1}\mathbf{u} = \lambda\mathbf{u} \quad (\mathbf{B} + \mathbf{T} - \mathbf{T})\mathbf{T}^{-1}\mathbf{u} = \lambda\mathbf{u} \quad \mathbf{W}\mathbf{T}^{-1}\mathbf{u} = (1 - \lambda)\mathbf{u}$$

$$\mathbf{u} = \frac{1}{1 - \lambda} \mathbf{W}\mathbf{T}^{-1}\mathbf{u} \quad \mathbf{B}\mathbf{W}^{-1}\mathbf{u} = \frac{\lambda}{1 - \lambda} \mathbf{u}$$

Cette propriété permet de relier les tests de Hotelling, Pillai, Wilks et Roy classiques en analyse de variance multivariée (¹¹ p. 157).

Pour simplifier, on ne garde donc qu'une des deux approches équivalentes et on s'en tient aux quatre définitions :

1 — L'ACP inter-classes ou ACP du nuage des centres de gravité est l'analyse du triplet $(\mathbf{P}_A \mathbf{Y}, \mathbf{I}_p, \mathbf{D}_n)$. Elle est présentée dans ⁹ p. 335 comme un cas particulier des analyses lissées.

2 — L'ACP intra-classe est l'analyse du triplet $(\mathbf{Y} - \mathbf{P}_A \mathbf{Y}, \mathbf{I}_p, \mathbf{D}_n)$. Elle est attribuée dans ¹² p. 104 à Thorpe ¹³.

3 — L'analyse discriminante classique est celle du triplet $(\mathbf{P}_A \mathbf{Y}, \mathbf{T}^{-1}, \mathbf{D}_n)$.

4 — L'analyse anti-discriminante¹⁴ est celle du triplet $(\mathbf{Y} - \mathbf{P}_A \mathbf{Y}, \mathbf{T}^{-1}, \mathbf{D}_n)$

On cherche alors à étendre ces définitions au cas de l'AFC.

2 — Discrimination et analyse des correspondances

2.1 — Les triplets de l'AFC

Considérons un tableau \mathbf{X} de nombres positifs ou nuls supportant une analyse des correspondances. L'introduction d'une partition de l'ensemble des lignes en groupes détruit, comme tout apport d'information sur une seule des deux marges, la symétrie naturellement en jeu dans le schéma classique de l'AFC.

Il y a sous-jacent à une analyse des correspondances 4 triplets statistiques significatifs¹⁵. On utilise les notations classiques (⁹ pp. 67-107). \mathbf{X} a n lignes et p colonnes.

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

$$f_{i.} = \sum_{j=1}^p f_{ij} \quad \mathbf{D}_n = \text{Diag}[f_{i.}] \quad f_{.j} = \sum_{i=1}^n f_{ij} \quad \mathbf{D}_p = \text{Diag}[f_{.j}]$$

Les diagonales des poids dérivent des données. Le triplet habituel est (Cf. ¹⁶) :

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

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

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

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

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

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

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

Les valeurs du tableaux sont $f_{ij} - f_{i.} f_{.j}$, les écarts au modèle d'indépendance et les deux métriques sont celles de Mahalanobis associées aux indicatrices des classes (modèle de l'AFC comme analyse canonique explicité dans ¹⁷ et utilisé dans ¹⁸).

C'est la présence implicite de ces quatre triplets qui donne à l'AFC une place centrale en analyse des données. La section 1.3 de ⁹ fait le tour de cette question définitivement. Ces éléments sont également détaillés dans ¹⁹.

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

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

On peut également considérer le triplet $\left(\mathbf{D}_n^{-1} \mathbf{F}, \mathbf{D}_p^{-1}, \mathbf{D}_n \right)$. Le non centrage introduit un vecteur propre artificiel associé à la valeur propre 1 classique dans ²⁰ qui est ensuite éliminée explicitement. Le nuage est dans le sous-espace affine de \mathbb{R}^p défini par l'équation $\sum_{j=1}^p x_j = 1$. Ce point de vue a l'intérêt de simplifier la présentation théorique mais est maintenant abandonné dans les programmes.

Soit alors $\mathbf{P}_{A_0} = \mathbf{P}_A - \mathbf{P}_{\mathbf{1}_n} = \mathbf{A} \left(\mathbf{A}^t \mathbf{D}_n \mathbf{A} \right)^{-1} \mathbf{A}^t \mathbf{D}_n - \mathbf{U}_{nn} \mathbf{D}_n$ le sous-espace de \mathbb{R}^n associé à la partition des lignes. Pour que la projection ait un sens statistique il faut que la dimension de ce sous-espace soit largement inférieure à n . Dans ces conditions, la matrice $\mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F}$ est de rang plein et son inverse symétrique et positive est un produit scalaire de \mathbb{R}^p .

On peut alors définir deux nouvelles analyses. La première est associée au triplet $\left(\mathbf{P}_{A_0} \mathbf{D}_n^{-1} \mathbf{F}, \mathbf{D}_p^{-1}, \mathbf{D}_n \right)$. C'est tout simplement celui de l'AFC du tableau des sommes par classes des lignes du tableau \mathbf{F} , qui est encore celui de l'AFC inter-classes. Il s'associe avec l'analyse du triplet $\left(\mathbf{D}_n^{-1} \mathbf{F} - \mathbf{P}_{A_0} \mathbf{D}_n^{-1} \mathbf{F}, \mathbf{D}_p^{-1}, \mathbf{D}_n \right)$ qui est celui de l'analyse intra-classe²¹. Le couple est largement utilisé en hydrobiologie ²².

La seconde est associée au triplet $\mathbf{P}_{A_0} \mathbf{D}_n^{-1} \mathbf{F}, \left(\mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F} \right)^{-1}, \mathbf{D}_n$ qui est celui de l'analyse discriminante associée à l'analyse des correspondances. On l'appellera analyse discriminante des correspondances.

2.2 — Propriété principale de l'ADC

On utilise ici les propriétés générales des schémas de dualité (Cf. ^{6, 8}). Comme

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

le centrage du nuage des lignes est restauré par la projection dans l'espace des colonnes.

Un axe principal est un vecteur de \mathbb{R}^p \mathbf{a} qui est $\left(\mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F} \right)^{-1}$ -normé.

Son image $\mathbf{b} = \left(\mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F} \right)^{-1} \mathbf{a}$ est une forme linéaire exprimée dans la base duale de la base canonique. Elle est $\left(\mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F} \right)$ -normée. C'est donc un score des colonnes qui a pour propriété :

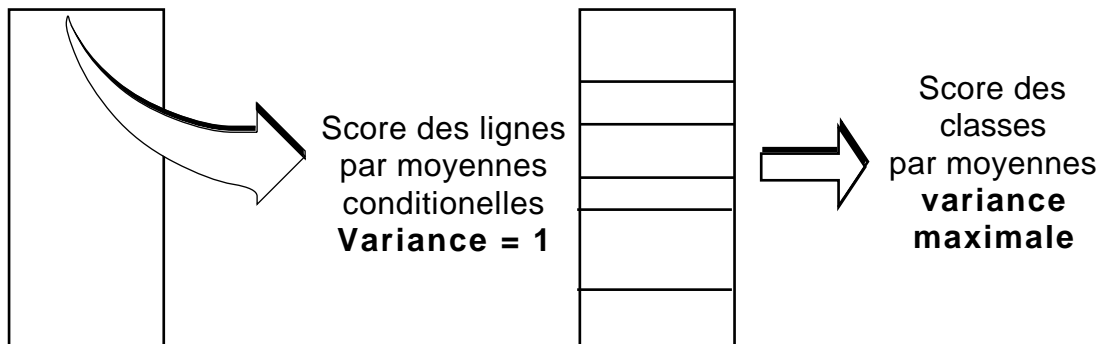
$$\|\mathbf{b}\|_{\mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F}}^2 = \mathbf{b}^t \mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{F} \mathbf{b} = \mathbf{b}^t \mathbf{F}^t \mathbf{D}_n^{-1} \mathbf{D}_n \mathbf{D}_n^{-1} \mathbf{F} \mathbf{b} = \|\mathbf{D}_n^{-1} \mathbf{F} \mathbf{b}\|_{\mathbf{D}_n}^2 = 1$$

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

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

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

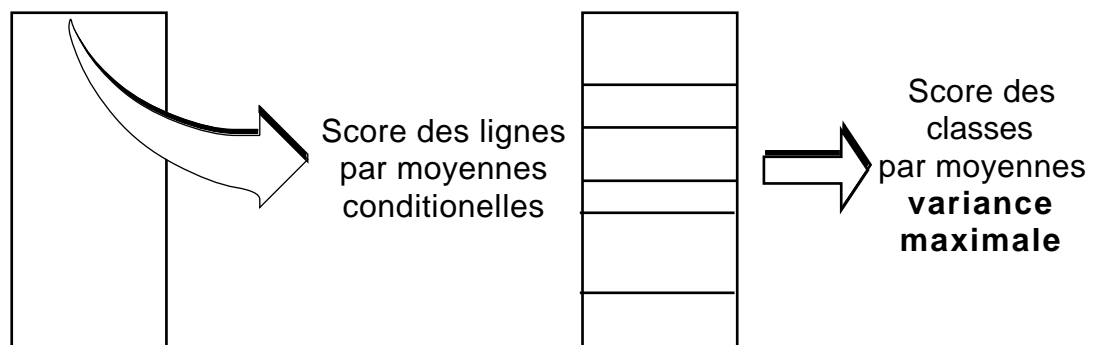
Score des colonnes



Dans le modèle habituel de l'analyse discriminante on a simplement introduit une opération d'averaging dans la constitution des variables canoniques. Il est important de comprendre que dans l'AFC inter-classe on a :

Score des colonnes

Variance = 1



Le principe est voisin mais on a déplacé une contrainte de fonctionnement. Le déplacement de cette contrainte est très importante quant aux résultats. On passe d'une logique d'analyse canonique (**l'analyse discriminante est une analyse canonique**) à une logique de variables instrumentales (**l'analyse inter-classe est une ACPVI**). C'est pourquoi nous pensons que le terme établi d'analyse canonique des correspondances¹ est très malheureux dans l'édifice théorique. L'analyse canonique des correspondances est une AFCVI, c'est-à-dire une ACPVI appliquée à un triplet d'AFC et non une analyse canonique appliquée à un triplet d'AFC. On a ici une analyse discriminante

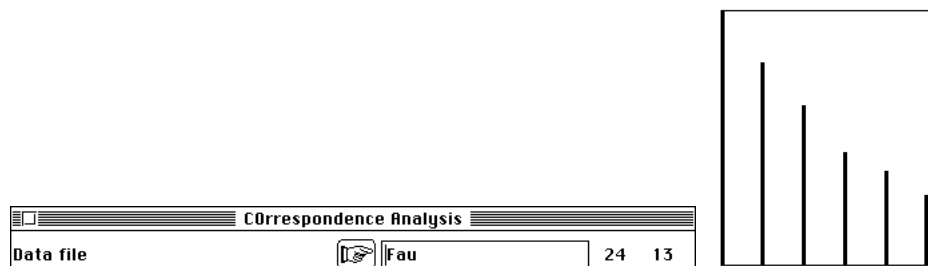
appliquée à un triplet d'AFC qu'on peut appeler analyse discriminante des correspondances qui est différente d'une analyse inter-classe appliquée à un triplet d'AFC, donc une AFC inter-classe.

2.3 — Mise en œuvre

Après cette comparaison des principes théoriques, une comparaison des résultats pratiques semble utile. Nous savons faire facilement une AFC inter-classe. Utiliser la carte Méaudret+2 pour obtenir les fichiers Fau (24-13) et Station (24-1). Lire ce dernier avec CategVar : Read Categ File.

2.3.1 — L'AFC inter-classe

Faire l'AFC du tableau Fau (COA : COrrépondence Analysis) :



Initialiser le lien entre cette analyse et la partition des relevés par dates (Discrimin : Initialize/LinkPrep) :

Initialize: LinkPrep

Statistical triplet 24 13

Categories file (.cat)

Selected variable (default=1)

Output file name

Exécuter l'analyse (Discrimin : Between analysis/Run) :

Between analysis: Run

--.dis input file

```
File FCInterSta.beta contains multivariate means by classes
It has 6 rows and 13 columns
File FCInterSta.bepl contains class weights
It has 6 rows and 1 column
File FCInterSta.bepl contains column weights
It has 13 rows and 1 column
This statistical triplet is the gravity centers one
between-class inertia 0.337235 (ratio: 0.423520)
```

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+1.3912E-01	+0.4125	+0.4125	02	+1.1259E-01	+0.3339	+0.7464
03	+4.1995E-02	+0.1245	+0.8709	04	+2.7070E-02	+0.0803	+0.9512
05	+1.6466E-02	+0.0488	+1.0000	06	+0.0000E+00	+0.0000	+1.0000

Il y a deux axes explicites. Les espèces donnent un score de variance unité :

```
File FCInterSta.becl contains column scores with unit norm
It has 13 rows and 2 columns
File :FCInterSta.becl
|Col.| Mini | Maxi |
|----|-----|-----|
| 1 | -3.001e+00 | 2.718e+00 |
```

2	-1.721e+00	2.414e+00
-----	-----	-----

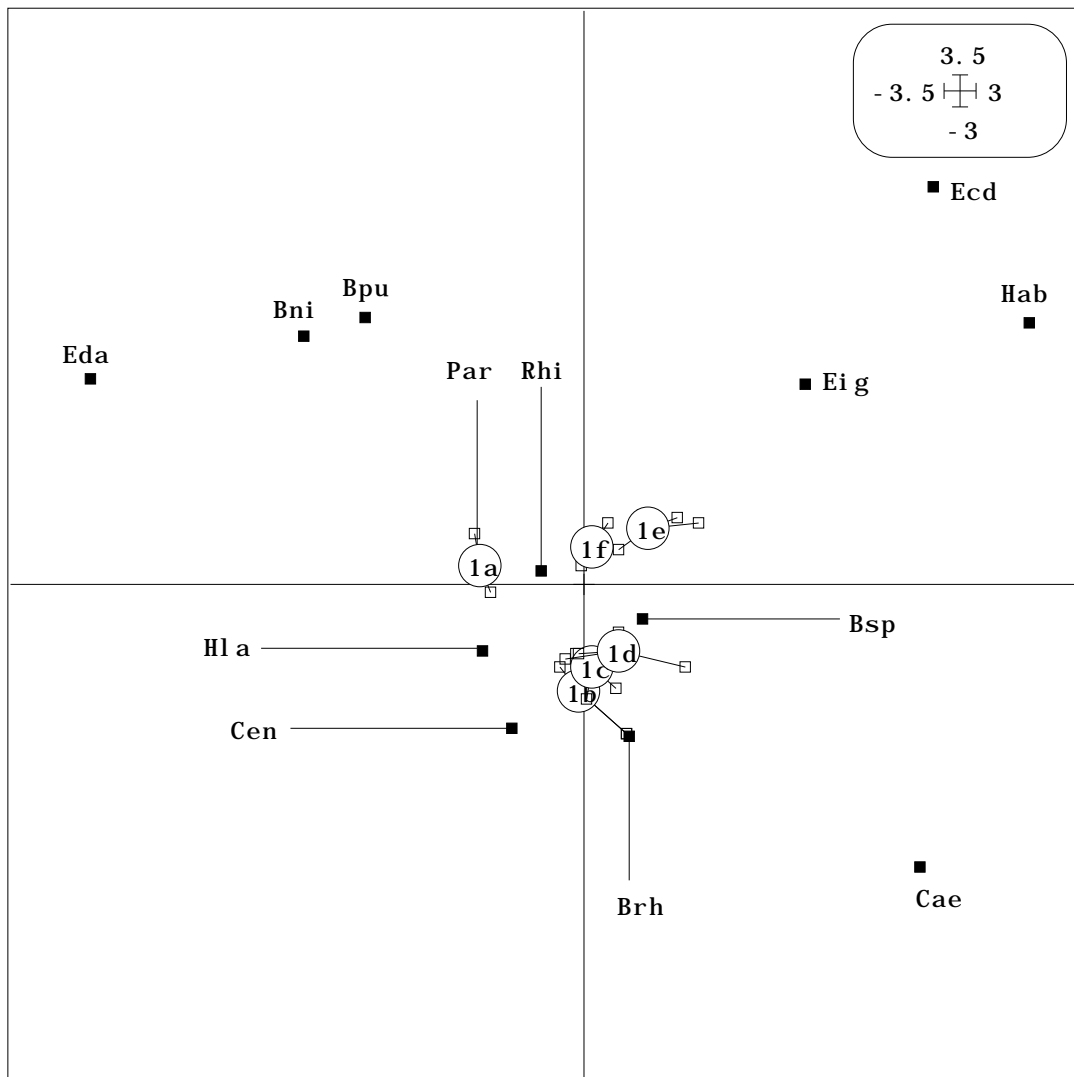


Figure 1 : Principe de fonctionnement de l'AFC inter-classe. Position des espèces par des codes de variance 1 (carrés noirs). Position des relevés à la moyenne de leur profils (carrés blancs). Position des stations à la moyenne des relevés correspondants. On maximise une variance inter-classe, c'est-à-dire le produit d'une variance et d'un pourcentage de variance expliquée.

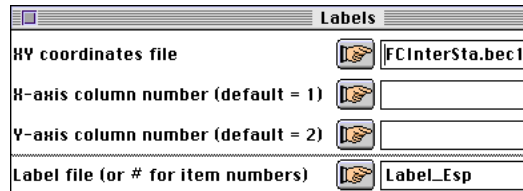
Les relevés sont à la moyenne de ces scores pour la distribution ligne correspondante :

File FCInterSta.bels contains standard row scores with lambda norm
It has 24 rows and 2 columns

```
File :FCInterSta.bels
```

Col.	Mini	Maxi
1	-6.564e-01	7.089e-01
2	-9.216e-01	4.018e-01

Utiliser d'abord Scatters : Labels pour positionner les taxons :



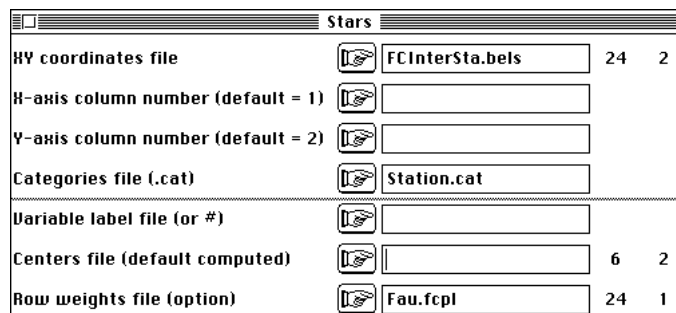
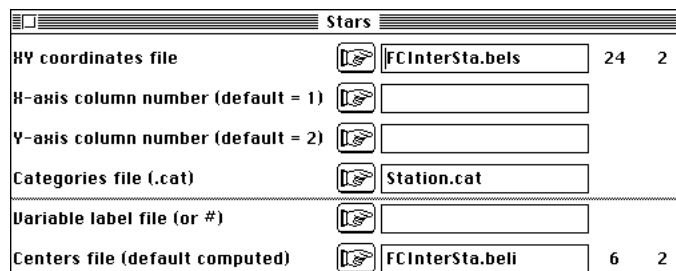
Chaque station est enfin à la moyenne des relevés qui la concerne :

File FCInterSta.beli contains standard gravity center scores with lambda norm
It has 6 rows and 2 columns

File :FCInterSta.beli

Col.	Mini	Maxi
1	-6.223e-01	4.157e-01
2	-6.716e-01	3.289e-01

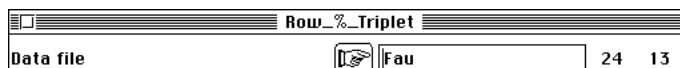
Superposer relevés et stations dans ScatterClass : Stars. Les deux options donnent le même résultat dès qu'on veille à introduire la bonne pondération. Dans la première on utilise les centres calculés dans le module de calcul, dans l'autre on calcule les centres dans le module de dessin :



Superposer enfin les deux graphiques dans la figure 1.

2.3.2 — L'analyse discriminante des correspondances

Pour obtenir les résultats correspondants en analyse discriminante, il n'est pas possible d'utiliser le même triplet de départ. Réinitialiser l'analyse de départ avec COA : Row % Triplet :



Correspondance analysis with row profile triplet
Input file: Fau
Number of rows: 24, columns: 13
File Fau.rcpl contains the edge distribution of rows
It has 24 rows and 1 column

```
File Fau.rcpc contains the Khi2 scalar product DJ-1
It has 13 rows and 1 column
File Fau.rcta contains the centred table DI-1*P -1IJ*DJ
It has 24 rows and 13 columns
```

```
-----
DiagoRC: General program for two diagonal inner product analysis
Input file: Fau.rcta
--- Number of rows: 24, columns: 13
-----
Total inertia: 0.796269
-----
```

L'inertie totale est celle de l'AFC ordinaire obtenue par COA : CO Re spondence Analysis . De même, les valeurs propres sont inchangées :

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+1.9305E-01	+0.2424	+0.2424	02	+1.5343E-01	+0.1927	+0.4351
03	+1.2175E-01	+0.1529	+0.5880	04	+8.5576E-02	+0.1075	+0.6955
...							
11	+1.0398E-02	+0.0131	+0.9960	12	+3.1519E-03	+0.0040	+1.0000
13	+0.0000E+00	+0.0000	+1.0000				

Par contre, les coordonnées des colonnes ne sont pas standard. Elles sont fournies par la diagonalisation générale d'un triplet. Noter ici qu'ADE-4 ne contient qu'une seule procédure d'analyse d'un triplet et est exactement la transcription algorithmique des travaux sur le schéma de dualité dont un des textes fondateurs est ¹⁶. Les coordonnées des colonnes ne sont pas interprétables :

```
File Fau.rcco contains the column scores
--- It has 13 rows and 3 columns
File :Fau.rcco
|Col.|   Mini   |   Maxi   |
|----|-----|-----|
|  1 | -6.705e-02 | 3.832e-02 |
|  2 | -4.324e-02 | 3.466e-02 |
|  3 | -8.215e-02 | 4.917e-02 |
|----|-----|-----|
```

On utilisera par contre, ce qu'on appelle les facteurs dans la théorie du schéma de dualité. L'existence des quatre familles de vecteurs axes principaux, facteurs principaux, composantes principales, cofacteurs principaux est introduite dans ²³: Les facteurs principaux de cette analyse sont les axes principaux de l'AFC ordinaire. Pour l'utilisateur, ce sont les codes des colonnes, centrés et normés (on dit aussi coordonnées de variance unité) qui maximisent la variance des moyennes conditionnelles par lignes (modèle dit de discrimination en AFC, Cf. ⁹ Ch.1.3) introduit en écologie par ²⁴.

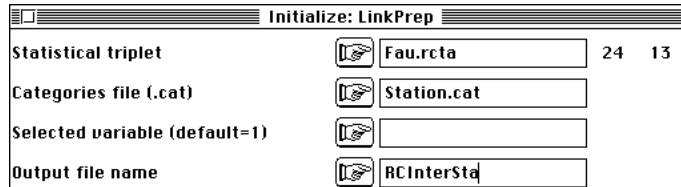
```
File Fau.rccl contains the column scores with unit norm (DJ scalar product)
It has 13 rows and 3 columns
File :Fau.rccl
|Col.|   Mini   |   Maxi   |
|----|-----|-----|
|  1 | -3.602e+00 | 1.578e+00 |
|  2 | -4.881e+00 | 2.091e+00 |
|  3 | -1.683e+00 | 2.382e+00 |
|----|-----|-----|
```

Les coordonnées des lignes sont alors celles de l'analyse ordinaire. On superposera les points représentés par ces deux fichiers pour illustrer le principe de fonctionnement :

```
File Fau.rcli contains the row scores
--- It has 24 rows and 3 columns
File :Fau.rcli
|Col.|   Mini   |   Maxi   |
```

1	-9.223e-01	6.808e-01
2	-1.145e+00	4.918e-01
3	-9.097e-01	4.219e-01

Cette version particulière de l'AFC est couplée avec la variable qualitative Station (Discrimin : Initialize/LinkPrep) :



Noter qu'on retrouve les valeurs d'inertie inter et intra déjà rencontrée :

```
New TEXT file RCInterSta.dis contains the parameters:
input file: Fau.rcta
categorical variable file: Station.cat
n° of categorical variable used: 1
```

```
-----
Between and Within-class inertia
Categories defined by column 1 of file Station
Input statistical triplet: table Fau.rcta
total inertia: 0.796269
between class inertia 0.337235 (ratio: 0.423520)
within class inertia 0.459033 (ratio: 0.576480)
-----
```

Poursuivre par le module d'analyse discriminante :



```
File RCInterSta.dima contains the parameters:
input file: Fau.rcta
categorical variable file: Station
n° of categorical variable used: 1
```

```
Discriminant analysis
Categories defined by column 1 of file Station
Input statistical triplet: table Fau.rcta
Number of rows: 24, columns: 13
total inertia (norm C- generalised inverse) = rank of the data matrix:
12.000000

between-class inertia (norm C-): 3.733622 (ratio: 0.311135)
```

Cette édition ne se comprend que par référence à la théorie. On étudie le schéma de dualité $(\mathbf{P}_A \mathbf{Y}, \mathbf{T}^{-1}, \mathbf{D}_n)$ avec $\mathbf{T}^{-1} = (\mathbf{Y}^t \mathbf{D}_n \mathbf{Y})^{-1}$. La matrice \mathbf{T} est diagonalisable $\mathbf{T} = \mathbf{Y}^t \mathbf{D}_n \mathbf{Y} = \mathbf{U} \mathbf{U}^t$. Donc :

$$\mathbf{T}^{-1} = \mathbf{U} \begin{matrix} -\frac{1}{2} & -\frac{1}{2} \end{matrix} \mathbf{U}^t$$

On en déduit qu'on retrouve dans l'étude de $(\mathbf{P}_A \mathbf{Y}, \mathbf{T}^{-1}, \mathbf{D}_n)$ celle de $\mathbf{P}_A \mathbf{Y} \mathbf{U} \begin{matrix} -\frac{1}{2} \\ -\frac{1}{2} \end{matrix}, \mathbf{I}_p, \mathbf{D}_n$.

L'analyse discriminante est une analyse inter-classe sur le tableau des coordonnées factorielles normalisée du schéma $(\mathbf{Y}, \mathbf{I}_p, \mathbf{D}_n)$. L'inertie totale est égale au rang du tableau \mathbf{Y} , soit ici 12. Les inerties ordinaires sur ces coordonnées sont des inerties sur le tableau utilisant la métrique \mathbf{T}^{-1} . En cas de non-inversibilité, dans ADE-4 on utilise les inverses généralisés de Burnaby²⁵, dont l'usage est explicité par Gower²⁶ qui montre que les inverses généralisées obtenues par diagonalisation sont celles de normes de hilbert-Schmidt minimales (p.3) et que ce n'est pas seulement *of academic interest* (p.5) mais fournit des moyens de calcul simples et efficaces. Cette approche est cohérente avec la conservation des seuls facteurs associés aux variances non nulles dans

$$\mathbf{P}_A \mathbf{Y} \mathbf{U}_f \frac{1}{f^2}, \mathbf{I}_p, \mathbf{D}_n \quad \text{ou la définition de l'inverse généralisée par } \mathbf{T}^- = \mathbf{U}_f \frac{1}{f} \mathbf{U}_f^t.$$

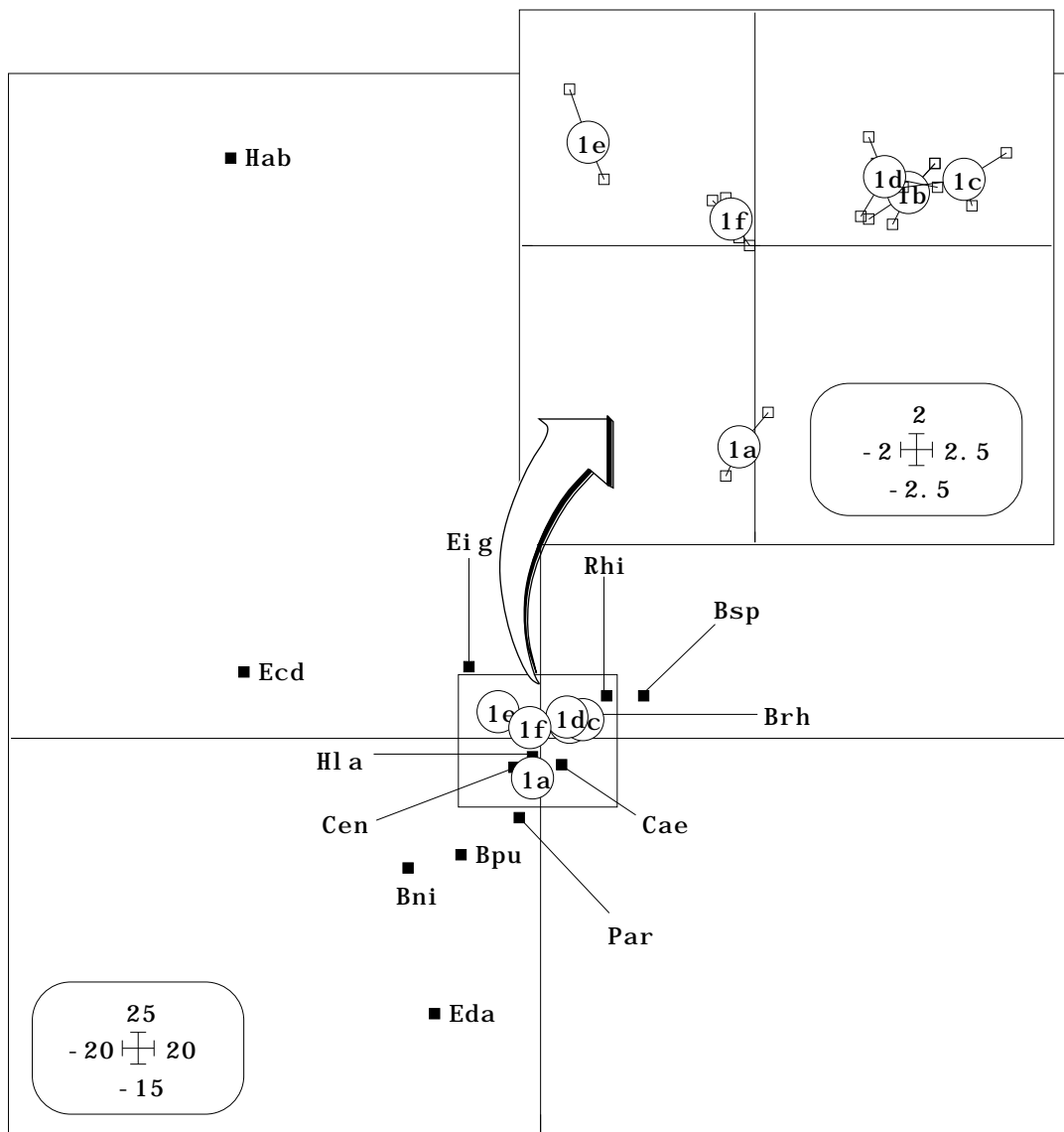


Figure 2 : Principe de fonctionnement de l'Analyse Discriminante des Correspondances. Position des espèces par des codes de variance arbitraires (carrés noirs). Position des relevés à la moyenne de leur profils (carrés blancs) avec une variance unité. Position des stations à la moyenne des relevés correspondants. On maximise une variance inter-classe sur une variance de 1, c'est-à-dire un pourcentage de variance expliquée. La discrimination est parfaite mais il faut un agrandissement pour en voir le principe.

La référence aux inverses généralisées est systématique dans ADE-4 et l'utilisateur n'a pas à se soucier des questions de rang omniprésentes dans les données qualitatives, distributionnelles ou floues.

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+9.7808E-01	+0.2620	+0.2620	02	+9.5522E-01	+0.2558	+0.5178
03	+7.5269E-01	+0.2016	+0.7194	04	+6.2765E-01	+0.1681	+0.8875
05	+4.1997E-01	+0.1125	+1.0000	06	+0.0000E+00	+0.0000	+1.0000

Les valeurs propres sont des inerties inter-classes pour des variables canoniques de variances initiales unité, donc des rapports de corrélation. On observe ici 97.8% et 95.5%, donc des discriminations presque parfaites liées au grand nombre de colonnes.

File RCInterSta.divp contains the eigenvalues and relative inertia for each axis
It has 6 rows and 2 columns

File RCInterSta.difa contains coefficient of discriminant scores
It has 13 rows and 2 columns

File :RCInterSta.difa

Col.	Mini	Maxi
1	-1.167e+01	3.956e+00
2	-1.045e+01	2.181e+01

Les facteurs dans ---.difa sont des codes des espèces donnant par averaging des codes de variance unité qui sont dans la fichier ---.dili. Ces deux fichiers donnent la figure 2.

File RCInterSta.dili contains canonical row scores with unit norm
It has 24 rows and 2 columns

File :RCInterSta.dili

Col.	Mini	Maxi
1	-1.586e+00	2.135e+00
2	-1.946e+00	1.332e+00

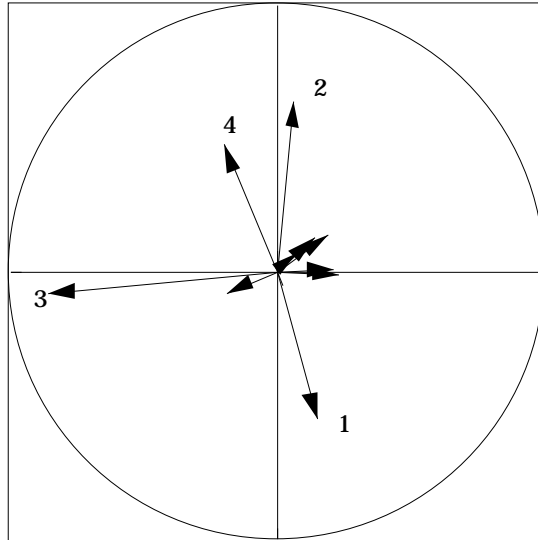
3 — Précautions d'utilisation

Il se pose par rapport à cette analyse les problèmes habituels des stratégies avec inversion de norme ou avec projecteur. Ici on a les deux effets : l'inversion de norme est explicite sur le tableau traité et le projecteur est explicite sur la variable qualitative de discrimination. On a deux problèmes de même nature mais pas de même conséquence.

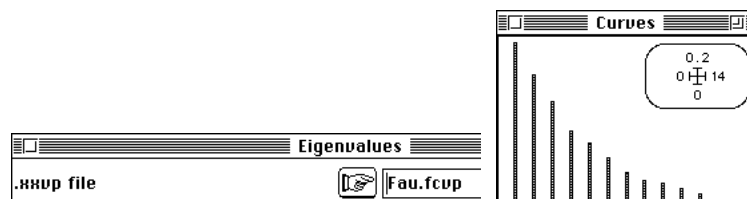
Le projecteur définit le premier problème. Il est associé à la variable de contrôle comme dans une régression. Les variables définissant le sous-espace de projection (les indicatrices des classes) sont orthogonales et en petit nombre (4 donnant une dimension de sous-espace de 3). Il est aisé de vérifier qu'on a une bonne signification statistique (Discrimin : Between analysis/Test) :



L'inter-classe cherche ses variables de synthèse proche des directions d'inertie grande et fait en même temps l'AFC et la discrimination. L'analyse discriminante n'est pas contrainte à cela, mais on s'en bien qu'en utilisant des variables canoniques qui n'ont aucun rapport avec les premières coordonnées de l'analyse de départ on risque de dire n'importe quoi. Il est intéressant de vérifier si les variables canoniques sont corrélées à des facteurs lointains :



Les variables canoniques sont donc une combinaison des quatre premiers axes de l'AFC. Curieusement, la répartition de l'inertie dans l'AFC est ambiguë :



C'est la présence des effets stationnels et temporels mélangés qui rend cette structure compliquée. En utilisant un mélange des quatre premières directions l'ADC ne présente pas de comportement anormal. Dans cet exemple, l'AFC inter-classe s'impose pour des raisons de simplicité et de disponibilité des références. On s'en est servi comme point de repère pour tester une procédure originale qui s'est avéré fonctionner correctement.

Augmenter encore le nombre de colonnes dans l'AFC de départ rendrait cette procédure sans objet. Pour des tableaux comportant beaucoup plus de lignes que de colonnes, elle devrait rendre des services spécifiques. Pour conclure on peut dire que la pratique habituelle de discrimination sur les coordonnées de l'AFC mérite d'être confrontée, en terme de contenu théorique, aux deux versions étudiées dans cette fiche.

Références

¹ Ter_Braak, C.J.F. (1986) Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology* : 69, 69-77.

² 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.

- ³ Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya, A* : 26, 329-359.
- ⁴ Obadia, J. (1978) L'analyse en composantes explicatives. *Revue de Statistique Appliquée* : 24, 4, 5-28.
- ⁵ Wollenberg, A.L. (1977) Redundancy analysis, an alternative for canonical analysis. *Psychometrika* : 42, 2, 207-219.
- ⁶ 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.
- ⁷ Dolédec, S. & Chessel, D. (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatique I- Description d'un plan d'observations complet par projection de variables. *Acta Œcologica, Œcologia Generalis* : 8, 3, 403-426.
- ⁸ Cailliez, F. & Pages, J.P. (1976) *Introduction à l'analyse des données*. SMASH, 9 rue Duban, 75016 Paris. 1-616.
- ⁹ Lebart, L., Morineau, A. & Piron, M. (1995) *Statistique exploratoire multidimensionnelle*. Dunod, Paris. 1-439.
- ¹⁰ Mahalanobis, P.C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* : 12, 49-55.
- ¹¹ Tomassone, R., Danzard, M., Daudin, J.-J. & Masson, J.P. (1988) *Discrimination et classement*. Masson, Paris. 1-173.
- ¹² Yoccoz, N. (1988) *Le rôle du modèle euclidien d'analyse des données en biologie évolutive*. Thèse de doctorat, Université Lyon 1. 1-254.
- ¹³ Thorpe, R.S. (1976) Biometric analysis of geographic variation and racial affinities. *Biological Reviews* : 51, 407-452.
- ¹⁴ Sabatier, R. (1987) *Méthodes factorielles en analyse des données : approximations et prise en compte de variables concomitantes*. Thèse de doctorat d'état. Université de Montpellier. 1-224.
- ¹⁵ Cazes, P., Chessel, D. & Dolédec, S. (1988) L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. *Revue de Statistique Appliquée* : 36, 39-54.
- ¹⁶ Escoufier, Y. (1982) L'analyse des tableaux de contingence simples et multiples. *Metron* : 40, 53-77.
- ¹⁷ Lebart, L., Morineau, L. & Warwick, K.M. (1984) *Multivariate descriptive analysis: correspondence and related techniques for large matrices*. John Wiley and Sons, New York. 1-231.
- ¹⁸ Thioulouse, J. & Chessel, D. (1992) A method for reciprocal scaling of species tolerance and sample diversity. *Ecology* : 73, 670-680.
- ¹⁹ Rouanet, H. & Le Roux, B. (1993) *Analyse des données multidimensionnelles*. Dunod, paris. 1-310.

²⁰ Benzecri, J.P. & Coll. (1973) *L'analyse des données. II L'analyse des correspondances*. Bordas, Paris. 1-620.

²¹ Benzecri, J.P. (1983) Analyse de l'inertie intra-classe par l'analyse d'un tableau de correspondances. *Les Cahiers de l'Analyse des données* : 8, 3, 351-358.

²² Lair, N. & Sargos, D. (1993) A 10 years study at four sites of the middle course of the River Loire. I - Patterns of change in hydrological, physical and chemical variables in relation to algal biomass. *Hydroécologie Appliquée* : 5, 1-27.

Bornette, G., Amoros, C. & Collilieux, G. (1994) Role of seepage supply on the aquatic vegetation dynamics in former river channels: prediction testing using a hydroelectric construction. *Environmental Management* : 18, 223-234.

Bornette, G. & Heiler, G. (1994) Environmental and biological responses of former channels to river incision: a diachronic study on the upper Rhône river. *Regulated rivers: Research and Management* : 9, 72-92.

Bornette, G. & Large, A.R.G. (1995) Groundwater-surface water ecotones at the upstream part of confluences in former river channels. *Hydrobiologia* : 310, 123-137.

Franquet, E., Cellot, B., Pont, D. & Bournaud, M. (1995) Environmental and macroinvertebrate dynamics in the Lower Rhône River and a lateral dike field: a study matching two functioning descriptors. *Hydrobiologia* : 308, 207-217.

²³ Tenenhaus, M. & Young, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika* : 50, 1, 91-119.

²⁴ 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.

²⁵ Burnaby, T.P. (1966) Growth-invariant discriminant functions and generalized distances. *Biometrics* : 22, 96-110.

²⁶ Gower, J.C. (1976) Growth-free canonical variates and generalized inverses. *Bulletin of the Geological Institutions of the University of Uppsala*, N.S : 7, 1-10.