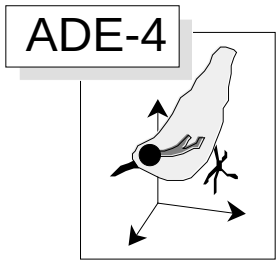


Analyse des correspondances avec colonne de référence

Résumé

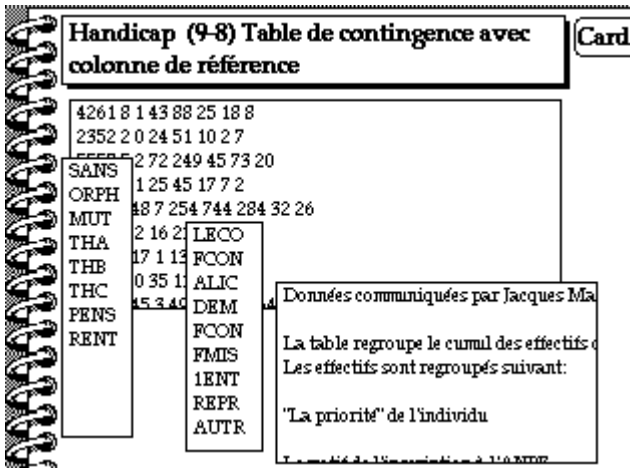
Quand une table de contingence contient une colonne de poids très élevé, cette colonne peut servir de point de référence. La distribution associée à la colonne de référence définit le poids des lignes, l'origine dans l'ensemble des profils colonnes et la métrique du Khi2 dans cet espace. L'inertie est alors une somme de Khi2 d'ajustement. La fiche donne donc un exemple d'analyse des correspondances sur modèles de B. Escofier (Analyse factorielle en référence à un modèle. Applications à l'analyse d'un tableau d'échanges. *Revue de Statistique Appliquée* : 32, 4, 25-36, 1984).

Plan

D. Chessel

1 — Définition du problème

La question ici traitée est posée par Jacques Malgras dans la carte de données Handicap :



Le tableau est une table de contingence qui regroupe le cumul des effectifs d'inscription à l'ANPE pendant l'année 1997 sur l'ensemble de la région Auvergne. (Allier+Creuse+Puy de Dôme+Haute Loire). En ligne, on trouve les motifs d'inscription à l'ANPE qui sont au nombre de 9 :

1	LECO	Licenciement économique hors fin de convention de conversion
2	FCON	Fin de convention de conversion
3	ALIC	Autres licenciements
4	DEM	Démission
5	FCON	Fin contrat
6	FMIS	Fin mission
7	1ENT	Première entrée
8	REPR	Reprise d'activité
9	AUTR	Autres cas

En colonne, est indiqué le statut de la personne par rapport à la notion de handicap. On a 8 modalités caractérisant la priorité accordée à cette personne, à savoir :

1	SANS	Sans priorité (valide)
2	ORPH	Orphelin
3	MUT	Mutilé de guerre
4	THA	Travailleur handicapé de catégorie A
5	THB	Travailleur handicapé de catégorie B
6	THC	Travailleur handicapé de catégorie C
7	PENS	Pension d'invalidité
8	RENT	Rente accident du travail

L'objectif est de comparer les populations handicapée et valide vis à vis de l'emploi (ou plutôt du non emploi) et de répondre aux questions

"y a t'il des différences entre les deux populations ? si oui, quels couples de modalités ont un comportement franchement différent ?

L'objectif, à terme, est de déterminer les catégories de travailleurs prioritaires pour lesquelles une action spécifique devra être menée (information, formation, aides ...)

Les marges du tableau ont une propriété très particulière associée au poids dominant d'une colonne. En effet, les personnes participant à l'enquête sont au nombre de

1	SANS	111075
2	ORPH	136
3	MUT	17
4	THA	972
5	THB	2892
6	THC	850
7	PENS	295
8	RENT	103

L'énorme différence entre poids marginaux pose des questions pour l'utilisation de l'analyse des correspondances mise en évidence dans une note de Lebart (1979)[1]. Dans l'exemple donné par L. Lebart, une colonne est formée de la répartition de 10 479 000 personnes dans 88 département, chaque autre colonne donnant le nombre de personnes décédées pour une cause donnée dans chacun de ces mêmes départements.

Dans les deux cas, le tableau contient une colonne de référence qui doit servir de centre à l'analyse. On a besoin d'une analyse des correspondances modifiée à centre imposé dans la même idée que celle de PCA : Decentring X[i,j] - Model[j] pour l'ACP.

2 — AFC à centre imposé

On peut considérer l'analyse suivante. Bien que le tableau traité contienne l'information en un seul bloc, au niveau conceptuel, on séparera ce qui revient à la colonne de référence et ce qui touche au tableau des observations elles-mêmes. Notons, comme d'habitude **N** la tableau de contingence *ne contenant pas la colonne de référence*, **P** la table de fréquence associée, I et J le nombres de lignes et de colonnes, $\mathbf{D}_I$ et $\mathbf{D}_J$ les diagonales des pondération marginales. La colonne de référence donne elle-même une distribution de fréquences à I composantes rangées dans une matrice diagonale **D** (trace unité, $\mathbf{D} = Diag(r_1, \quad , r_i, \quad r_I)$).

L'écart entre la distribution de fréquence associée à une colonne de **N** et la distribution de fréquence associée à la colonne de référence se mesure et se teste par un Khi2 d'ajustement :

$$\chi^2_j = \sum_{i=1}^{I} \frac{\left(n_{ij} - n_{.j} r_i\right)^2}{n_{.j} r_i} = n_{..} p_{.j} \sum_{i=1}^{I} \frac{\left(\frac{n_{ij}}{n_{.j}} - r_i\right)^2}{r_i}$$

Au facteur $n_{..}$ près, cette valeur est l'inertie associée à la colonne j de l'analyse du triplet :

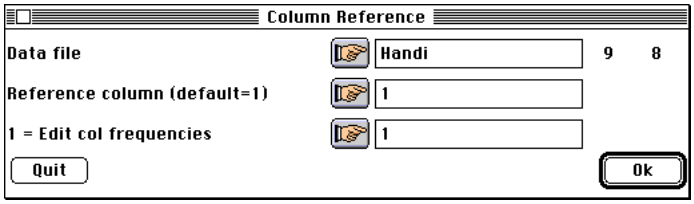
$$\left(\left[p_{i/j} - r_i\right], \mathbf{D}_J, \mathbf{D}^{-1}\right)$$

Un triplet équivalent s'écrit $\left(\mathbf{D}^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}, \mathbf{D}_J, \mathbf{D}\right)$ (voir la plasticité des schémas de l'AFC dans [2]).

Les composantes principales sont des scores des lignes **D**-normés (centrés, réduits et non corrélés pour la pondération de référence car $\left(\mathbf{D}^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{IJ}\right)^t \mathbf{D}\mathbf{1}_I = \mathbf{0}_J$)

maximisant la moyenne (au sens de $\mathbf{D}_J$) des carrés des écarts entre l'origine et la moyenne par colonne de **N**.

L'option COA : Column Reference permet de faire les calculs.




```
rq/COA with column reference
Input file: Handi
Number of rows: 9, columns: 8
Column reference: 1
Total:       5265

Column profiles (unit = 1/10000)
---|-----|-----|-----|-----|-----|-----|-----|-----|
num|*REF*|    2|    3|    4|    5|    6|    7|    8|
---|-----|-----|-----|-----|-----|-----|-----|-----|
  1|  384|  588|  588|  442|  304|  294|  610|  777|
  2|  212|  147|    0|  247|  176|  118|   68|  680|
  3|  500|  368| 1176|  741|  861|  529| 2475| 1942|
  4|  367|   74|  588|  257|  156|  200|  237|  194|
  5| 3253| 3529| 4118| 2613| 2573| 3341| 1085| 2524|
  6|  257|  294| 1176|  165|   80|    0|  102|   97|
  7| 1279| 1250|  588|  134|  277|  729|  305|   97|
  8|  349|  441|    0|  360|  467|  506|  847|  485|
  9| 3398| 3309| 1765| 5041| 5107| 4282| 4271| 3204|
---|-----|-----|-----|-----|-----|-----|-----|-----|
Fre|    0|  258|   32| 1846| 5493| 1614|  560|  196|
---|-----|-----|-----|-----|-----|-----|-----|-----|
```

Le programme édite les profils par colonne en 1 pour dix mille. Il y a 32.53% de fin de contrats dans la population de référence (personnes valides) et 26.13% dans l'ensemble des travailleur handicapé de catégorie A (colonne 4).

Le triplets annoncé est ensuite constitué :

```
File Handi.rqpl contains the margin distribution of rows
It has 9 rows and 1 column
Frequency distribution from column reference

File Handi.rqpc contains the margin distribution of columns (without
reference)
It has 8 rows and 1 column

File Handi.rqta contains the table [P(j/i)/R(j)-1]
It has 9 rows and 8 column
```

Les khi2 sont calculés par colonne :

```
Col:1 Khi2 =  0    DDL = 8 Proba =     1 Trace Cont. =         0 référence
Col:2 Khi2 = 6.187 DDL = 8 Proba = 0.628 Trace Cont. =   0.00445
Col:3 Khi2 = 10.86 DDL = 8 Proba = 0.209 Trace Cont. =    0.0078
Col:4 Khi2 = 208.3 DDL = 8 Proba =     0 Trace Cont. =      0.15
Col:5 Khi2 = 680.7 DDL = 8 Proba =     0 Trace Cont. =     0.489
Col:6 Khi2 = 79.59 DDL = 8 Proba =     0 Trace Cont. =    0.0572
Col:7 Khi2 = 332.9 DDL = 8 Proba =     0 Trace Cont. =     0.239
Col:8 Khi2 = 73.03 DDL = 8 Proba =     0 Trace Cont. =    0.0525
```

On peut vérifier les calculs sur la colonne 4 :

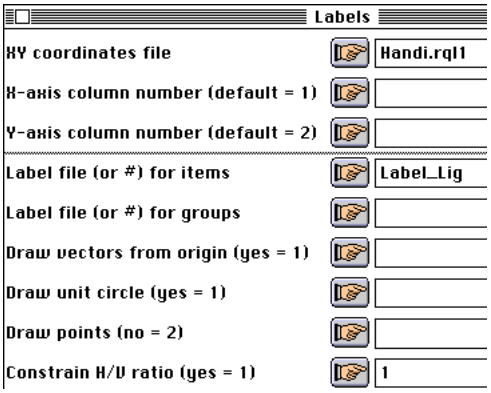
Obs	Proba	Theo	Obs-Theo	khi2
43	0.0384	37.3248	5.6752	.863
24	0.0212	20.6064	3.3936	.559
72	0.05	48.6	23.4	11.267
25	0.0367	35.6724	-10.6724	3.193
254	0.3253	316.1916	-62.1916	12.232
16	0.0257	24.9804	-8.9804	3.228
13	0.1279	124.3188	-111.3188	99.678
35	0.0349	33.9228	1.0772	.034
490	0.3398	330.2856	159.7144	77.232
Tot=972				Tot=208.287

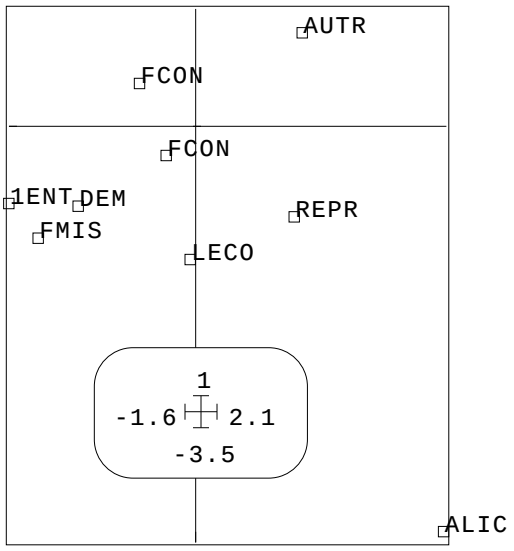
Les profils des classes de personnes handicapées sont très sensiblement différentes du profil des personnes valides servant de référence. On n'accordera aucune importance à la position des classes 2 et 3 sans signification. L'analyse du triplet est effectuée :

```
-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Handi.rqta
--- Number of rows: 9, columns: 8
-----------------------
Total inertia: 0.264296
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.1296E-01 +0.8057 +0.8057   |02   +3.8362E-02 +0.1451 +0.9509 |
03   +6.9374E-03 +0.0262 +0.9771   |04   +3.2972E-03 +0.0125 +0.9896 |
05   +1.4436E-03 +0.0055 +0.9951   |06   +7.4172E-04 +0.0028 +0.9979 |
07   +5.5819E-04 +0.0021 +1.0000   |08   +0.0000E+00 +0.0000 +1.0000 |

File Handi.rqvp contains the eigenvalues and relative inertia for each
axis
--- It has 8 rows and 2 columns

File Handi.rqco contains the column scores
--- It has 8 rows and 2 columns
File :Handi.rqco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.086e-01| 8.263e-01|
|   2|-6.572e-01| 1.346e-01|
|----|----------|----------|

File Handi.rqli contains the row scores
--- It has 9 rows and 2 columns
File :Handi.rqli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-7.333e-01| 9.590e-01|
|   2|-6.709e-01| 1.526e-01|
|----|----------|----------|

File Handi.rql1 contains the row scores with unit norm
It has 9 rows and 2 columns
File :Handi.rql1
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.589e+00| 2.078e+00|
|   2|-3.425e+00| 7.789e-01|
|----|----------|----------|

File Handi.rqc1 contains the column scores with unit norm
It has 8 rows and 2 columns
File :Handi.rqc1
```

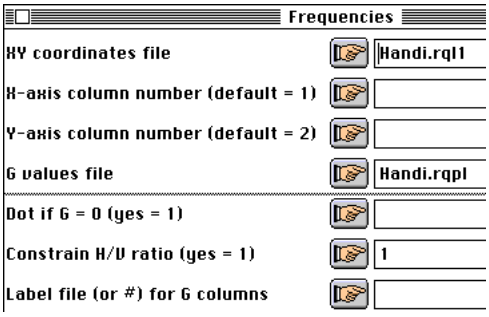
Col.	Mini	Maxi
1	-2.352e-01	1.791e+00
2	-3.355e+00	6.872e-01

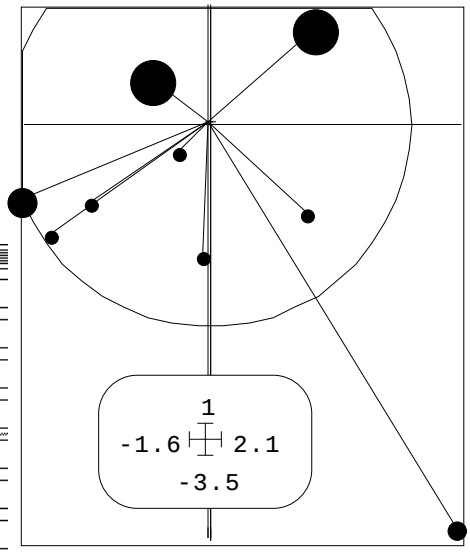
On utilisera essentiellement les notion d'averaging pour dépouiller. Positionner les lignes avec leurs scores normalisés :



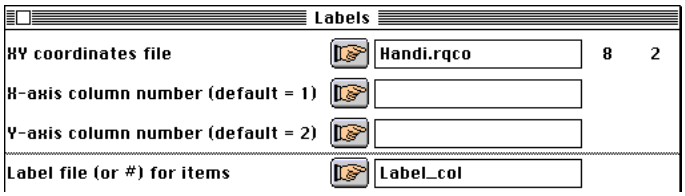


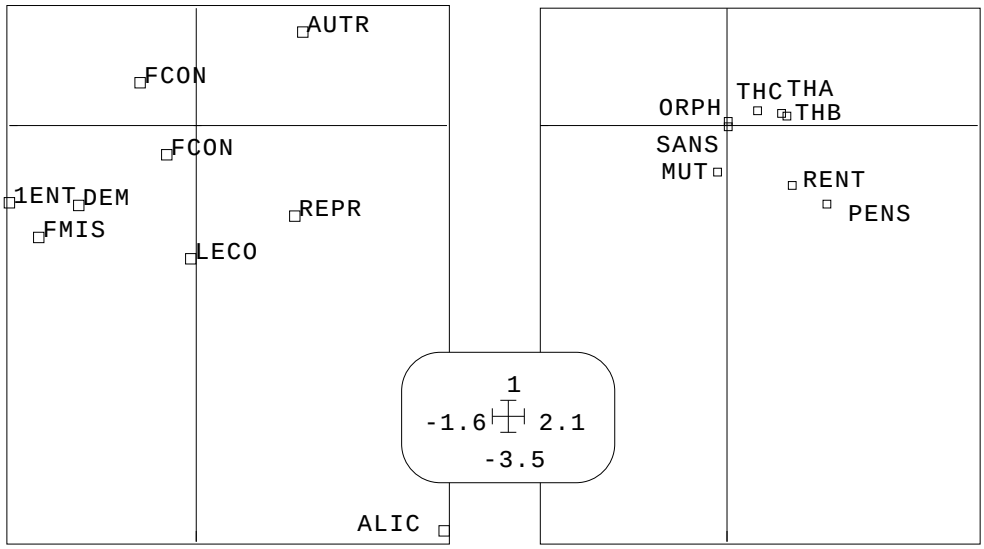
Vérifier que ces scores sont centrés et orthogonaux pour la pondération des lignes (fréquences de référence) par ScatterDistri : Frequencies, Stars, Ellipses :





Positionner dessus les colonnes par averaging (centres de gravités des distributions de fréquence par colonnes) :





La lecture simultanée des deux représentations attire l'attention sur deux faits majeurs.

La sur-représentation de la catégorie "autres causes d'inscription" est manifeste pour les travailleurs handicapés (THA 50%, THB 51%, THC 43%) et les pensionnés d'invalidité (PENS 42%) contre 34% dans la population de référence, ce qui laisse tout de suite penser à des interactions de codage qui pourrait être réduites.

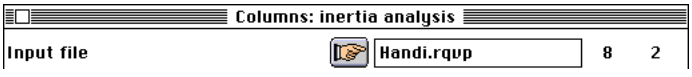
Rentiers d'accidents et pensionnés d'invalidité sont d'autre part associés à la catégorie "autres licenciements" dans des proportions très supérieures à la moyenne (RENT 25%, PENS 19% contre 5% en général).

La distribution de référence à l'origine permet donc la lecture directe de tous les écarts sans difficultés.

3 — Une AFC sur modèles

On notera pour conclure que l'analyse proposée ici appartient à la classe définie pat B. Escofier [3]. Cette classe très large de triplet s'écrit $\left(\mathbf{D}_I^{-1}(\mathbf{P}-\mathbf{M})\mathbf{D}_J^{-1}, \mathbf{D}_J, \mathbf{D}_I\right)$ où tous les termes peuvent être indépendants. On a seulement des relations d'averaging quand les pondérations marginales sont en relation avec les tableaux et l'interprétation n'est simple que si les tableaux (**P** pour les observations et **M** pour le modèle) ont les mêmes marges. On retrouve l'AFC avec $\mathbf{M} = \mathbf{D}_I \mathbf{1}_{IJ} \mathbf{D}_J$ (modèle $p_{i.}p_{.j}$ pour p_{ij}).

Dans le cas présent, une des marges est en commun avec le modèle $\mathbf{M} = \mathbf{D}\mathbf{1}_{IJ}\mathbf{D}_J$ pour le schéma $\left(\mathbf{D}(\mathbf{P}-\mathbf{M})\mathbf{D}_J^{-1}, \mathbf{D}_J, \mathbf{D}\right)$ et on a une relation d'averaging. L'analyse d'inertie du nuage des colonnes (qui redonne les khi2 d'ajustement est ainsi parfaitement justifiée :



On retrouve une partie des résultats déjà édités :

```
Input file: Handi.rqta
Number of rows: 9, columns: 8
```

```
Inertia: Two diagonal norm inertia analysis
Total inertia: 0.264296 - Number of axes: 2

File Handi.rqcc contains the contribution of columns to the trace
It has 8 rows and 1 column

Column inertia
All contributions are in 1/10000

----------Absolute contributions----------
|Num  |Fac 1|Fac 2|
|    1|    0|    0|
|    2|    0|    9|
|    3|    1|  119|
|    4| 1633|  542|
|    5| 5877| 1000|
|    6|  422|  762|
|    7| 1796| 6308|
|    8|  268| 1256|

----------Relative contributions----------
|Num  |Fac 1|Fac 2||Remains| Weight | Cont.|
|    1|    0|    0||10000  |    0   |10000 |
|    2|   30|  320|| 9648  |  258   |   44 |
|    3|  184| 2218|| 7596  |   32   |   78 |
|    4| 8792|  526||  681  | 1846   | 1496 |
|    5| 9681|  296||   21  | 5492   | 4891 |
|    6| 5947| 1934|| 2118  | 1614   |  571 |
|    7| 6051| 3827||  120  |  560   | 2392 |
|    8| 4123| 3475|| 2400  |  195   |  524 |
```

On retrouve les principaux éléments pour l'interprétation. Notons enfin que la situation expérimentale ici étudiée n'est pas celle de l'AFC décentrée définie dans [4] mais ceci montre que le modèle général du schéma de dualité permet de s'adapter à chaque cas précis.

Références

[1] Lebart, L. (1979) Exemple d'analyse des correspondances d'un tableau dont l'une des colonnes a un poids prédominant. *Les Cahiers de l'Analyse des Données* : 4, 417-422.

[2] 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.

[3] Escofier, B. (1984) Analyse factorielle en référence à un modèle. Applications à l'analyse d'un tableau d'échanges. *Revue de Statistique Appliquée* : 32, 4, 25-36.

[4] Dolédec, S., Chessel, D. & Olivier, J.M. (1995) L'analyse des correspondances décentrée: application aux peuplements ichtyologiques du haut-Rhône. *Bulletin Français de la Pêche et de la Pisciculture* : 336, 29-40.