Fiche de Biostatistique

Analyse des correspondances simples

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

Résumé

La fiche regroupe les principales définitions de l'analyse des correspondances. Elles sont repérées par rapport à une procédure de base.

Plan

1. INTRODUCTION 2
2. PROCEDURE DE REFERENCE 4
   2.1. Schéma de base 4
   2.2. Symétrie lignes-colonnes 6
   2.3. Propriétés élémentaires des coordonnées 6
   2.4. Exemple 7
3. CORRELATIONS ENTRE VARIABLES QUALITATIVES 9
   3.1. Exemple 9
   3.2. Corrélation canonique 15
   3.3. Réorganisation de tableaux 16
4. GEOMETRIE DE DEUX NUAGES 18
   4.1. Exemple 18
   4.2. Double analyse d'inertie 18
   4.3. Relations entre cartes factorielles 23
   4.4. Moyennes conditionnelles 23
   4.5. Dilatation 24
5. DOUBLE DISCRIMINATION 26
6. UN SCHEMA DE DUALITE PEUT EN CACHER UN AUTRE 28
7. REFERENCES 30

1. Introduction

Le terme *Analyse Factorielle des Correspondances*, réduit aux initiales AFC, recouvre un ensemble de résultats théoriques, de pratiques statistiques et d'exemples d'utilisation ayant suscité de nombreuses explications de son fonctionnement. Nishisato, dans son ouvrage de référence[1], l'appelle *dual scaling* mais cite (p. 11) les noms de :

— *the method of reciprocal averages*
— *additive scoring*
— *appropriate scoring*
— *canonical scoring*
— *Guttman weighting*
— *principal component analysis of qualitative data*
— *optimal scaling*
— *Hayashi's theory of quantification*
— *simultaneous linear regression*
— *correspondence factor analysis*
— *biplot*

La passionnante analyse bibliographique présentée dans ce livre, qui recouvre largement celle de Buyse [2], montre ses origines lointaines [3], puis les redécouvertes, les enrichissements et les approfondissements successifs. Le processus bibliographique dérive d'une part d'une approche progressive de toutes les facettes d'un même modèle, d'autre part du développement de l'informatique d'abord centralisée, maintenant personnalisée.

Le lien entre la majorité des approches mathématiques se fait clairement dans le schéma de dualité mais l'article de Williams [4], le chapitre 33 (*Categorized data*) de Kendall & Stuart [5] et la communication d'Hathaway (1971)[6] indiquent clairement qu'on connaît la méthode et sa fonction avant la thèse d'Escofier. La diffusion en direction des expérimentateurs est entreprise par Benzécri (1973) [7] et largement connue par l'ouvrage de référence de Greenacre [8]. L'AFC prend le nom d'*homogeneity analysis* dans l'ouvrage de Rijckevorsel [9] qui cite les plus importantes revues sur l'histoire de la méthode et analyse une sélection de citations croisées.

Les extensions, généralisations, utilisations particulières, modalités d'intervention dans chaque discipline, sont tellement nombreuses qu'établir la liste des approches plus ou moins indépendantes n'est plus un objectif raisonnable.

Prenons l'exemple de l'écologie. L'AFC y joue un rôle particulièrement important pour une raison essentielle : l'écologie factorielle, dans son objectif de description de la faune, de la flore, et de leurs relations avec le milieu, s'appuie sur la pratique des relevés et fournit nombre de tableaux dits écologiques. En lignes se trouvent les relevés (placette, prélèvement, piège, sondage, station, point, district, surface, quadrat, segment, échantillon ponctuel, volume d'eau, de sol, d'air,...). En colonnes, se présentent les espèces de la faune ou de la flore étudiée (présence-absence du taxon, effectif des individus, note d'abondance conventionnelle, quantification en pourcentage, en échelle

logarithmique, ...). Les tableaux floro-faunistiques (relevés-taxons) sont analysables par l'AFC (Roux & Roux 1967) : la plupart des milieux et des groupes taxonomiques ont fournit des analyses de ce type. La méthode est particulièrement populaire en phytosociologie [10]. La carte factorielle des espèces et celle des relevés sont les sorties habituellement utilisées.

L'analyse est introduite en hydrobiologie [11], en ornithologie [12], en planctonologie [13]. La représentation des coordonnées factorielles en fonction du temps [14] ou de l'espace [15] introduit en écologie la notion de discrimination par l'AFC. Le modèle d'ordination réciproque est repéré par Hill (1973)[16] et utilisé, par exemple, par Bates & Brown [17] en phytoécologie ou par Prodon & Lebreton (1981)[18] en ornithologie.

Indépendamment, Feoli & Orlóci [19] s'attribuent la procédure sous le nom de *analysis of concentration*, en partant de l'article de Williams [20] qui parle de *analysis of association*, alors que Noy-Meir [21] y voit une analyse en composantes principales doublement standardisée, en partant de l'article de Benzécri (1969) [22]. Des dizaines d'articles utilisent, précisent et commentent la méthode.

Quatre éléments contribuent au succès de la méthode. Le premier a trait à l'énorme diversité des contraintes numériques : un tableau espèces-relevés sera aussi bien constitué de 300 espèces et 15 relevés en forêt dense que de 50 espèces et 300 relevés en steppe aride.

Le second, qui ne lui est pas étranger, concerne la "discrétion" de la méthode en ce qui touche aux notions de variables et d'individus. Typologie d'espèces par les relevés, typologie de relevés par un groupe taxonomique, typologie réciproque sont des objectifs distincts : l'emploi de l'AFC évite, fondamentalement, de se poser la question.

Le troisième est lié à la diversité des modèles justificatifs : parce qu'on peut justifier l'algorithme de multiples façons, parce que ces justificatifs correspondent, même implicitement, à des objectifs précis (l'utilisation sur des tableaux disjonctifs complets, observée comme pertinente, a précédé les théorèmes preuves de cette pertinence), l'AFC est riche de possibilités aptes à restituer la multiplicité des structures observées dans la nature.

La dernière, sans doute décisive, est d'ordre biologique. L'écologie, par principe, utilise comme éléments de base, les correspondances entre individus, entre espèces, entre caractéristiques de leur habitat. Nombre de problèmes écologiques s'expriment *a priori* en termes de correspondances. Citons les premiers mots de l'ouvrage de Guinochet (1973 p.1):

> *"La notion d'association végétale résulte de l'observation suivante : pour quelqu'un qui connaît suffisamment les plantes dans la nature, le simple rappel du nom de l'une d'elles évoque instantanément dans son esprit, non seulement son image, mais encore celle d'un certain nombre d'autres que l'on trouve ordinairement dans les mêmes endroits qu'elle."*

Les exigences écologiques d'une espèce, comme sa valeur indicatrice, recouvrent l'ensemble des correspondances entre la présence d'individus de cette espèce et les

modalités de milieu identifiées aux mêmes places. La structure taxonomique, spatiale ou temporelle, d'une biocénose est exactement l'ensemble des correspondances entre individus de divers taxons, concordances en un lieu ou à une époque, présences simultanées, d'organismes vivants dans les mêmes conditions. L'écologie factorielle échantillonne moins des unités spatio-temporelles que des ensembles de correspondances entre individus, entre espèces, entre modalités d'habitat et entre ces éléments. Ce n'est donc pas la nature formelle des données numériques (tableaux de nombres positifs) qui justifie l'emploi de la méthode pour leur dépouillement, mais la finalité de leur acquisition.

C'est aussi pourquoi l'exécution de l'analyse comporte, pour une part absolument irréductible, l'intervention du langage expérimental proprement dit. En dépit d'une même connaissance des modèles, d'une même maîtrise des organigrammes et d'une même exécution des programmes, le dépouillement des mêmes résultats conduit rarement deux expérimentateurs à une expression identique des structures recherchées: comme partie de l'expérience, l'analyse n'induit pas une solution réglementaire, un résultat qui serait juste à l'exclusion des autres, un résumé qui serait irréductible, exhaustif et indiscutable. Chaque analyse concrète est riche d'une information unique liée à la fois au matériel et à son examen. Un tableau donné ne permet ni d'épuiser une partie des modèles mathématiques sous-jacents ni inversement de se ramener à l'un ou l'autre d'entre eux. Un exemple, quel qu'il soit, oblige soit à réduire soit à dépasser l'expression des fondements de la méthode.

Pour faciliter les comparaisons, il devient alors nécessaire d'appeler AFC une procédure de référence, puis d'expliciter la vocation des résultats obtenus. Toutes les versions de cette procédure ne sont pas identiques, en étant équivalentes. Nous choisirons la présentation d'Y. Escoufier [23], qui est le premier à introduire le double centrage initial explicite, lequel clarifie l'exposé.

2. Procédure de référence

2.1. Schéma de base

On considère un tableau **T** de nombres positifs ou nuls, comportant I lignes et J colonnes. On note n_{ij} son terme générique, $n_{i.}$ et $n_{.j}$ les sommes marginales, n la somme de tous les éléments du tableau :

$$n_{i.} = \sum_{j=1}^{J} n_{ij} \quad n_{.j} = \sum_{i=1}^{I} n_{ij} \quad n = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j}$$

On calcule les fréquences conjointes p_{ij}, les fréquences marginales $p_{i.}$ et $p_{.j}$:

$$p_{ij} = \frac{n_{ij}}{n} \quad p_{i.} = \frac{n_{i.}}{n} \quad p_{.j} = \frac{n_{.j}}{n}$$

On note **P** le tableau des p_{ij} , $\mathbf{D}_I$ et $\mathbf{D}_J$ les matrices diagonales :

$$\mathbf{D}_I = Diag\left(p_{1.},\ldots,p_{I.}\right) \quad \mathbf{D}_J = Diag\left(p_{.1},\ldots,p_{.J}\right)$$

Soit alors **Z** le tableau :

$$\mathbf{Z} = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ} \quad \mathbf{D}_I^{-1} = Diag\left(1/p_{1.},\ldots,1/p_{I.}\right) \quad \mathbf{D}_J^{-1} = Diag\left(1/p_{.1},\ldots,1/p_{.J}\right)$$

Le terme général de **Z** s'écrit simplement :

$$z_{ij} = \frac{p_{ij}}{p_{i.}p_{.j}} - 1 = \frac{p_{ij} - p_{i.}p_{.j}}{p_{i.}p_{.j}}$$

On notera que :

$$\mathbf{D}_I^{1/2} = Diag\left(\sqrt{p_{1.}},\ldots,\sqrt{p_{I.}}\right) \quad \mathbf{D}_I^{-1/2} = Diag\left(1/\sqrt{p_{1.}},\ldots,1/\sqrt{p_{I.}}\right)$$

Par définition, l'AFC du tableau **T** est l'analyse du triplet (**Z**, $\mathbf{D}_J$, $\mathbf{D}_I$) :

$$\begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{Z}^t = \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1} - \mathbf{1}_{JI} \uparrow & & \downarrow \mathbf{Z} = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array}$$

Pour obtenir les éléments propres du schéma, il suffit de suivre la procédure :

— Calcul de $\mathbf{H} = \mathbf{D}_J^{1/2}\mathbf{Z}^t\mathbf{D}_I\mathbf{Z}\mathbf{D}_J^{1/2}$

— Diagonalisation de **H**, matrice symétrique réelle et conservation des *K* premières valeurs propres non nulles dans $\Lambda_K = \mathrm{Diag}\left(\lambda_1,\ldots,\lambda_K\right)$ et des *K* premiers vecteurs propres associés, orthonormés pour la métrique canonique, en colonne dans $\mathbf{U}_K$. $\mathbf{U}_K$ a *J* lignes et *K* colonnes et vérifie $\mathbf{U}_K^t\mathbf{U}_K = \mathbf{I}_K$. En toute généralité, on pourrait rencontrer des valeurs propres multiples. C'est très rarement le cas dans la pratique statistique et on supposera, sauf avis contraire, dans tout ce qui suit, que les espaces propres associés aux valeurs propres non nulles sont de dimension 1.

— Calcul des axes principaux de norme $\sqrt{\lambda_k}$ en colonnes dans la matrice :

$$\tilde{\mathbf{A}}_K = \mathbf{D}_J^{-1/2}\mathbf{U}_K\mathbf{\Lambda}_K^{1/2}$$

Les colonnes de $\tilde{\mathbf{A}}_k$ sont appelées coordonnées des colonnes : la matrice a *J* lignes et *K* colonnes. A la ligne *j* et à la colonne *k* on y trouve la coordonnée de la colonne *j* de rang *k*.

— Calcul des composantes principales de norme $\sqrt{\lambda_k}$ en colonnes dans la matrice :

$$\tilde{\mathbf{C}}_K = \mathbf{Z}\mathbf{D}_J^{1/2}\mathbf{U}_K$$

Les colonnes de $\tilde{\mathbf{C}}_k$ sont appelées coordonnées des lignes : la matrice a *I* lignes et *K* colonnes. A la ligne *i* et à la colonne *k* on y trouve la coordonnée de la ligne *i* de rang *k*.

Ces calculs sont exécutés dans la plupart des programmes d'analyse des correspondances et ils donnent des coordonnées factorielles de normes égales aux valeurs propres. Ils ne préjugent pas de l'emploi qui en sera fait suivant le problème traité.

2.2. Symétrie lignes-colonnes

Si $\lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$ alors l'AFC de $\mathbf{T}$ et l'AFC de $\mathbf{T}^t$ donnent des résultats identiques, à la permutation lignes-colonnes près.

En effet, supposons exécutée l'AFC d'ordre K et exécutons l'AFC de $\mathbf{T}^t$. La permutation laisse inchangée les marges du tableau comme les fréquences. L'AFC de $\mathbf{T}^t$ est la décomposition canonique du schéma ($\mathbf{Z}^t$, $\mathbf{D}_I$ $\mathbf{D}_J$). L'opérateur VQ d'un schéma étant égal à l'opérateur WD de l'autre, les valeurs propres sont conservées et les axes d'une analyse sont les composantes de l'autre et réciproquement. Chaque sous-espace propre étant de dimension 1, l'unicité (au signe près, cependant) du vecteur propre garantit l'identité des deux procédures.

Il s'en suit que la formulation axes-composantes n'a pas grande signification, les axes de l'AFC de $\mathbf{T}$ étant les composantes de l'AFC de $\mathbf{T}^t$. Pratiquement, on diagonalise dans la plus petite des deux dimensions I ou J. On notera par cohérence avec le schéma général $\tilde{\mathbf{A}}_K$ (respectivement $\mathbf{A}_K$) les coordonnées des colonnes de norme $\sqrt{\lambda_k}$ (respectivement 1) et $\tilde{\mathbf{C}}_K$ (respectivement $\mathbf{C}_K$) les coordonnées des lignes de norme $\sqrt{\lambda_k}$ (respectivement 1).

2.3. Propriétés élémentaires des coordonnées

Les coordonnées principales des lignes $\tilde{\mathbf{A}}_K$ de l'AFC de $\mathbf{T}$ sont des variables de $\mathbb{R}^I$ centrées, de variance λ_k, de covariance nulle deux à deux. Les coordonnées principales des lignes $\tilde{\mathbf{C}}_K$ sont des variables de $\mathbb{R}^J$ centrées, de variance λ_k, de covariance nulle deux à deux.

Les vecteurs colonnes $\tilde{\mathbf{A}}_K$ sont propres de $\mathbf{VQ} = \mathbf{Z}^{\mathrm{t}}\mathbf{D}_I\mathbf{Z}\mathbf{D}_J$. Or :

$$\mathbf{Z}\mathbf{D}_J\mathbf{1}_J = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}\begin{bmatrix} p_{.1} \\ \vdots \\ p_{.J} \end{bmatrix} - \mathbf{1}_{IJ}\begin{bmatrix} p_{.1} \\ \vdots \\ p_{.J} \end{bmatrix} = \mathbf{1}_I - \mathbf{1}_I = \mathbf{0}_I$$

$\mathbf{1}_J$ étant dans le noyau, les vecteurs propres associés aux valeurs propres non nulles lui sont orthogonaux au sens de $\mathbf{D}_J$, c'est-à-dire centrés pour la pondération marginale. Les carrés des normes sont donc des variances et les produits scalaires des covariances. La propriété dérive du fait que les axes principaux forment une base orthogonale.

2.4. Exemple

On utilise la table de contingence sur la couleur des yeux et des cheveux chez des Ecossais de Caithness (lignes = couleurs des yeux, colonnes = couleurs des cheveux) [24].

Dans R :

```
> data(caith)
> caith
       fair red medium dark black
blue    326  38    241  110     3
light   688 116    584  188     4
medium  343  84    909  412    26
dark     98  48    403  681    85
```

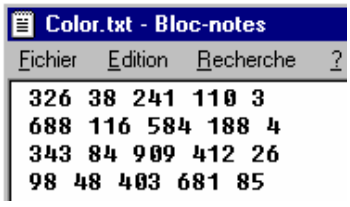
```
> library(MASS)  S original by Venables & Ripley.
  R port by Brian Ripley <ripley@stats.ox.ac.uk>, following earlier
  work by Kurt Hornik and Albrecht Gebhardt.
```

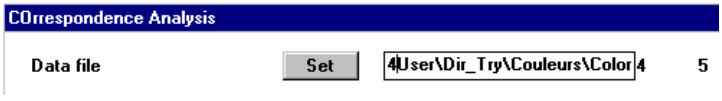
```
> corresp(caith,nf=2)
First canonical correlation(s): 0.4464 0.1735

 Row scores:
          [,1]    [,2]
blue   -0.8968  0.9536
light  -0.9873  0.5100
medium  0.0753 -1.4125
dark    1.5743  0.7720

 Column scores:
           [,1]    [,2]
fair   -1.21871  1.0022
red    -0.52258  0.2783
medium -0.09415 -1.2009
dark    1.31888  0.5993
black   2.45176  1.6514
```

```
> library(multiv)  F. Murtagh (fmurtagh@eso.org), August 1994
```

```
> ca(as.matrix(caith))
$evals
[1] 1.992e-01 3.009e-02 8.595e-04 2.335e-17

$rproj
      Factor1  Factor2   Factor3    Factor4
[1,] -0.40030 -0.16541  0.064158 -2.634e-16
[2,] -0.44071 -0.08846 -0.031773  2.981e-17
[3,]  0.03361  0.24500  0.005553 -8.014e-17
[4,]  0.70274 -0.13391 -0.004345 -1.371e-16

$cproj
      Factor1  Factor2   Factor3    Factor4
[1,] -0.54400 -0.17384  0.012522 -1.383e-08
[2,] -0.23326 -0.04828 -0.118055 -1.605e-08
[3,] -0.04202  0.20830  0.003236 -1.603e-08
[4,]  0.58871 -0.10395  0.010116 -2.175e-08
[5,]  1.09439 -0.28644 -0.046136 -2.547e-08
```

Dans ADE-4 :





```
------------------------------------------------
D:\ADE4USER\DIR_TRY\COULEURS\Color.fcli - 4 rows, 2 cols.
   1 | -0.4003  0.1654
   2 | -0.4407  0.0885
   3 |  0.0336 -0.2450
   4 |  0.7027  0.1339

------------------------------------------------
D:\ADE4USER\DIR_TRY\COULEURS\Color.fcco - 5 rows, 2 cols.
   1 | -0.5440  0.1738
   2 | -0.2333  0.0483
   3 | -0.0420 -0.2083
   4 |  0.5887  0.1040
   5 |  1.0944  0.2864

------------------------------------------------
D:\ADE4USER\DIR_TRY\COULEURS\Color.fcvp - 4 rows, 2 cols.
   1 |  0.1992  0.8656
   2 |  0.0301  0.1307
   3 |  0.0009  0.0037
   4 |  0.0000  0.0000
```

Qu'on se rassure, il s'agit de détails :

```
> cor1 <- corresp(caith,nf=2)
> names(cor1)
[1] "cor"    "rscore" "cscore" "Freq"
> cor1$cor^2
[1] 0.19924 0.03009
> cor1$rscore[,1]*cor1$cor[1]
    blue    light   medium     dark
-0.40030 -0.44071  0.03361  0.70274
> cor1$rscore[,2]*cor1$cor[2]
    blue    light   medium     dark
 0.16541  0.08846 -0.24500  0.13391
> cor1$cscore[,1]*cor1$cor[1]
    fair      red   medium     dark    black
-0.54400 -0.23326 -0.04202  0.58871  1.09439
> cor1$cscore[,2]*cor1$cor[2]
    fair      red   medium     dark    black
 0.17384  0.04828 -0.20830  0.10395  0.28644
```

Le premier programme conserve les racines carrées des valeurs propres, le second conserve les valeurs propres. Le premier conserve les coordonnées normées à 1, le second conserve les coordonnées normées à la valeur propre de même rang. Le rôle de ces calculs est l'objet de ce qui suit. Le débat est ouvert dans la documentation de R :

> *nf: The number of factors to be computed. Note that although 1 is the most usual, one school of thought takes the first two singular vectors for a sort of biplot.*

On retiendra donc que “l'AFC-programme” exige un tableau de nombres positifs et, par souci d'efficacité, diagonalise une matrice de dimension Min(*I*,*J*). Les termes lignes et colonnes sont donc arbitraires ou interchangeables, ce qui n'est pas toujours vrai dans l'interprétation.

Cette propriété fondamentale de la procédure ne préjuge pas de la dissymétrie éventuelle des objets concrets représentés numériquement. Interpréter l'analyse c'est utiliser les coordonnées factorielles (ou facteurs), produits numériques de l'algorithme, pour organiser la lecture des données, en préparer un résumé aussi précis que possible, éventuellement faire émerger de cette lecture et de ce résumé un modèle de la structure interne du tableau. Il est bien des manières d'opérer.

3. Corrélations entre variables qualitatives

La première des fonctions de l'AFC est de proposer une mesure de l'intensité de la relation entre deux variables qualitatives. Pour deux variables quantitatives nous avons vu l'usage du coefficient de corrélation et pour une variable quantitative et une variable qualitative celui du rapport de corrélation. La distinction entre qualitatif et quantitatif n'est d'ailleurs pas aussi clair qu'on pourrait le penser.

3.1. Exemple

Examinons l'exemple suivant d'une remarquable simplicité apparente. Legay et Pontier **[25]** ont noté l'âge et la fécondité (nombre de chatons produits dans l'année) pour 350 chattes domestiques. La répartition de 350 chattes en fonction de l'âge (1 an à 8 ans et plus) et du nombre de chatons produits dans l'année 1 ou 2 (1.5), 3 ou 4 (3.5), ..., 13 ou 14 (13.5) est :

```
> chats
   0 1-2 3-4 5-6 7-8 9-10 11-12 13-14
A1 8  15  44  11   7    4     0     0
A2 6  12  36  21  11    6     1     1
A3 4   7  18  13  12    4     0     2
A4 2   8   7   3   7    5     1     0
A5 2   3   5   3   4    6     0     0
A6 1   0   5   3   2    2     0     1
A7 2   2   8   3  12    8     1     1
```

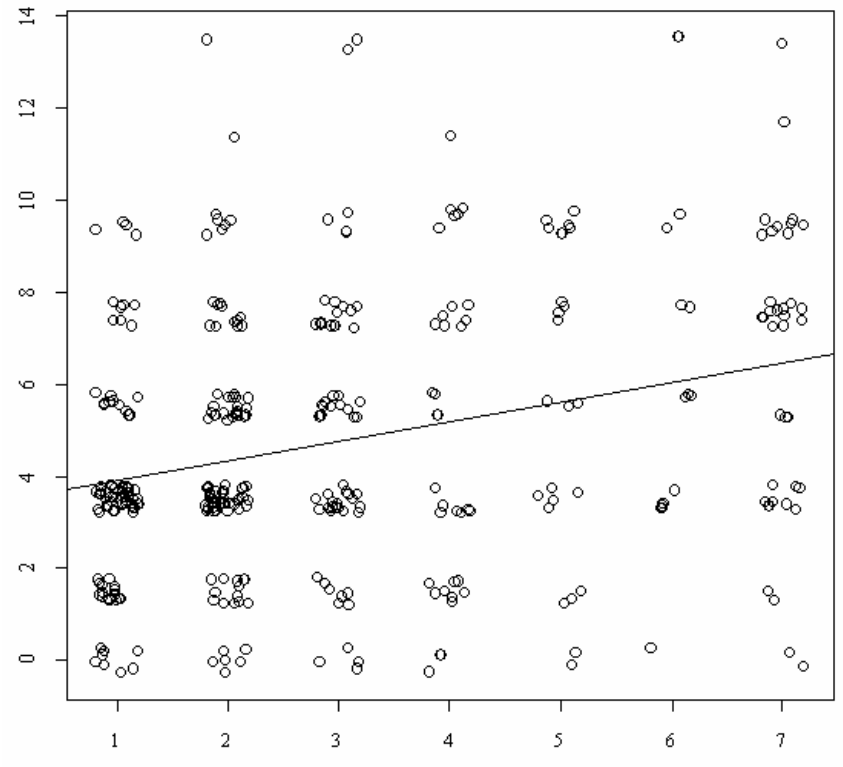
Il est question d'étudier la fécondité en fonction de l'âge, une augmentation simultanée étant un cas fréquent chez les mammifères. L'âge peut être une variable qualitative ou une variable quantitative :

```
> chats.mat <- as.matrix(chats)
> age <- rep(row(chats.mat),as.vector(chats.mat))
> age
  [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 4 4 5 5 6 7 7 1 1 1 1 1 1 1 1 1 1
1
…
[297] 7 7 7 7 7 7 7 7 7 7 7 1 1 1 1 2 2 2 2 2 2 3 3 3 3 4 4 4 4 4 5 5 5 5 5
6
[334] 6 7 7 7 7 7 7 7 7 2 4 7 2 3 3 6 7

> age.fac <- rep(row.names(chats)[row(chats.mat)],as.vector(chats.mat))
> age.fac
  [1] "A1" "A1" "A1" "A1" "A1" "A1" "A1" "A1" "A2" "A2" "A2" "A2" "A2" "A2"
"A3"
…
[331] "A5" "A5" "A6" "A6" "A7" "A7" "A7" "A7" "A7" "A7" "A7" "A7" "A2" "A4"
"A7"
[346] "A2" "A3" "A3" "A6" "A7"
```

La fécondité peut être une variable qualitative ou une variable quantitative :

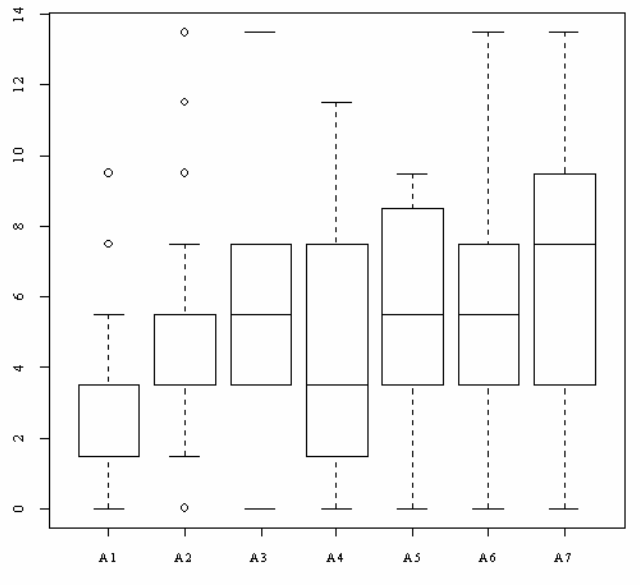
```
> w0 <- c(0,1.5,3.5,5.5,7.5,9.5,11.5,13.5)
> feco <- rep(w0[col(chats.mat)],as.vector(chats.mat))
> feco
  [1]  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
0.0
…
[331]  9.5  9.5  9.5  9.5  9.5  9.5  9.5  9.5  9.5  9.5  9.5  9.5 11.5 11.5
11.5
[346] 13.5 13.5 13.5 13.5 13.5
> feco.fac <- rep(names(chats)[col(chats.mat)],as.vector(chats.mat))
```

```
> feco.fac
  [1] "0"     "0"     "0"     "0"     "0"     "0"     "0"     "0"     "0"
…
[334] "9-10"  "9-10"  "9-10"  "9-10"  "9-10"  "9-10"  "9-10"  "9-10"  "9-10"
[343] "11-12" "11-12" "11-12" "13-14" "13-14" "13-14" "13-14" "13-14"

> age.fac <- factor(age.fac)
> feco.fac <- factor(feco.fac)

> plot(jitter(age),jitter(feco))
> abline(lm(feco~age))
> cor(age,feco)
[1] 0.2784
> cor(age,feco)^2
[1] 0.07752
> var(predict(lm(feco~age)))/var(feco)
[1] 0.07752
```



La variable **x** est considérée comme une variable quantitative (1 an, 2 ans, ...) soit comme une variable qualitative (classe 1, classe 2, ...).

```
> boxplot(split(feco,age.fac))
```



On obtient un rapport de corrélation

```
> var(predict(lm(feco~age.fac)))/var(feco)
[1] 0.08202
```

La droite de régression (meilleur prédicteur linéaire de **y** par **x**) a pour équation $y = 0.425x + 3.476$, ce qui correspondant à un coefficient de corrélation de 0.278, dont le carré donne la part de variance expliquée par la prédiction linéaire, à savoir :

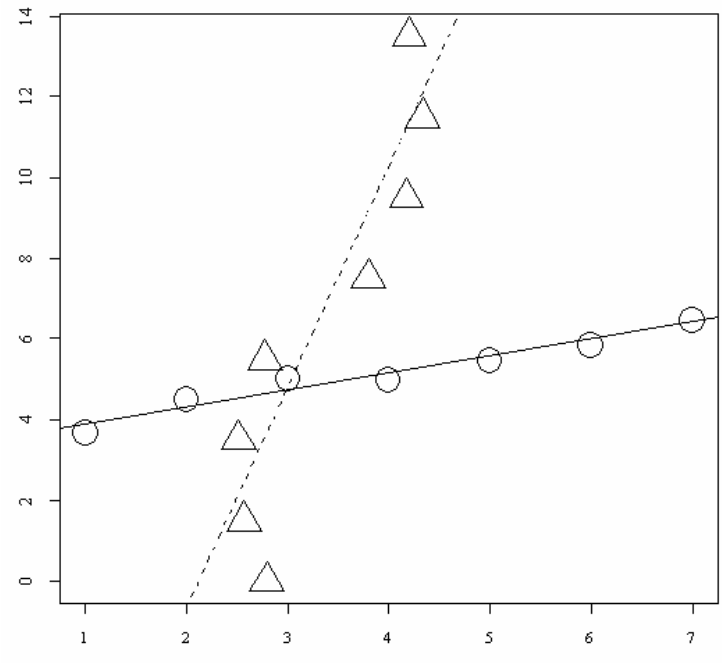
$$r^2(\mathbf{x},\mathbf{y}) = 0.0775 \leq \eta^2_{\mathbf{yx}} = 0.082$$

.

On pourrait tester le coefficient de corrélation et dire qu'il est significatif au risque de 10^{-5}. La faible différence entre $r^2(\mathbf{x},\mathbf{y})$ et $\eta^2_{\mathbf{yx}}$ et l'écart modéré entre ligne et droite de régression devrait témoigner, en première approche, d'une liaison linéaire entre les deux variables.

```
> predict(lm(feco~age),newdata=list(age=1:7))
    1     2     3     4     5     6     7
3.901 4.326 4.751 5.176 5.600 6.025 6.450
> tapply(feco,age.fac,mean)
   A1    A2    A3    A4    A5    A6    A7
3.680 4.511 5.000 4.985 5.457 5.821 6.446
> predict(lm(age~feco),newdata=list(feco=unique(feco)))
    1     2     3     4     5     6     7     8
2.125 2.399 2.764 3.129 3.494 3.859 4.224 4.589
> tapply(age,feco.fac,mean)
    0   1-2 11-12 13-14   3-4   5-6   7-8  9-10
2.800 2.574 4.333 4.200 2.512 2.772 3.800 4.171

> plot(age,feco,type="n")
> abline(lm(feco~age))
> points(1:7,tapply(feco,age.fac,mean),cex=3)
> lm(age~feco)

Call:
lm(formula = age ~ feco)

Coefficients:
(Intercept)         feco
      2.125        0.182

> abline(coef=c(-2.125/0.182,1/0.182),lty=2)
> tapply(age,feco.fac,mean)
    0   1-2 11-12 13-14   3-4   5-6   7-8  9-10
```

```
2.800 2.574 4.333 4.200 2.512 2.772 3.800 4.171
> points(tapply(age,feco.fac,mean)[names(chats)],w0,cex=3,pch=2)
```



On a obtenu les deux droites de régression et les deux courbes de régression (remarquer la bizarrerie de la seconde). Ceci introduit à la plus ancienne définition mathématique connue de l'AFC, solution donnée par Hirschfeld (1935, op. cit.) à la question:

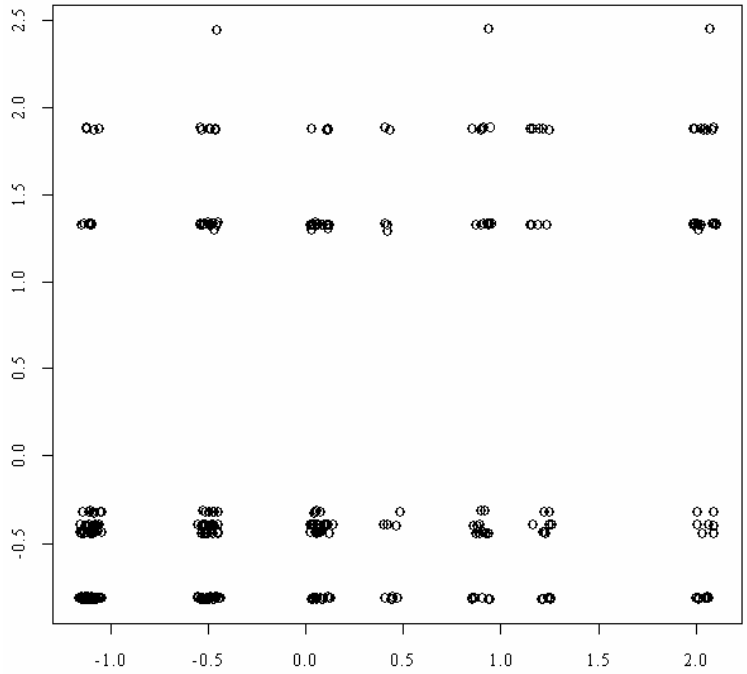
"Etant donnée une distribution discrète p_{ij} est-il possible d'introduire de nouvelles valeurs pour x_i et y_j de telle manière que les deux régressions soient linéaires?" (in Nishisato op. cit. p.13).

Le tableau est soumis a la procédure de l'AFC.

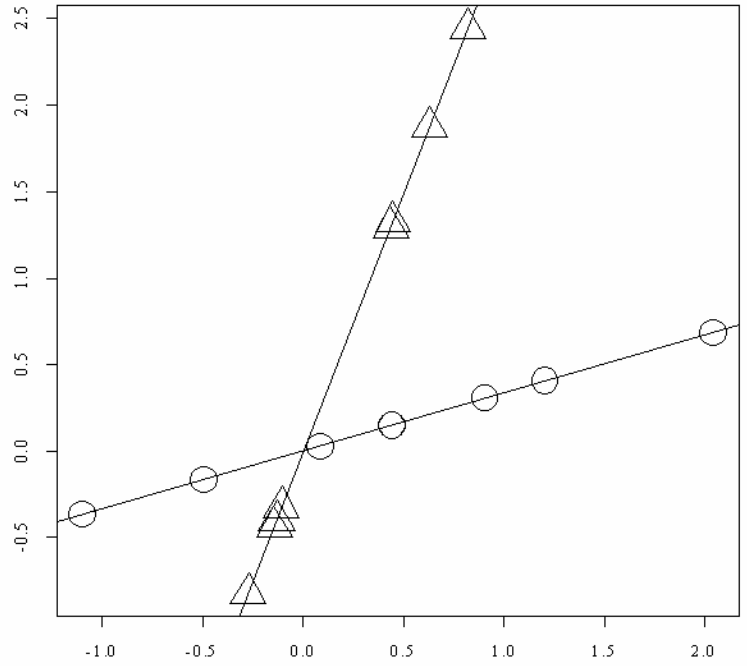
```
> library(MASS)
> afccor <- corresp(chats)
> afccor
First canonical correlation(s): 0.3351

 Row scores:
      A1       A2       A3       A4       A5       A6       A7
-1.10321 -0.49372  0.08393  0.90511  1.20630  0.44213  2.04745

 Column scores:
      0     1-2     3-4     5-6     7-8    9-10   11-12   13-14
-0.3214 -0.4401 -0.8213 -0.3984  1.3278  1.8750  2.4462  1.2916
```

On reproduit *exactement* les mêmes opérations en substituant aux valeurs observées de x (respectivement y), à savoir 1 an, 2 ans, ... (respectivement 1.5 chatons, 3.5 chatons, ...) la valeur de la première coordonnée de la colonne (respectivement la ligne) correspondante.

```
> agearti <- as.vector(afccor$rscor[age.fac])
> fecoarti <- as.vector(afccor$cscor[as.character(feco.fac)])
> plot(jitter(agearti),jitter(fecoarti))
```



```
> plot(agearti,fecoarti,type="n")
> abline(lm(fecoarti~agearti))
> points(afccor$rscor,tapply(fecoarti,age.fac,mean),cex=3)
> lm(agearti~fecoarti)
```



```
Call:
lm(formula = agearti ~ fecoarti)

Coefficients:
(Intercept)     fecoarti
  -1.82e-16     3.35e-01
> abline(coef=c(0,1/0.335))
> points(tapply(agearti,feco.fac,mean)[levels(feco.fac)],
     afccor$cscor[levels(feco.fac)],cex=3,pch=2)
```

La figure obtenue possède des propriétés de deux types. Les unes sont propres à la méthode, les autres donnent de l'information sur les données traitées. On peut observer d'abord (mais la démonstration est plus convaincante) que :

— Les moyennes marginales sont nulles (les coordonnées sont centrées pour les distributions marginales) ;

```
> mean(fecoarti)
[1] 2.236e-17
> mean(agearti)
[1] -1.719e-16
```

— Les variances marginales sont égales à 1 (les coordonnées sont normées pour les distributions marginales) ;

```
> var(fecoarti)*349/350
[1] 1
> var(agearti)*349/350
[1] 1
```

— Les points représentatifs des moyennes conditionnelles sont alignés: droite et courbe de régression sont confondues dans les deux sens ;

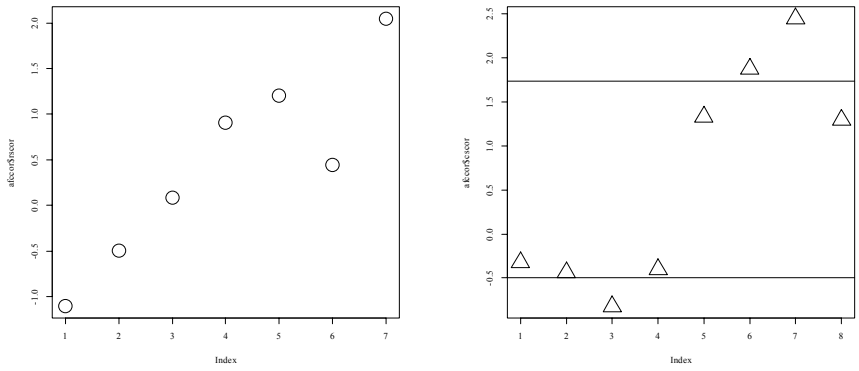
— Les variances marginales de x et y, le rapport de corrélation de y en x (et celui de x en y), le carré du coefficient de corrélation de x et y sont tous égaux à λ_1 (0.1123). La différence numérique avec le même calcul sur les données de base est bien mince mais la figure est totalement différente.

```
> afccor$cor^2
[1] 0.1123
> var(predict(lm(fecoarti~agearti)))*349/350
[1] 0.1123
> var(predict(lm(fecoarti~age.fac)))*349/350
[1] 0.1123
> var(predict(lm(agearti~fecoarti)))*349/350
[1] 0.1123
> var(predict(lm(agearti~feco.fac)))*349/350
[1] 0.1123
```

Notons que:

— Le codage ligne est sensiblement (à une transformation linéaire près) l'âge mesuré. Les rares inversions invitent à regrouper par classe pour assurer une précision constante et on gardera les classes 1 an, 2 ans, 3 ans, 4 à 6 ans, 7 ans et plus.

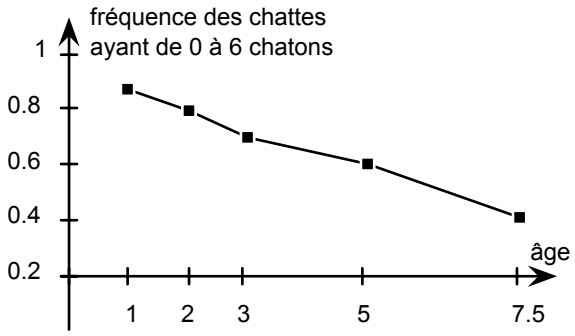
```
> plot(afccor$rscor,cex=3)
> plot(afccor$cscor,cex=3,pch=2)
> abline(h=mean(afccor$cscor[1:4]))
> abline(h=mean(afccor$cscor[5:8]))
```



— Le codage colonne impose deux groupes respectivement de 0 à 6 chatons et de 7 à 14 chatons. Il y a là un exemple remarquable d'une recherche de liaison doublement

linéaire qui renvoie à une ordination sur une variable et à une partition sur une autre, image caractéristique de la pensée de Benzécri pour qui les modèles doivent émerger des données, et non l'inverse.

Ne retenons de l'analyse que la simplification du tableau qu'elle nous propose :



Il existe deux classes de chattes et la proportion de la première décroît continûment avec l'âge: ce pourrait être alors le nombre de portées par an qui caractériserait correctement l'évolution de la fécondité. Pour affiner l'interprétation il conviendra d'analyser la liaison trivariée âge-nombre de portées-nombre de chatons, ce qui renvoie à l'analyse des correspondances multiples. Nous retiendrons que la première fonction de l'AFC est celle de la description d'une corrélation par double codage (*dual scaling*) et qu'il peut être utile de recoder numériquement des valeurs quantitatives pour apprécier correctement une relation bivariée. Nous avons utilisé dans la discussion la propriété fondamentale suivante.

3.2. Corrélation canonique

Si on cherche un code numérique des lignes et un code numérique des colonnes respectivement centrés et normés pour les distribution marginales qui maximisent la corrélation associée à la table de contingence, il suffit de prendre les premières coordonnées de norme 1 de l'AFC.

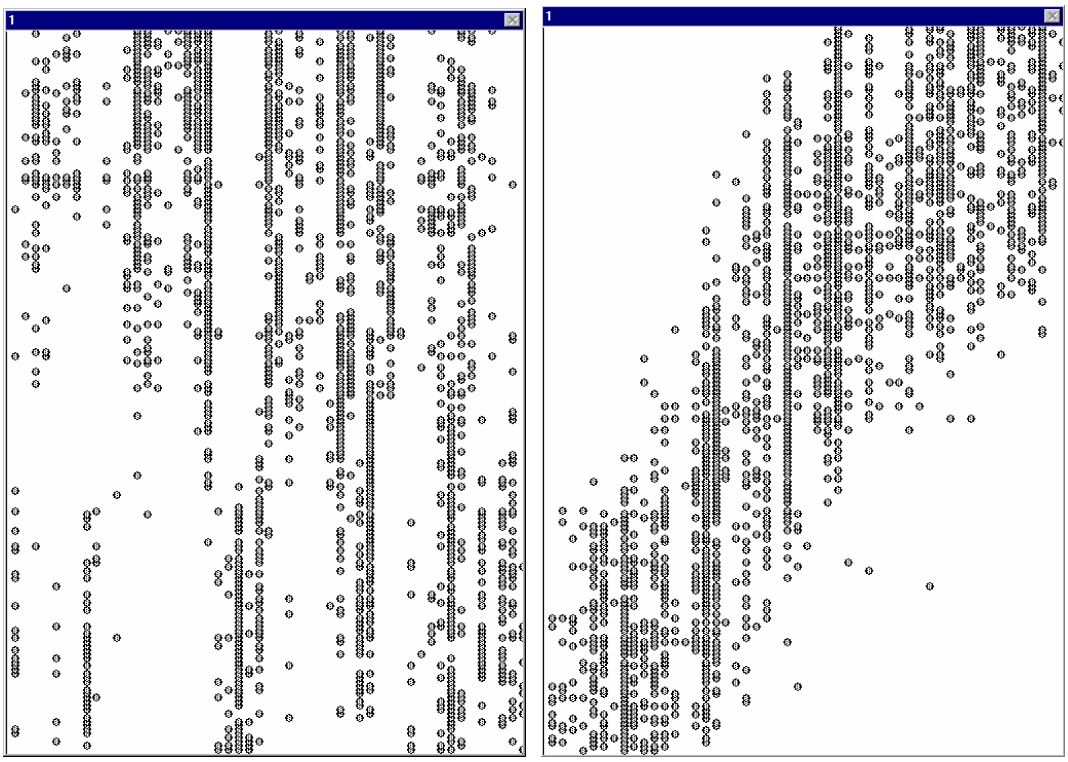
C'est une conséquence directe des propriétés générales des triplets statistiques (schéma de dualité).

Souvent considérée comme une méthode "lourde" pour l'analyse des grands tableaux, l'AFC peut être d'abord comprise comme une méthode élémentaire de statistique bivariée permettant d'examiner rapidement toute table de contingence pour laquelle on se serait contenté d'un test khi2. Ce point de vue renvoie l'AFC à ses origines inférentielles puisque Kendall et Stuart (1961, p. 571) la définissent comme test d'hypothèse contre une corrélation nulle dans une distribution binormale groupée en classes en comparant la première valeur propre à un khi2 à I+J-3 degrés de liberté.

Récupérée comme technique de double codage déjà implicite dans le point de vue de Kendall et Stuart (on estime alors les centres des classes du groupement) l'AFC facilite la description des tables de contingence de toutes dimensions. Pour les plus petites d'entre elles la réécriture du tableau ordonné par la première coordonnée factorielle suffit.

3.3. Réorganisation de tableaux

Soit un tableau écologique (présences-absences) comportant 182 relevés et 51 espèces d'oiseaux [26].



On a simplement permuté les lignes et les colonnes du tableau pour les placer par valeurs croissantes de la première coordonnée de l'AFC du tableau.

[27] Données artificielles dans

```
> a
   a b c d e f g h i j k l m n o p
1  1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
2  1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
3  0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
4  0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0
5  0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0
6  0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
7  0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0
8  0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0
9  0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0
10 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
11 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
> b <- a[,sample(16,16,replace=F)]
> b
   l d c m h p a o k b i f j g n e
1  0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0
2  0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0
3  0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 1
4  0 1 1 0 1 0 0 0 0 0 1 1 0 1 0 1
5  0 1 0 0 1 0 0 0 0 0 1 1 1 1 0 1
6  1 0 0 0 1 0 0 0 1 0 1 1 1 1 0 1
7  1 0 0 1 1 0 0 0 1 0 1 0 1 1 0 0
8  1 0 0 1 0 0 0 0 1 0 1 0 1 0 1 0
9  1 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0
10 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 0
11 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0
> c <- b[sample(11,11,replace=F),]
> c
   l d c m h p a o k b i f j g n e
8  1 0 0 1 0 0 0 0 1 0 1 0 1 0 1 0
```

```
4  0 1 1 0 1 0 0 0 0 0 1 1 0 1 0 1
3  0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 1
10 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 0
5  0 1 0 0 1 0 0 0 0 0 1 1 1 1 0 1
1  0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0
7  1 0 0 1 1 0 0 0 1 0 1 0 1 1 0 0
6  1 0 0 0 1 0 0 0 1 0 1 1 1 1 0 1
9  1 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0
11 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0
2  0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0
> library(MASS)
> cor <- corresp(c)
> d <- c[order(cor$rscore),order(cor$cscore)]
> d
   p o n m l k j i h g f e d c b a
11 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
10 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
9  0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
8  0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0
7  0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0
6  0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
5  0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0
4  0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0
3  0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0
2  0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
1  0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
```

L'AFC est un cas particulier de l'analyse canonique (voir couplages de tableaux). Dès que le volume des données augmente, la structure de la relation entre les deux variables ne peut s'exprimer qu'à travers plusieurs codages successifs et l'usage des cartes factorielles.

4. Géométrie de deux nuages

4.1. Exemple

Le tableau donne la répartition de 5694 couples formés de deux conjoints actifs en fonction de la catégorie socio-professionnelle de la femme (en lignes) et du mari (en colonnes) en 1982. Le code des lignes et des colonnes est commun :

1 — Agriculteur exploitant et salarié agricole
2 — Patron (commerce et industrie), profession libérale et cadre supérieur
3 — Cadre moyen
4 — Employé
5 — Ouvrier et personnel de service

Il est extrait de l'article de VALLET (1986 [28]) et simplifié pour des raisons pédagogiques.

```
> mariages <- read.table("mariages.txt",sep=",")
> mariages
   V1 V2  V3  V4  V5  V6  V7 V8 V9
1 420  3   8   2   2   4  20  0  0
2   2 12   0   0   0   1  10  0  0
3   9  1 333  38  24  22  48  7  3
4   3  1  18 158  53  18  24  2  5
5  19  4  61 201 309 113 225 14 29
6  20  9 111 159 323 348 756 34 65
7  13 14  41  24  79 118 795 22 19
8   8 11  30  21  61  82 382 44 20
9   0  0   1   3   2   1   2  0  6
```

4.2. Double analyse d'inertie

Dans les notations du cas général, nous pouvons d'abord considérer le tableau des distributions conditionnelles par lignes, $\mathbf{L} = \mathbf{D}_I^{-1}\mathbf{P}$, de terme général :

$$p_{j|i} = \frac{p_{ij}}{p_{i.}}$$

```
> P <- mariages/sum(mariages)
> print(P,digits=2)
       V1      V2      V3      V4      V5      V6      V7      V8      V9
1 0.07179 0.00051 0.00137 0.00034 0.00034 0.00068 0.00342 0.00000 0.00000
2 0.00034 0.00205 0.00000 0.00000 0.00000 0.00017 0.00171 0.00000 0.00000
3 0.00154 0.00017 0.05692 0.00650 0.00410 0.00376 0.00821 0.00120 0.00051
4 0.00051 0.00017 0.00308 0.02701 0.00906 0.00308 0.00410 0.00034 0.00085
5 0.00325 0.00068 0.01043 0.03436 0.05282 0.01932 0.03846 0.00239 0.00496
6 0.00342 0.00154 0.01897 0.02718 0.05521 0.05949 0.12923 0.00581 0.01111
7 0.00222 0.00239 0.00701 0.00410 0.01350 0.02017 0.13590 0.00376 0.00325
8 0.00137 0.00188 0.00513 0.00359 0.01043 0.01402 0.06530 0.00752 0.00342
9 0.00000 0.00000 0.00017 0.00051 0.00034 0.00017 0.00034 0.00000 0.00103
```

Les sommes par lignes sur $\mathbf{L}$ sont égales à 1, soit $\mathbf{L}\mathbf{1}_J = \mathbf{1}_I$. Donnons à chaque ligne le poids marginal $p_{i.}$. Nous avons un nuage de I points de $\mathbb{R}^J$ pondérés par $\mathbf{D}_I = Diag\left(p_{1.},\ldots,p_{I.}\right)$. Calculons le centre de gravité de ce nuage :

$$\sum_{i=1}^{I} p_{i.} \frac{p_{ij}}{p_{i.}} = \sum_{i=1}^{I} p_{ij} = p_{.j} \Rightarrow g_J = \sum_{i=1}^{I} p_{i.} \mathbf{L}_i = \begin{bmatrix} p_{.1} \\ \\ p_{.J} \end{bmatrix} = \mathbf{D}_J \mathbf{1}_J$$

et centrons le tableau **L** par :

$$\mathbf{L}_0 = \mathbf{L} - \mathbf{1}_{IJ} \mathbf{D}_J$$

```
> DI <- apply(P,1,sum)
> DI
       1          2          3          4          5          6          7          8
9
0.078462  0.004274  0.082906  0.048205  0.166667  0.311966  0.192308  0.112650
0.002564
Madame est employée dans 31.2% des couples
> DJ <- apply(P,2,sum)
> DJ
      V1         V2         V3         V4         V5         V6         V7         V8
V9
0.084444  0.009402  0.103077  0.103590  0.145812  0.120855  0.386667  0.021026
0.025128
Monsieur est employée dans 12.9% des couples
> L <- P/DI
> L
       V1         V2        V3         V4         V5         V6        V7         V8
V9
1  0.91503  0.006536  0.01743  0.004357  0.004357  0.008715  0.04357  0.000000
0.000000
2  0.08000  0.480000  0.00000  0.000000  0.000000  0.040000  0.40000  0.000000
0.000000
3  0.01856  0.002062  0.68660  0.078351  0.049485  0.045361  0.09897  0.014433
0.006186
4  0.01064  0.003546  0.06383  0.560284  0.187943  0.063830  0.08511  0.007092
0.017730
5  0.01949  0.004103  0.06256  0.206154  0.316923  0.115897  0.23077  0.014359
0.029744
6  0.01096  0.004932  0.06082  0.087123  0.176986  0.190685  0.41425  0.018630
0.035616
7  0.01156  0.012444  0.03644  0.021333  0.070222  0.104889  0.70667  0.019556
0.016889
8  0.01214  0.016692  0.04552  0.031866  0.092564  0.124431  0.57967  0.066768
0.030349
9  0.00000  0.000000  0.06667  0.200000  0.133333  0.066667  0.13333  0.000000
0.400000
> apply(L,1,sum)
1 2 3 4 5 6 7 8 9
1 1 1 1 1 1 1 1 1
Quand Madame est cadre supérieur, dans 56% des cas Monsieur l'est aussi.
> C <- P/rep(DJ,rep(9,9))
> C
        V1       V2        V3         V4        V5        V6         V7       V8
V9
1  0.850202  0.05455  0.013267  0.003300  0.002345  0.005658  0.0088417  0.00000
0.00000
2  0.004049  0.21818  0.000000  0.000000  0.000000  0.001414  0.0044209  0.00000
0.00000
3  0.018219  0.01818  0.552239  0.062706  0.028136  0.031117  0.0212202  0.05691
0.02041
4  0.006073  0.01818  0.029851  0.260726  0.062134  0.025460  0.0106101  0.01626
0.03401
5  0.038462  0.07273  0.101161  0.331683  0.362251  0.159830  0.0994695  0.11382
0.19728
6  0.040486  0.16364  0.184080  0.262376  0.378664  0.492221  0.3342175  0.27642
0.44218
7  0.026316  0.25455  0.067993  0.039604  0.092614  0.166902  0.3514589  0.17886
0.12925
8  0.016194  0.20000  0.049751  0.034653  0.071512  0.115983  0.1688771  0.35772
0.13605
9  0.000000  0.00000  0.001658  0.004950  0.002345  0.001414  0.0008842  0.00000
0.04082
```

```
> apply(C,2,sum)
V1 V2 V3 V4 V5 V6 V7 V8 V9
 1  1  1  1  1  1  1  1  1
Quand Monsieur est cadre supérieur, dans 26% des cas Madame l'est aussi.
>                                  label                                    <-
c("Agri_exp","Agri_sal","Patron","Cadre_Sup","Cadre_moy","Employ","Ouvrier","S
ervice","Autre")
```

Les lignes de $\mathbf{L}_0$ forment un nuage de I points de $\mathbb{R}^J$ pondérés par $\mathbf{D}_I = Diag\left(p_{1.},\ldots,p_{I.}\right)$. Ce nuage est entièrement dans l'hyperplan :

$$\sum_{i=1}^{J} x_j = 0$$

Pour en faire une analyse d'inertie, il suffit de munir $\mathbb{R}^J$ d'un produit scalaire. Nous prendrons le produit scalaire diagonal défini par :

$$\mathbf{D}_J^{-1} = Diag\left(1/p_{.1},\ldots,1/p_{.J}\right)$$

On obtient alors la décomposition canonique du triplet :

$$\left(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_{IJ}\mathbf{D}_J, \mathbf{D}_J^{-1}, \mathbf{D}_I\right)$$

L'opérateur associé est :

$$\mathbf{WD} = \left(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_{IJ}\mathbf{D}_J\right)\mathbf{D}_J^{-1}\left(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_{IJ}\mathbf{D}_J\right)^{\mathrm{t}}\mathbf{D}_I$$

qui s'écrit encore :

$$\mathbf{WD} = \left(\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}\right)\mathbf{D}_J\left(\mathbf{D}_J^{-1}\mathbf{P}^{\mathrm{t}}\mathbf{D}_I^{-1} - \mathbf{1}_{JI}\right)\mathbf{D}_I$$

c'est-à-dire :

$$\mathbf{WD} = \left(\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}\right)\mathbf{D}_J\left(\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}\right)^{\mathrm{t}}\mathbf{D}_I$$

Les triplets ($\mathbf{Z}$, $\mathbf{D}_J$, $\mathbf{D}_I$) et $\left(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_{IJ}\mathbf{D}_J, \mathbf{D}_J^{-1}, \mathbf{D}_I\right)$ ont donc les mêmes composantes principales, d'où :

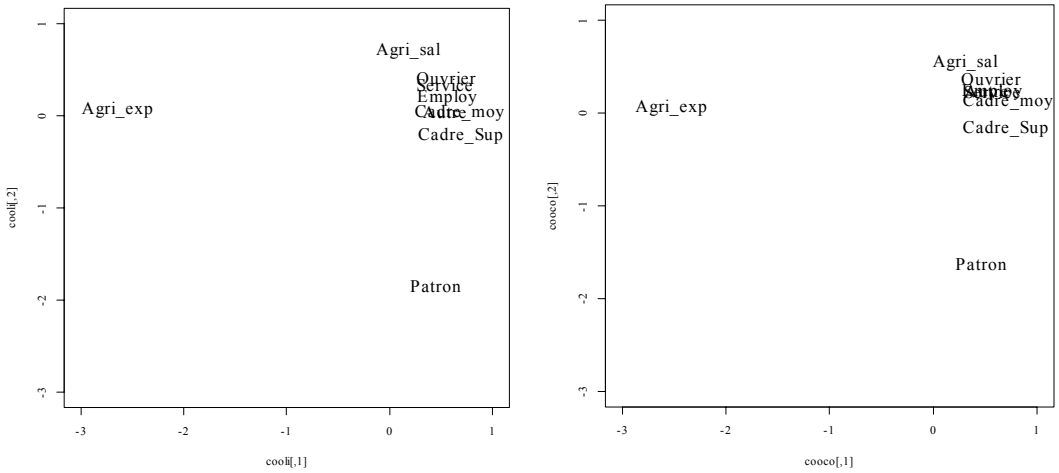
Les coordonnées des lignes fournies par la procédure générale sont celles de l'analyse d'inertie du nuage des I distributions conditionnelles par ligne, pondéré et centré par la distribution marginale des lignes utilisant la métrique des inverses des poids des colonnes.

Il s'en suit par symétrie que :

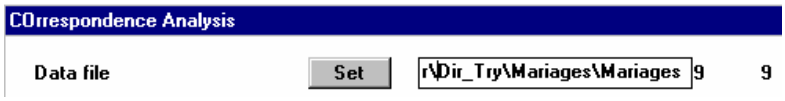
Les coordonnées des colonnes fournies par la procédure générale sont celles de l'analyse d'inertie du nuage des J distributions conditionnelles par colonne, pondéré et centré par la distribution marginale des colonnes utilisant la métrique des inverses des poids des lignes.

```
> cooli <- marcor$rscor*rep(marcor$cor,rep(9,2))
> cooco <- marcor$cscor*rep(marcor$cor,rep(9,2))
> cooli
     [,1]      [,2]
1 -2.9877  0.034733
2 -0.1310  0.673422
3  0.1977 -1.870622
```

```
4  0.2794 -0.244576
5  0.2442  0.005416
6  0.2699  0.165144
7  0.2606  0.382632
8  0.2600  0.312807
9  0.3251  0.011703
> cooco
        [,1]     [,2]
V1 -2.870291  0.03198
V2 -0.003025  0.50860
V3  0.210267 -1.65237
V4  0.284431 -0.18752
V5  0.286641  0.09324
V6  0.277572  0.20332
V7  0.264129  0.34854
V8  0.295512  0.19844
V9  0.301896  0.20513

> plot(cooli,type="n",xlim=c(-3,1),ylim=c(-3,1))
> text(cooli,label,cex=1.5,adj=0)
> plot(cooco,type="n",xlim=c(-3,1),ylim=c(-3,1))
> text(cooco,label,cex=1.5,adj=0)
```

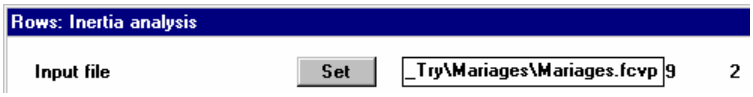


La figure donne les cartes factorielles, projections des deux nuages sur les axes d'inertie. La lecture de ces cartes doit explicitement intégrer cette notion de poids. En ACP, avec une pondération uniforme des lignes et des colonnes, le problème ne se pose pas. Ici il est prépondérant. On utilise les statistiques d'inertie :




```
Binary  input  file:  D:\ADE4USER\DIR_TRY\MARIAGES\Mariages.fcli  -  9  rows,  2
cols.
   1 |  2.9877 -0.0347
   2 |  0.1310 -0.6734
   3 | -0.1977  1.8706
   4 | -0.2794  0.2446
   5 | -0.2442 -0.0054
   6 | -0.2699 -0.1651
   7 | -0.2606 -0.3826
   8 | -0.2600 -0.3128
   9 | -0.3251 -0.0117

Binary  input  file:  D:\ADE4USER\DIR_TRY\MARIAGES\Mariages.fcco  -  9  rows,  2
cols.
   1 |  2.8703 -0.0320
   2 |  0.0030 -0.5086
   3 | -0.2103  1.6524
```

```
   4 | -0.2844  0.1875
   5 | -0.2866 -0.0932
   6 | -0.2776 -0.2033
   7 | -0.2641 -0.3485
   8 | -0.2955 -0.1984
   9 | -0.3019 -0.2051
```




```
Row inertia
All contributions are in 1/10000

----------Absolute contributions----------
|Num  |Fac 1|Fac 2|
|    1| 9202|    2|
|    2|    0|   56|
|    3|   42| 8464|
|    4|   49|   84|
|    5|  130|    0|
|    6|  298|  248|
|    7|  171|  821|
|    8|  100|  321|
|    9|    3|    0|

----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
|    1| 9997|    1||    1  |  784   | 4698 |
|    2|    7|  188|| 9803  |   42   |  688 |
|    3|  105| 9437||  457  |  829   | 2061 |
|    4|  327|  251|| 9421  |  482   |  770 |
|    5| 1363|    0|| 8635  | 1666   |  488 |
|    6| 5215| 1952|| 2831  | 3119   |  292 |
|    7| 1410| 3042|| 5546  | 1923   |  620 |
|    8| 1847| 2674|| 5477  | 1126   |  276 |
|    9|  176|    0|| 9823  |   25   |  103 |

> mariages <- as.matrix(read.table("mariages.txt"))
> ca(mariages)
$evals
[1] 0.761033 0.342714 0.214863 0.100492 0.038923 0.015453 0.009521


$rcntr
        Factor1   Factor2   Factor3   Factor4   Factor5   Factor6   Factor7
 [1,] 9.203e-01 2.762e-04 8.356e-05 7.711e-04 4.262e-05 4.876e-05 7.800e-06
 [2,] 9.643e-05 5.655e-03 2.452e-02 9.403e-01 2.015e-02 1.029e-10 4.694e-03
 [3,] 4.256e-03 8.465e-01 6.522e-02 2.125e-04 4.099e-04 3.998e-05 4.287e-06
 [4,] 4.945e-03 8.414e-03 4.068e-01 3.355e-02 4.296e-01 1.847e-02 4.713e-03
 [5,] 1.306e-02 1.426e-05 2.465e-01 6.005e-04 1.429e-01 1.115e-01 5.710e-02
 [6,] 2.986e-02 2.483e-02 3.982e-04 1.960e-02 1.643e-01 1.397e-02 1.308e-01
 [7,] 1.715e-02 8.215e-02 1.858e-01 4.783e-03 2.067e-01 3.268e-02 1.086e-01
 [8,] 1.001e-02 3.216e-02 6.688e-02 6.093e-06 1.856e-02 4.999e-02 6.779e-01
 [9,] 3.561e-04 1.025e-06 3.900e-03 1.813e-04 1.727e-02 7.733e-01 1.615e-02

$ccntr
        Factor1   Factor2   Factor3   Factor4   Factor5   Factor6   Factor7
 [1,] 9.142e-01 0.0002520 3.398e-04 7.739e-04 1.087e-06 1.418e-05 1.988e-05
 [2,] 1.130e-07 0.0070961 2.759e-02 9.394e-01 1.607e-02 1.499e-04 8.746e-07
 [3,] 5.988e-03 0.8211855 6.788e-02 2.420e-09 4.222e-04 6.712e-05 4.780e-04
 [4,] 1.101e-02 0.0106282 5.367e-01 2.864e-02 2.856e-01 9.243e-03 2.435e-03
 [5,] 1.574e-02 0.0036987 1.476e-01 1.869e-03 3.457e-01 1.546e-01 3.398e-02
 [6,] 1.224e-02 0.0145773 6.657e-06 1.157e-02 1.416e-01 3.686e-03 7.999e-02
 [7,] 3.545e-02 0.1370612 2.079e-01 1.389e-02 1.424e-01 1.138e-02 1.380e-02
 [8,] 2.413e-03 0.0024158 9.186e-03 1.235e-03 3.217e-03 2.621e-02 8.633e-01
 [9,] 3.009e-03 0.0030853 2.805e-03 2.592e-03 6.500e-02 7.946e-01 5.951e-03
```

Ajouter les définitions des statistiques d'inertie.

4.3. Relations entre cartes factorielles

Il est établi que la procédure d'AFC conduit à deux analyses d'inerties de deux nuages dans deux espaces avec deux métriques et deux pondérations, tous les paramètres utilisés dérivant du tableau initial. En fait chacune de ces analyses pourrait elle-même être une double analyse d'inertie du nuage des lignes et des colonnes associées au même triplet. Il y aurait donc quatre analyses d'inertie dont deux seraient celles des lignes et des colonnes du triplet de référence. La pratique collective n'a conservé que deux des quatre possibles, pour la simplicité de leur conception.

Les deux cartes des deux analyses retenues entretiennent des relations très fortes qui sont une des caractéristiques de l'AFC.

4.4. Moyennes conditionnelles

Les lignes positionnées par les coordonnées de variances λ_k (colonnes de $\tilde{\mathbf{C}}_K$) sont à la moyenne conditionnelle des colonnes positionnées par les coordonnées de variances 1 (colonnes de $\mathbf{A}_k$) et réciproquement, les colonnes positionnées par les coordonnées de variances λ_k (colonnes de $\tilde{\mathbf{A}}_K$) sont à la moyenne conditionnelle des lignes positionnées par les coordonnées de variances 1 (colonnes de $\mathbf{C}_K$)

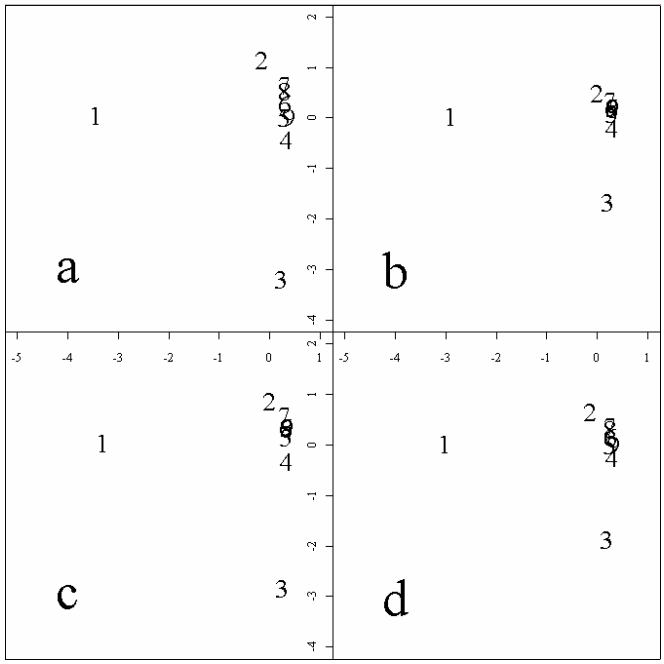
Il s'agit d'une conséquence directe des théorèmes de transition. Par exemple, la relation générale (Fiche BSA p. 5) :

$$\mathbf{B}_r = \mathbf{XQA}_r \Lambda_r^{-\frac{1}{2}}$$

s'écrit :

$$\tilde{\mathbf{C}}_K = \left(\mathbf{D}_I^{-1}\mathbf{PD}_J^{-1} - \mathbf{1}_{IJ}\right)\mathbf{D}_J\mathbf{A}_K = \mathbf{D}_I^{-1}\mathbf{PA}_K - \mathbf{1}_{IJ}\mathbf{D}_J\mathbf{A}_K = \mathbf{D}_I^{-1}\mathbf{PA}_K$$

car les coordonnées des colonnes sont $\mathbf{D}_J$ centrées. L'opérateur $\mathbf{D}_I^{-1}\mathbf{P}$ associe à un vecteur à *J* composantes le vecteur des moyennes par lignes pour le tableau **T**. La figure illustre cette propriété. La symétrie lignes-colonnes est alors rompue et on peut hésiter entre les deux représentations par moyennes conditionnelles.



a - Position des lignes (femmes) avec les coordonnées de variance 1. b - Position des colonnes (hommes) avec les coordonnées de variance lambda . c - Position des colonnes avec les coordonnées de variance 1. d - Position des lignes avec les coordonnées de variance lambda . On passe de a à b par les moyennes des distributions colonnes. On passe de b à c par une dilatation sur chaque axe. On passe de c à d par les moyennes des distributions lignes. On passe de d à a par une dilatation sur chaque axe. On revient en un tour au même endroit parce qu'on utilise des vecteurs propres.

La figure est obtenue par :

```
mar <- function() {
   par(mfrow=c(2,2))
   par(mai=c(0,0,0,0))
   labelnum <- as.character(1:9)
   plot(marcor$rscor,xlim=c(-5,1),ylim=c(-4,2),type="n")
   text(marcor$rscor,labelnum,cex=2)
   text(-4,-3,"a",cex=4)

   plot(cooco,xlim=c(-5,1),ylim=c(-4,2),type="n",cex=2)
   text(cooco,labelnum,cex=2)
   text(-4,-3,"b",cex=4)

   plot(marcor$cscor,xlim=c(-5,1),ylim=c(-4,2),type="n",cex=2)
   text(marcor$cscor,labelnum,cex=2)
   text(-4,-3,"c",cex=4)

   plot(cooli,xlim=c(-5,1),ylim=c(-4,2),type="n",cex=2)
   text(cooli,labelnum,cex=2)
   text(-4,-3,"d",cex=4)
}
```

4.5. Dilatation

On ne peut placer simultanément les lignes à la moyenne des colonnes et les colonnes à la moyenne des lignes.

Si on superpose les cartes construites avec les coordonnées de $\tilde{\mathbf{C}}_K$ et $\tilde{\mathbf{A}}_K$, on a une double relation de transition qui s'écrit simplement :

$$\begin{cases} \tilde{C}_K(i,k) = \dfrac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{J} p_{j/i} \tilde{A}_K(j,k) \\ \tilde{A}_K(j,k) = \dfrac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{I} p_{i/j} \tilde{C}_K(i,k) \end{cases}$$

Cette propriété est optimale, en ce sens que, si on a un code numérique **x** des lignes $\mathbf{D}_I$-centré et un code numérique des colonnes **y** $\mathbf{D}_J$-centré présentant une double relation en moyenne conditionnelle dilatée, alors :

$$\begin{cases} \mathbf{x}_i = a \sum_{j=1}^{J} p_{j/i} \mathbf{y}_j \\ \mathbf{y}_j = a \sum_{i=1}^{I} p_{i/j} \mathbf{x}_i \end{cases} \Rightarrow \begin{cases} \mathbf{x} = a \mathbf{P} \mathbf{D}_J^{-1} \mathbf{y} \\ \mathbf{y} = a \mathbf{P}^t \mathbf{D}_I^{-1} \mathbf{x} \end{cases} \Rightarrow \mathbf{x} = a^2 \mathbf{P} \mathbf{D}_J^{-1} \mathbf{P}^t \mathbf{D}_I^{-1} \mathbf{x}$$

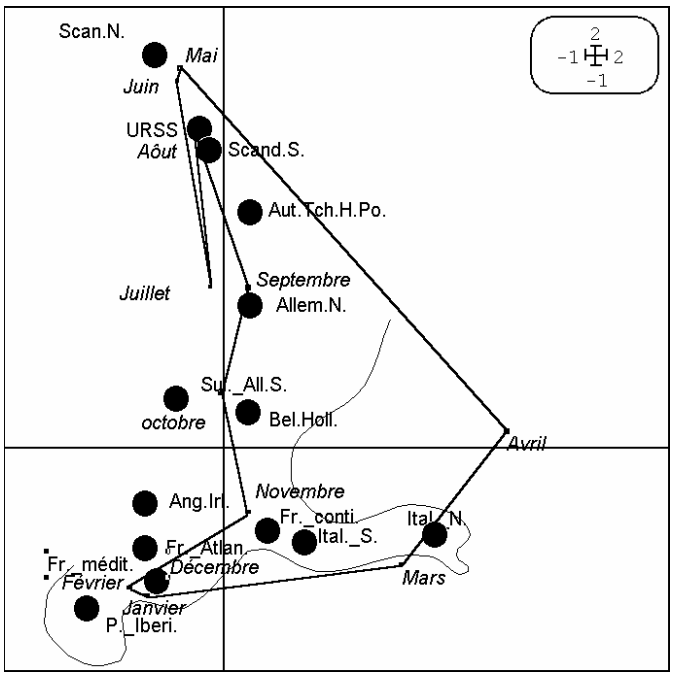
La dilatation est la moins sensible si *a* est le plus proche possible de l'unité, ce qui est réalisé pour $a = \sqrt{\lambda_1}$ et les deux premiers scores de norme 1.

Considérons par exemple, les données [29] consignées dans le tableau ci-dessous, dont tous les éléments sont extraits de [13].

	Jui	Aou	Sep	Oct	Nov	Déc	Jan	Fév	Mar	Avr	Mai	Jun
Fr._médit.	0	12	44	44	48	69	179	500	147	4	0	0
Fr._conti.	4	4	30	16	20	11	17	75	70	1	0	0
Fr._Atlan.	0	7	9	13	7	14	18	74	16	0	0	0
P._Iberi.	1	0	0	3	5	6	11	98	6	0	0	0
Ital._N.	0	4	38	18	21	9	16	87	218	51	2	0
Ital._S.	0	1	2	5	3	7	1	11	14	2	0	0
Ang.Irl.	1	0	3	2	0	7	5	1	0	1	0	0
Bel.Holl.	0	7	28	17	9	4	9	9	10	2	0	0
Sui._All.S.	0	4	8	12	2	3	5	5	1	0	0	1
Allem.N.	1	12	20	13	9	2	0	0	2	1	0	0
Aut.Tch.H.Po.	1	43	31	7	2	1	1	1	4	5	0	1
Scand.S.	4	68	53	15	3	0	0	0	0	5	25	7
Scan.N.	0	14	3	0	0	0	0	0	0	0	7	1
URSS	3	184	105	34	5	0	0	0	0	11	83	13

Table de contingence croisant régions et mois de capture pour 3049 reprises de bagues de Sarcelles d'hiver.

Dans le tableau à I = 14 lignes et J = 12 colonnes sont inscrits les effectifs n_{ij} de Sarcelles d'hiver (*Anas C. Crecca*) dont la bague a été récupérée dans la région i au cours du mois j (n=3049). L'exécution traditionnelle de l'analyse consiste à replacer chacune des lignes et des colonnes sur un plan à l'aide de ses coordonnées factorielles. On retrouve ainsi une carte d'Europe approximative (Lebreton op. cit.):



Carte factorielle F1-F2 de l'AFC du tableau Sarcelles. Superposition des lignes (régions) et des colonnes (dates) et lecture de la migration en boucle sur une carte factorielle représentation "topologiquement" correcte de l'espace. Cercles : position des lignes (régions) avec les coordonnées normées à la valeur propre. Carrés reliés par une trajectoire : position des colonnes (dates) avec les coordonnées normées à la valeur propre.

5. Double discrimination

Les approches qui précèdent insistent sur les notions de moyennes. Les variances associées aux distributions conditionnelles jouent également un grand rôle dans la conception de l'AFC.

Considérons la table de contingence **P** et un code numérique des lignes **x** (vecteur de $\mathbb{R}^I$). Chaque colonne du tableau est une pondération associée à **x** qui définit une moyenne et une variance. La variance marginale $v_{\mathbf{x}}$ se décompose en variance des moyennes conditionnelles, ou variance inter-colonne $b_{\mathbf{x}}$ et moyenne des variances conditionnelles ou variance intra-colonne $w_{\mathbf{x}}$.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_I \end{bmatrix} \begin{bmatrix} p_{1/1} & p_{1/j} & p_{1/J} \\ \vdots & \vdots & \vdots \\ p_{i/1} & p_{i/j} & p_{i/J} \\ \vdots & \vdots & \vdots \\ p_{I/1} & p_{I/j} & p_{I/J} \end{bmatrix} \begin{bmatrix} x_1 \; p_{1.} & \\ \vdots \; \vdots & \\ x_i \; p_{i.} & \rightarrow \; \overline{x} \\ \vdots \; \vdots & \quad v_{\mathbf{x}} \\ x_I \; p_{I.} & \end{bmatrix}$$

$$\begin{bmatrix} \downarrow & \downarrow & \downarrow \\ p_{.1} & p_{.j} & p_{.J} \\ \overline{x}_{/1} & \overline{x}_{/i} & \overline{x}_{/J} \\ p_{.1} & p_{.j} & p_{.J} \\ v_{\mathbf{x}/1} & v_{\mathbf{x}/j} & v_{\mathbf{x}/J} \end{bmatrix} \begin{matrix} \\ \\ \rightarrow \; \overline{\mathbf{x}} \; b_{\mathbf{x}} \\ \\ \rightarrow \; w_{\mathbf{x}} \end{matrix}$$

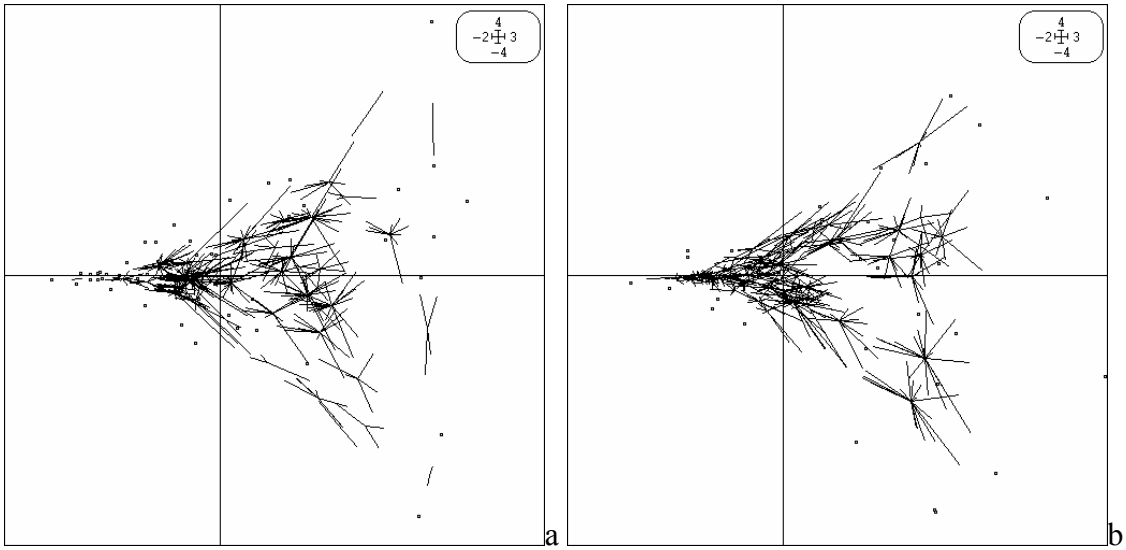
Le rapport de corrélation est la part de variance inter dans la variance totale :

$$\eta^2_{\mathbf{x}} = \frac{b_{\mathbf{x}}}{v_{\mathbf{x}}}$$

Un bon code des lignes optimise cette quantité, qui reste inchangée par $\mathbf{D}_I$-centrage. Si on cherche un code numérique des lignes qui maximise ce rapport de corrélation, il suffit de prendre la première coordonnée factorielle de l'AFC. En effet, c'est une application simple du théorème d'optimisation des formes quadratiques associée à la décomposition des schémas de dualité . Soit un vecteur $\mathbf{D}_I$ -normé.

$$Q_{\mathrm{F}}(\mathbf{x}) = \left\| \mathbf{X}^{\mathrm{t}} \mathbf{D}(\mathbf{x}) \right\|^2_{\mathbf{Q}} = \left\| \left(\mathbf{D}_J^{-1} \mathbf{P}^{\mathrm{t}} \mathbf{D}_I^{-1} - \mathbf{1}_{JI} \right) \mathbf{D}_I \mathbf{x} \right\|^2_{\mathbf{D}_J} = b_{\mathbf{x}}$$

Cette quantité est optimisée par la première composante principale.



a b

Analyse des correspondances d'un tableau avifaunistique (51 relevés et 40 espèces [30]). a - Position des relevés avec un score unité (carrés). Etoiles : position des espèces à la moyenne (scores de variance lambda). Les relevés sont positionnés pour maximiser la variance des positions des espèces. b - Position des espèces avec un score unité (carrés). Etoiles : position des espèces à la moyenne (scores de variance lambda). Les espèces sont positionnées pour maximiser la variance des positions des relevés. La relation ne semble pas symétrique.

La caractéristique de l'AFC est d'être une **double** analyse discriminante. L'avantage est de ne pas préciser dans quel sens doit se faire la discrimination. L'inconvénient est que si on doit privilégier un point de vue, l'autre sera en œuvre automatiquement. On trouvera des compléments dans le chapitre sur l'analyse canonique.

6. Un schéma de dualité peut en cacher un autre

Nous avons vu que l'AFC-programme est l'analyse du schéma :

$$\begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1} - \mathbf{1}_{IJ} \uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array} \quad \textbf{[Schéma 1]}$$

Les liens entre les points de vue se font simplement en jouant sur une réécriture qui déplace les matrices. On utilisait avant les travaux d'Escoufier le schéma :

$$\begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1} \uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array} \quad \textbf{[Schéma 2]}$$

Dans le second le vecteur $\mathbf{1}_J$ est axe principal pour la valeur propre 1 (on l'éliminait volontairement), dans le premier il est axe principal pour la valeur propre 0 (il s'élimine de lui-même). Le reste est inchangé.

Le schéma du nuage des profils lignes est :

$$\begin{array}{ccc} & \mathbf{D}_J^{-1} & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{P}^t\mathbf{D}_I^{-1} - \mathbf{D}_J\mathbf{1}_{IJ} \uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_{IJ}\mathbf{D}_J \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array} \quad \textbf{[Schéma 3]}$$

Il a les mêmes composantes principales que le schéma 1 ($\mathbf{D}_J\mathbf{D}_J^{-1}\mathbf{D}_J = \mathbf{D}_J$) ce qui donne les coordonnées du nuage des profils lignes.

Le schéma du nuage des profils colonnes est :

$$\begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{D}_J^{-1}\mathbf{P}^t - \mathbf{1}_{IJ}\mathbf{D}_I \uparrow & & \downarrow \mathbf{P}\mathbf{D}_J^{-1} - \mathbf{D}_I\mathbf{1}_{IJ} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I^{-1} & \end{array} \quad \textbf{[Schéma 4]}$$

Il a les mêmes axes principaux que le schéma 1 ($\mathbf{D}_I\mathbf{D}_I^{-1}\mathbf{D}_I = \mathbf{D}_I$) ce qui donne les coordonnées du nuage des profils colonnes. On voit ainsi que l'AFC est une demi-analyse d'inertie de deux schémas associés et non une double analyse d'inertie du même schéma comme dans une ACP.

On retrouvera enfin dans l'analyse canonique :

$$\begin{array}{ccc} & \mathbf{D}_J^{-1} & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{P}^t \uparrow & & \downarrow \mathbf{P} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I^{-1} & \end{array} \quad \textbf{[Schéma 5]}$$

L'exécution de l'une des cinq analyses donnent les éléments propres (axes, composantes, facteurs, cofacteurs) des autres à une transformation linéaire simple (mettant en jeu une matrice diagonale) près. Seul le schéma de dualité rend très simplement compte des relations entre les points de vue.

Pour conclure, on notera, en mettant en parallèle ACP et AFC que la même procédure de calcul lue sur le schéma rend compte de raisonnements concrets en apparence éloignés.

$$\begin{array}{ccc} & \mathbf{Q} = \mathbf{I}_p & \\ \boxed{p} & \rightarrow & \boxed{p} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \boxed{n} & \leftarrow & \boxed{n} \\ & \mathbf{D} = \frac{1}{n}\mathbf{I}_n & \end{array} \qquad \begin{array}{ccc} & \mathbf{D}_J & \\ \boxed{J} & \rightarrow & \boxed{J} \\ \mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{D}_I^{-1} \uparrow & & \downarrow \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} \\ \boxed{I} & \leftarrow & \boxed{I} \\ & \mathbf{D}_I & \end{array}$$

A gauche, on prend un vecteur $\mathbf{a}$, $\mathbf{I}_p$-normé (poids des variables), on en tire une combinaison linéaire des variables $\mathbf{y} = \mathbf{X}\mathbf{a}$ de variance maximale ($\|\mathbf{y}\|^2_{\mathbf{D}}$). On normalise ($\frac{1}{\sqrt{\lambda}}\|\mathbf{y}\|^2_{\mathbf{D}}$) et on a une combinaison linéaire des variables de variance 1 qui maximise la somme des carrés des corrélations avec les variables de départ ($\|\mathbf{X}^t\mathbf{y}\|^2_{\mathbf{Q}}$). On normalise à nouveau et on est au point de départ (vecteur propre).

A droite, on prend un vecteur $\mathbf{a}$, $\mathbf{D}_J$-normé (score des colonnes), on en tire la moyenne conditionnelle par profils lignes $\mathbf{y} = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1}\mathbf{D}_J\mathbf{a} = \mathbf{D}_I^{-1}\mathbf{P}\mathbf{a}$ de variance maximale ($\|\mathbf{y}\|^2_{\mathbf{D}_I}$). On normalise ($\frac{1}{\sqrt{\lambda}}\|\mathbf{y}\|^2_{\mathbf{D}_I}$) et on a un score des lignes de variance 1 qui maximise la variance des moyennes conditionnelles par colonne ($\|\mathbf{D}_J^{-1}\mathbf{P}^t\mathbf{y}\|^2_{\mathbf{D}_J}$). On normalise à nouveau et on est au point de départ (vecteur propre).

Ce principe permet de créer des analyses originales pour des structures de données particulières et de prévoir le comportement d'un jeu de paramètres dans le schéma général.

7. Références

1 Greenacre, M. (1984) *Theory and applications of correspondence analysis*. Academic Press, London. 1-364.

2 Buyse, M. (1983) Les différentes approches conduisant à l'analyse des correspondances. *Biométrie-praximétrie* : 23, 1, 1-26.

3 Richardson, M. & Kuder, G.F. (1933) Making a rating scale the measures. *Personnel Journal* : 12, 36-40. Horst, P. (1935) Measuring complex attitude. *Journal of Social Psychology* : 6, 369-374. Hirschfeld, H.O. (1935) A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences* : 31, 520-524.

4 Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika* : 39, 274-289.

5 Kendall, D.G. & Stuart, A. (1961) *The advanced theory of statistics*. Vol 2: Inference and relationships. Cha. 33 : Categorized data. Griffin, London. 536-591.

6 Hatheway, W.H. (1971) Contingency table analysis of rain forest vegetation. In : *Statistical Ecology. III Many species populations ecosystems and systems analysis*. Patil, G.P., Pielou, E.C. & Waters, W.E. (Eds.) Pennsylvania State University Press. 271-314.

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

8 Greenacre, M. (1984) *Theory and applications of correspondence analysis*. Academic Press, London. 1-364.

9 Rijckevorsel, J. van. (1987) *The application of fuzzy coding and hoerseshoes in multiple correspondence analysis*. DSWO Press, Leiden. 1-272.

10 Guinochet, M. (1973) *Phytosociologie*. Masson, Paris. 1-228.

11 Lévèque, C. & Gaborit, M. (1972) Utilisation de l'analyse factorielle des correspondances pour l'étude des peuplements en Mollusques benthiques du lac Tchad. *Cahiers ORSTOM, Série Hydrobiologie* : 4, 1, 47-66.

12 Lebreton, J.D. (1973) Etude des déplacements saisonniers des Sarcelles d'hiver, Anas c. crecca L., hivernant en Camargue à l'aide de l'analyse factorielle des correspondances. *Compte rendu hebdomadaire des séances de l'Académie des sciences*. Paris, D : III, 277, 2417-2420.

13 Ibanez, F. & Seguin, G. (1972) Etude du cycle annuel du zooplancton d'Abidjan. Comparaison de plusieurs méthodes d'analyse multivariable. *Investigacion pesquera* : 36, 81-108.

14 Dessier, A. & Laurec, A. (1978) Le cycle annuel du zooplancton à Pointe-Noire (RP Congo). Description mathématique. *Oceanologica acta* : 1, 3, 285-304.

15 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.

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

17 Bates, J.W. & Brown, D.H. (1981) Epiphyte differentiation between Quercus petraea and Fraxinus excelsior trees in a maritime area of South West England. *Vegetatio* : 48, 61-70.

18 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. *Oïkos* : 37, 21-38.

19 Feoli, E. & Orloci, L. (1979) Analysis of concentration and detection of underlying factors in structured tables. *Vegetatio* : 40, 49-54.

20 Williams, E.J. (1952) Use of scores for the analysis of association in contingency tables. *Biometrika* : 39, 274-289.

21 Noy-Meir, I. (1973) Data transformations in ecological ordination. I. Some advantages of non-centering. *Journal of Ecology* : 61,329-341.

22 Benzecri, J.P. (1969) Statistical analysis as a tool to make patterns emerge from data. In : *Methodologies of pattern recognition*. Watanabe, S. (Ed.) Academic Press, New-York. 35-60.

23 Escoufier, Y. (1982) L'analyse des tableaux de contingence simples et multiples. *Metron* : 40, 53-77. Escoufier, Y. (1985) L'analyse des correspondances : ses propriétés et ses extensions. In : *Proceedings 45th session. Institut International de la Statistique*. Amsterdam. 28.2.1-28.2.16. 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.

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

25 Legay, J.M. & Pontier, D. (1985) Relation âge-fécondité dans les populations de Chats domestiques, Felis catus. *Mammalia* : 49, 395-402.

26 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. *Oïkos* : 37, 21-38.

27 Digby, P. G. N. & Kempton, R. A. . (1987) *Multivariate Analysis of Ecological Communities*. Chapman and Hall, Population and Community Biology Series, London. 1-205. p. 96.

28 Vallet, L.A. (1986) Activité professionnelle de la femme mariée et détermination de la position sociale de la famille. Un test empirique : la France entre 1962 et 1982. *Revue Française de Sociologie* : 27, 656-696.

29 Hoffmann, L. (1960) Untersuchungen an Enten in der Camargue. *Ornithologischer Beobatcher* : 57, 35-50.

30 Tatibouet, F., Chessel, D., Broyer, J. & Lebreton, J.D. (1980) Etude des peuplements d'oiseaux nicheurs de la zone urbaine de Lyon. Rapport final du Contrat Ecologie urbaine n° 237-01-78-00314. Ministère de l'Environnement. 106-156.