DEA Analyse et Modélisation des Systèmes Biologiques
Module 01 - 2001/2002

1. Tableau 2-4

Soit un tableau à 2 lignes et 4 colonnes :

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

1.1. Donner ses pondérations marginales.

1.2. Donner les valeurs propres de l'analyse des correspondances de **A**.

2. Essais avec n=p

Un expérimentateur avisé désire se faire une opinion personnelle du comportement de l'analyse des correspondances sur des tableaux artificiels. Il considère 3 tableaux comportant n = 16 lignes et p = 16 colonnes. Le premier est appelé **talea** car il a été généré par une procédure de tirage aléatoire.

```
> talea
   a b c d e f g h i j k l m n o p
1  1 0 0 1 1 0 1 0 0 0 0 0 0 0 1 1
2  1 1 1 1 1 0 1 0 0 1 1 0 1 1 1 1
3  1 1 1 1 0 0 0 1 0 0 1 0 0 1 1 0
4  1 1 1 0 0 1 0 1 0 1 0 0 0 1 1 0
5  1 0 0 0 0 0 1 0 0 0 0 0 1 1 0 0
6  1 0 0 0 0 0 1 0 1 0 1 1 1 0 1 1
7  1 0 1 1 0 0 0 1 1 0 0 0 1 0 1 1
8  0 1 0 0 1 1 1 1 0 0 0 0 0 1 1 1
9  0 0 1 0 0 0 0 0 1 1 1 0 1 1 0 0 1
10 0 1 0 0 1 1 0 1 0 1 1 1 1 1 1 0
11 0 0 1 0 1 0 0 0 1 0 1 0 0 0 0 1
12 0 0 1 0 1 0 1 0 0 1 0 1 0 1 1 1
13 1 1 1 0 1 1 0 0 1 0 0 0 1 0 0 0
14 0 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1
15 1 1 1 1 0 1 1 1 0 0 0 0 1 1 0 0
16 0 0 1 1 1 0 1 0 0 0 1 0 1 1 0 1
```

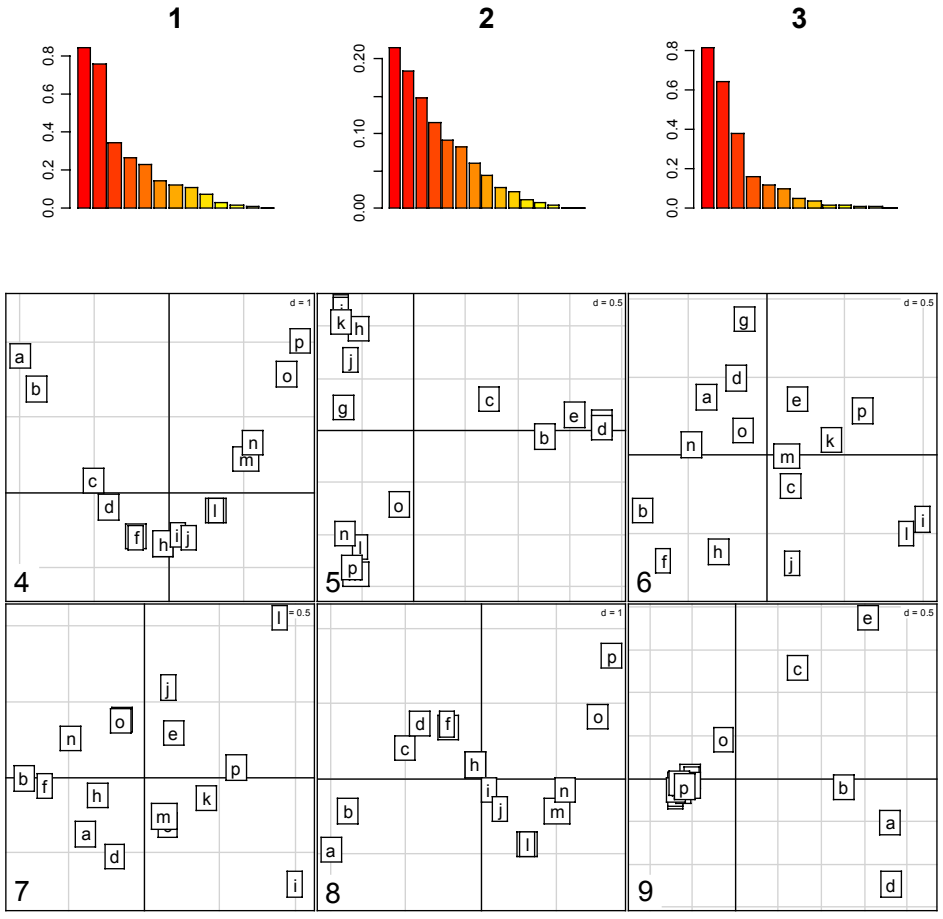
Le second est appelé **tgrad** car il représente une structure simple définie par un gradient.

```
> tgrad
   a b c d e f g h i j k l m n o p
1  1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2  1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
3  1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
4  0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
5  0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0
6  0 0 1 1 0 1 1 1 1 0 0 0 0 0 0 0
7  0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0
8  0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0
9  0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0
10 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0
11 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 0
12 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0
13 0 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0
14 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0
15 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
16 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1
```

Le troisième est appelé **tparti** car il représente une structure simple définie par une partition.

```
> tparti
   a b c d e f g h i j k l m n o p
1  0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
2  1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
3  0 1 1 0 0 0 0 1 0 0 0 0 0 0 1 0
4  1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
5  0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0
6  0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
7  0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 0
8  0 0 1 0 0 1 1 1 1 1 1 0 0 1 0 0
9  0 0 0 0 0 0 1 0 1 1 0 1 0 0 0 0
10 0 0 0 0 0 0 0 1 1 0 1 1 0 0 0 0
11 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0
12 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1
13 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1
14 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1
15 0 1 0 0 0 0 0 0 0 0 1 0 1 1 0 1 1
16 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1
```

On pourra penser que les tableaux simulent la présence-absence de 16 espèces (colonnes) sur un échantillon de 16 sites (lignes). Pour les trois analyses des correspondances on garde, dans un ordre arbitraire, le graphe des valeurs propres, la carte f1-f2 des colonnes et la carte f1-f3 des colonnes :



2.1. Quelles sont les trois figures qui appartiennent à l'analyse de talea) ?

2.2. Quelles sont les trois figures qui appartiennent à l'analyse de tparti) ?

2.3. Quelles sont les trois figures qui appartiennent à l'analyse de tgrad) ?

2.4. Laquelle de ces trois analyses invalide l'assertion "quand sont légitimement conservés deux axes, il existe deux faits marquants dans les données" ?

2.5. Commenter l'assertion "si on avait utilisé les tableaux transposés au lieu des tableaux initiaux on aurait obtenu strictement les mêmes résultats numériques".

3. Expression graphique

Une enquête d'opinion (période 1970-1980) portant sur 1000 personnes contient entre autres la question "A quelle famille politique vous rattachez vous ?" avec 5 modalités de réponse (Extrême gauche ; gauche ; centre ; droite et extrême droite) et la question " Que pensez vous de l'importance des syndicats ?" avec 4 modalités de réponse (importance trop grande ; importance convenable; importance insuffisante ; sans opinion). La répartition des réponses est consignée dans la table de contingence :

```
          trop convenable insuffisante sans_opinion
ExtGauche   10         31           85            4
Gauche      80         90          118           52
Centre      55         58           31           36
Droite      95         38           46           71
ExtDroite   40         23           10           27
```

L'analyse des correspondances de ce tableau donne des valeurs propres, des coordonnées des lignes et des coordonnées des colonnes :

```
$eig
[1] 1.614e-01 1.747e-02 7.105e-05

$li
            Axis1    Axis2
ExtGauche -0.8543  0.08628
Gauche    -0.1622 -0.04255
Centre     0.1836 -0.21241
Droite     0.3498  0.17555
ExtDroite  0.4570 -0.02403

$co
                Comp1    Comp2
trop          0.35645  0.04033
convenable   -0.06875 -0.23193
insuffisante -0.56140  0.09297
sans.opinion  0.41841  0.09163
```

3.1. Exprimer les résultats de l'analyse par une figure et sa légende.

3.2. Que suggère cette figure sur la catégorie "sans opinion".

4. Moyennes conditionelles

Soit une table de contingence **K** "obtenue en ventilant une population de 592 femmes suivant la couleur des yeux et la couleur des cheveux" (Lebart, L., A. Morineau, and M. Piron. 1995. Statistique exploratoire multidimensionnelle. Dunod, Paris. p. 68) :

```
         brun châtain roux blond
marron     68     119   26     7
noisette   15      54   14    10
vert        5      29   14    16
bleu       20      84   17    94
```

Le tableau des profiles-lignes est :

```
            brun châtain    roux   blond
marron   0.30909  0.5409 0.11818 0.03182
noisette 0.16129  0.5806 0.15054 0.10753
vert     0.07813  0.4531 0.21875 0.25000
bleu     0.09302  0.3907 0.07907 0.43721
```

L'analyse des correspondances de **K** donne des coordonnées des lignes

```
           Axis1    Axis2
marron   -0.4922  0.08832
noisette -0.2126 -0.16739
vert      0.1618 -0.33904
bleu      0.5474  0.08295
```

des coordonnées des colonnes

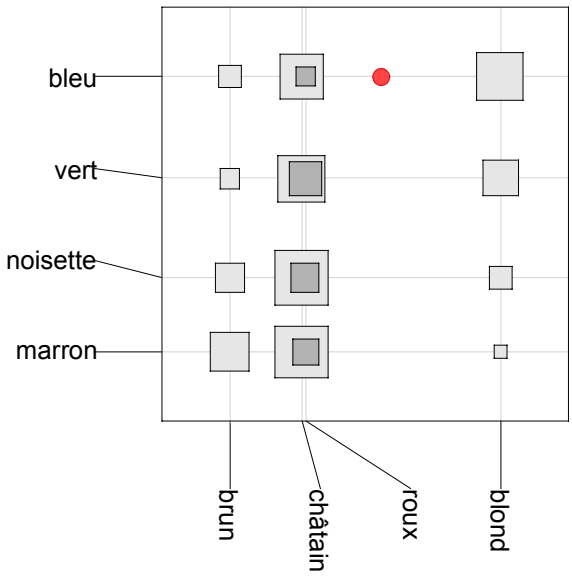
```
           Comp1    Comp2
brun     -0.5046  0.21482
châtain  -0.1483 -0.03267
roux     -0.1295 -0.31964
blond     0.8353  0.06958
```

et des valeurs propres :

```
0.208773 0.022227 0.002598
```

Sur le graphique les lignes et les colonnes sont respectivement positionnées par les coordonnées sur le premier axe. Les pointillés indiquent l'origine. Les carrés représentent les profils lignes. Le point donne la moyenne du score des colonnes pour la distribution de la ligne correspondante.

4.1. Placer sur la figure les points moyens du score des colonnes pour les autres distributions conditionnelles par ligne.



5. Comparaison

Dans une enquête de sociologie (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.) on a retenu 4365 mariages comportant deux époux salariés. En utilisant les catégories socio-professionnelles cadre supérieur, cadre moyen, employé, ouvrier et personnel de service (nomenclature de 1982) on a :

```
> m1
            H.Cadre_Sup H.Cadre_moy H.Employ H.Ouvrier H.Service
F.Cadre_Sup         158          53       18        24         2
F.Cadre_moy         201         309      113       225        14
F.Employ            159         323      348       756        34
F.Ouvrier            24          79      118       795        22
F.Service            21          61       82       382        44
```

L'analyse des correspondances avec la fonction corresp de R donne directement :

```
> corresp(m1)
First canonical correlation(s): 0.5051

 Row scores:
F.Cadre_Sup F.Cadre_moy    F.Employ   F.Ouvrier   F.Service
   -2.73503    -1.08275     0.06649     1.00773     0.80851

 Column scores:
H.Cadre_Sup H.Cadre_moy    H.Employ   H.Ouvrier   H.Service
    -2.1028     -0.7897      0.1072      0.7721      0.6720
```

L'analyse des correspondances avec la fonction COA d'ade4 donne :

```
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +2.5516E-01 +0.7957 +0.7957   |02   +4.4023E-02 +0.1373 +0.9329 |
03   +1.1230E-02 +0.0350 +0.9680   |04   +1.0273E-02 +0.0320 +1.0000 |
05   +0.0000E+00 +0.0000 +1.0000

------------------------------------------------
Binary input file: D:\Ade4\Dir_Try\Mariages\m1.fcli - 5 rows, 1 cols.
   1 | -1.3816
   2 | -0.5469
   3 |  0.0336
   4 |  0.5090
   5 |  0.4084

------------------------------------------------
Binary input file: D:\Ade4\Dir_Try\Mariages\m1.fcco - 5 rows, 1 cols.
   1 | -1.0622
   2 | -0.3989
   3 |  0.0542
   4 |  0.3900
   5 |  0.3395
```

5.1. Expliquer en quoi ces résultats en apparence différents sont en fait cohérents.

5.2. Comment retrouve-t-on dans ade4 les scores donnés par la fonction de R ?

6. Analyse canonique

Soit un tableau faunistique (Prodon, R., and J. D. Lebreton. 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) comportant 182 relevés (lignes) et 51 espèces d'oiseaux (colonnes) . On peut utiliser le score des 51 colonnes (espèces) pour représenter la moyenne et la variance de chacune des lignes (relevés) ou le score des 182 lignes (relevés) pour représenter la moyenne et la variance de chacune des colonnes (espèces). Dans le premier cas (ci-dessous, à gauche) on exprime la diversité des sites dans le gradient défini par les espèces, dans le second (ci-dessous, à droite) on exprime l'amplitude d'habitat des espèces dans le gradient défini par les sites.

6.1. Quel principe permet de faire les deux opérations simultanément ?

6.2. Quelle procédure permet dans ade4 de faire les représentations à deux dimensions associées à ce principe ?

7. Profils alléliques (pour ceux qui préfèrent l'approche mathématique)

Le polymorphisme biochimique des bovins domestiques fait l'objet de nombreuses études. D. Laloë (INRA, Jouy-en-Josas) propose un exemple pédagogique comportant 2 races taurines africaines (Taurins N'Dama et Baoulé), 2 races de Zébus (Zébu Azawak du Niger et Zébu malgache) et 2 races bovines européennes (Charolais et Salers). Le tableau donne les fréquences alléliques de 4 systèmes génétiques ($\alpha_{s1}-\mathbf{Cn}$, $\beta-\mathbf{Cn}$, $\kappa-\mathbf{Cn}$ et $\beta-\mathbf{Lg}$) définis par le polymorphisme des protéines du lait :

	alpha		beta			kappa			beta-lacto	
Ndama	0.89	0.11	0.60	0.37	0.03	0.27	0.73	0.00	0.10	0.90
Baoule	0.92	0.08	0.63	0.36	0.01	0.34	0.64	0.02	0.09	0.91
Zebu_a	0.22	0.78	0.08	0.86	0.06	0.83	0.17	0.00	0.14	0.86
Zebu_m	0.17	0.83	0.10	0.90	0.00	0.75	0.25	0.00	0.27	0.73
Charolais	0.92	0.08	0.10	0.76	0.14	0.49	0.51	0.00	0.67	0.33
Salers	0.96	0.04	0.19	0.70	0.11	0.54	0.46	0.00	0.64	0.36

Plus généralement, les tableaux de profils alléliques sont des juxtapositions du type :

$$\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2 \mid ... \mid \mathbf{X}_k]$$

dans laquelle la somme par lignes dans chacun des sous-tableaux vaut l'unité.

7.1. Quelle est, dans ce cas, la pondération des lignes ? Quelle est la somme des éléments du tableau ?

7.2. Si $p_1, p_2, ..., p_k$ sont les nombres de colonnes par sous-tableaux, que valent les produits $\mathbf{X}_j \mathbf{1}_{p_j}$?

7.3. Dans la situation présente, quelle est la propriété particulière du tableau $\mathbf{D}_I^{-1}\mathbf{P}$ dans les notations habituelles.

7.4. En utilisant les vecteurs $\mathbf{a}_j = \left(\underbrace{0,...,0}_{p_1}, \underbrace{0,...,0}_{p_2}, ..., \underbrace{1,...,1}_{p_j}, ..., \underbrace{0,...,0}_{p_k} \right)$ montrer que les coordonnées des colonnes de l'AFC du tableau sont centrées par blocs.

7.5. Donner l'illustration de cette propriété sur l'exemple numérique :

```
Tableau traité
    V1   V2   V3   V4   V5   V6   V7   V8   V9  V10
1 0.89 0.11 0.60 0.37 0.03 0.27 0.73 0.00 0.10 0.90
2 0.92 0.08 0.63 0.36 0.01 0.34 0.64 0.02 0.09 0.91
3 0.22 0.78 0.08 0.86 0.06 0.83 0.17 0.00 0.14 0.86
4 0.17 0.83 0.10 0.90 0.00 0.75 0.25 0.00 0.27 0.73
5 0.92 0.08 0.10 0.76 0.14 0.49 0.51 0.00 0.67 0.33
6 0.96 0.04 0.19 0.70 0.11 0.54 0.46 0.00 0.64 0.36

Poids des colonnes multiplié par k
      V1       V2       V3       V4       V5       V6       V7       V8
0.680000 0.320000 0.283333 0.658333 0.058333 0.536667 0.460000 0.003333
      V9      V10
0.318333 0.681667
```

Coordonnées des colonnes

	Comp1	Comp2
V1	-0.4910	0.11741
V2	1.0433	-0.24949
V3	-0.6078	-0.57284
V4	0.2784	0.17688
V5	-0.1891	0.78612
V6	0.3572	0.08638
V7	-0.4099	-0.09277
V8	-0.9432	-1.10472
V9	-0.1058	0.75786
V10	0.0494	-0.35392

Solution

1. Tableau 2-4

1.1. $\left(\frac{1}{2},\frac{1}{2}\right)$ *et* $\left(\frac{1}{4},\frac{1}{2},\frac{1}{4},0\right)$

1.2.

L'AFC donne des scores normés maximisant la variance des moyennes conditionnelles. Il n'y a avec 2 lignes qu'une seule façon de fabriquer un score centré réduit des lignes. Pour le score normé $(-1,1)$ *des lignes, les moyennes conditionnelles sont* $(-1,0,1,0)$. *Elles sont centrées pour la pondération des lignes et de variance* $\frac{1}{4}+\frac{1}{4}=0.5=\lambda_1$. *L'autre valeur propre est toujours nulle.*

2. Essais avec n=p

2.1.

La carte 4 est structurée comme un gradient de même que la 8. La carte 5 est structurée par une partition. Le graphe 2 est sans structure. Les graphes de l'analyse de talea sont donc 2, 6 et 7.

2.2.

La carte 5 est structurée par une partition. Il y a donc au moins deux valeurs propres ayant du sens. La carte 8 est celle de tgrad (effet Guttman), la carte 7 est celle de talea, la carte 9 est donc la 1-3 de tparti. Elle ne répète pas la partition. On peut donc prévoir que les graphes 1, 5 et 9 sont ceux de l'analyse de tparti.

2.3.

Les graphes 3, 4 et 8 sont celles de tgrad.

2.4.

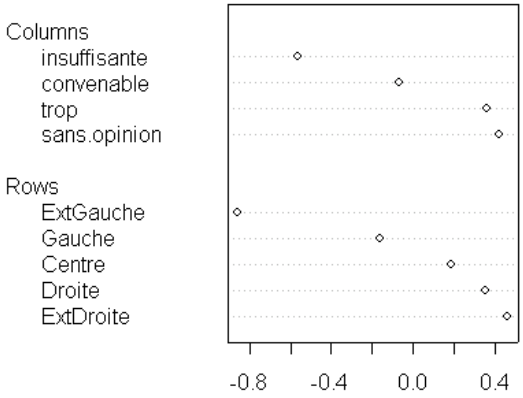
L'analyse de tgrad donne comme facteur 1 l'ordre du gradient, comme facteur 2 un polynôme de degré 2 et comme facteur 3 un polynôme de degré 3. On peut isoler 3 facteurs qui ne n'explicite qu'un seul fait (ordination). (Voir Jackson, D. A., and K. M. Somers. 1991. Putting things in order: the ups and downs of detrended correspondence analysis. The American Naturalist ***137****:707-712. Peet, R. K., R. G. Knox, J. S. Case, and R. B. Allen. 1988. Putting things in order : the advantages of detrended correspondence analysis. The American Naturalist* ***131****:924-934. Wartenberg, D., S. Ferson, and F. J. Ohlf. 1987. Putting things in order : a critique of detrended correspondence analysis. The American Naturalist* ***129****:434-448.)*

2.5.

Elle est parfaitement exacte, l'AFC étant totalement symétrique dans la notion lignes-colonnes.

3. Expression graphique

3.1.



Représentation des lignes et des colonnes du tableau par les coordonnées factorielles sur l'axe 1. Les scores utilisés maximisent la corrélation au sens de la table de contingence. L'ordre droite-gauche est celui de l'opinion négative-positive envers les syndicats.

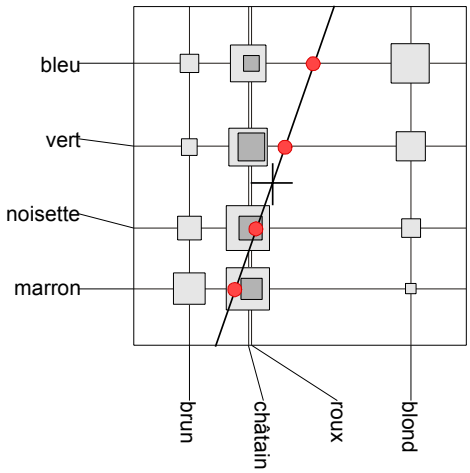
3.2.

Le profil politique des sans opinions sur les syndicats est le plus à droite des quatre. On peut dire que l'absence d'opinion est une opinion déguisée ou qu'une partie de la droite n'a pas d'opinion sur les syndicats.

4. Moyennes conditionelles

4.1.

Quand on utilise les scores de l'AFC, les régressions sont doublement linéaires. Les moyennes conditionnelles sont sur les droites de régression, lesquelles passent par l'origine (double centrage). Il suffit de tracer la droite et on obtient directement les points.



5. Comparaison

5.1.

La première corrélation canonique (R) est la racine carrée de la première valeur propre (ade4). Les scores dans R sont de variance 1, ceux de ade4 sont de variance λ. Ils ne diffère que d'une constante multiplicative ($\sqrt{\lambda}$).

5.2. Comment retrouve-t-on dans ade4 les scores donnés par la fonction de R ?

L'option Add normed scores du module DDUtil est faite pour cela.

```
*-------------------------------------------------------------*
| DDUtil: Add normed scores                 26/02/02  15/12  |
*-------------------------------------------------------------*
File D:\Ade4\Dir_Try\Mariages\m1.fcl1 contains the row scores with unit norm
It has 5 rows and 1 columns
File :D:\Ade4\Dir_Try\Mariages\m1.fcl1
|Col.|   Mini    |   Maxi    |
|----|-----------|-----------|
|   1|-2.735e+00| 1.008e+00|
|----|-----------|-----------|

File D:\Ade4\Dir_Try\Mariages\m1.fcc1 contains the column scores with unit norm
It has 5 rows and 1 columns
File :D:\Ade4\Dir_Try\Mariages\m1.fcc1
|Col.|   Mini    |   Maxi    |
|----|-----------|-----------|
|   1|-2.103e+00| 7.721e-01|
|----|-----------|-----------|
```

6. Analyse canonique

6.1.

L'analyse canonique du paquet des indicatrices des classes lignes et du paquet des indicatrices des classes colonnes donnent un score normalisé des correspondances (cases non nulles du tableau) sur lesquels on peut placer les moyennes et variances des groupes de correspondances associées soit à la même ligne soit à la même colonne.

6.2.

L'option COA: Reciprocal scaling donne les scores des correspondances et leur poids (p_{ij}) et met en place les indicatrices des classes par lignes et colonnes. Les options ScatterClass: Stars ou ScatterClass: Ellipses ou ScatterClass: Convex hulls permettent de représenter les moyennes et les variances (amplitude-diversité) à la même échelle pour les deux types d'objets.

7. Profils alléliques (pour ceux qui préfèrent l'approche mathématique)

7.1.

La somme par ligne vaut k (la somme par sous-tableau valant 1). La somme totale vaut kI (I est le nombre de lignes). La pondération marginale des lignes est uniforme.

7.2.

La somme par lignes des sous-tableaux valant 1, $\mathbf{X}_j \mathbf{1}_{p_j} = \mathbf{1}_I$.

7.3.

$\mathbf{D}_I^{-1}\mathbf{P}$ *est le tableau des profils lignes. Comme on a additionné les fréquences de k distributions, la somme d'une ligne vaut k, on divise les données brutes par k pour avoir le profil global et en re-multipliant par k on a exactement* $\mathbf{D}_I^{-1}\mathbf{P} = \mathbf{X}$.

7.4.

Le simple produit de matrices donne $\mathbf{X}\mathbf{a}_j = \mathbf{1}_I$. *Donc* $\left(\mathbf{D}_I^{-1}\mathbf{P}\mathbf{D}_J^{-1} - \mathbf{1}_{ij}\right)\mathbf{D}_J\mathbf{a}_j = \mathbf{0}_I$ *et chacun des k vecteurs* $\mathbf{a}_j$ *sont vecteurs propres pour la valeur propre 0. Ils sont indépendants et les axes principaux sont* $\mathbf{D}_J$ *aux* $\mathbf{a}_j$ *donc centrés par blocs.*

7.5.

Par exemple :

0.68*-0.4910+0.32*1.0433=0

0.283333*-0.6078+0.658333*0.2784+0.058333*-0.1891=0