Fiche TD avec le logiciel R : tdr52

Correspondances multiples

D. Chessel & A.B. Dufour

Introduction à l'analyse des correspondances multiples. Approche pratique des codages numériques. Tableaux de facteurs et tableaux disjonctifs complets. Approche procédurale et comparaison des fonctions `dudi.acm` et `mca`. Indications théoriques sur le schéma de dualité. Exemples et extensions.

Table des matières

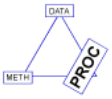
1 Introduction

Une variable qualitative ou variable nominale ou facteur (*factor*) est une mesure qui prend ses valeurs dans un ensemble d'items ou modalités ou valeurs ou niveaux (*levels*).

Une correspondance simple est la présence de deux modalités de deux facteurs différents chez le même individu.

Une correspondance multiple est la présence simultanée de k modalités de k facteurs différents chez le même individu. La description du mode d'assemblage des modalités chez les individus relève de l'analyse des correspondances simples (2 facteurs) ou multiples (plus de 2 facteurs).

On décrit ici quelques exemples de cette pratique en utilisant plusieurs librairies et en indiquant quelque repères théoriques.



2 Codages numériques

Il est d'abord utile de voir l'analyse des correspondances simples (nous dirons **AFC**) comme une pratique de codage numérique des lignes et colonnes d'un tableau de nombres positifs ou nuls. Cette définition peut paraître bien frustre mais elle a longtemps été efficace, particulièrement en écologie.

2.1 Codages de deux modalités : approche pratique

L'hebdomadaire *Telerama* publiait dans son numéro 2473 du 4 juin 1997 un article intitulé "*Dis moi qui t'informe et ...*" accompagné des résultats d'un sondage dit *Sortie des urnes* exécuté le 25 mai 1997 (voir Élections législatives françaises de 1997 sur Wikipedia).

```
sond <- read.table(url("http://pbil.univ-lyon1.fr/R/donnees/sond.txt"))
sond
```

	ExtGauche	PC	PS	Ecolo	Droite	ExtDroite	Divers
TF1	2	8	21	5	44	19	1
FR2	3	10	33	7	35	11	1
FR3	4	11	33	7	32	12	1
M6	3	11	27	10	30	16	3
La_Croix	2	3	21	6	64	2	2
Le_Figaro	1	3	9	3	73	11	0
France-Soir	4	9	27	8	32	19	1
L_Humanité	6	63	20	4	5	1	1
Libération	7	10	51	11	18	2	1
Le_Monde	5	9	36	8	34	7	1
Le_Parisien	3	11	33	6	26	20	1
Les_Echos	2	3	16	9	53	13	4
La_Tribune	2	10	27	6	39	13	3
Quotidien_régional	2	8	29	6	39	15	1
RadioFrance	3	8	29	9	39	10	2
RTL	3	10	23	5	42	16	1
Europe1	2	8	27	6	46	10	1
RMC	0	11	22	3	37	26	1
NRJ	3	8	26	7	37	17	2
Le_Canard	3	15	41	10	18	12	1
L_Evènement	2	9	44	7	30	7	1
L_Express	2	4	25	6	51	12	0
Le_Nouvel_Obs.	2	6	47	8	27	8	2
Marianne	4	9	46	11	22	5	3
Paris_Match	0	6	21	6	48	18	1
Le_Pélerin	1	3	27	6	53	10	0
Télérama	6	6	41	14	27	4	2
Le_Point	1	2	17	6	65	9	0
La_Vie	1	9	24	13	46	6	1
VSD	4	8	27	8	34	19	0

Sont utilisées 7 catégories de votes respectivement Extrême gauche, Parti Communiste, Parti Socialiste, Verts, Partis de droite, Extrême droite et Divers. Sont étudiés 30 media dont 4 chaînes de télévision, 10 quotidiens, 5 chaînes de radio et 11 hebdomadaires.

Le tableau donne des distributions de fréquences conditionnelles pour chaque media sous la forme de la distribution des votes de 100 consommateurs. Il ne s'agit évidemment pas des données initiales du sondage qui est le matériel du statisticien mais d'une image numérique de l'association entre un media et une opinion politique. Cette association n'est jamais négative. Quand elle vaut 63 pour L'Humanité/PC ou 73 pour Le Figaro/Droite tout le monde comprend de quoi il s'agit.

Il y a plusieurs procédures qui assure l'analyse des correspondances. Celle d'`amap[6]` présentait dans sa version 0.7-1 l'erreur suivante :

```
Package:        amap
Version:        0.7-1
Date:           2006-20-01
Title:          Another Multidimensional Analysis Package
Author:         Antoine Lucas
Maintainer:     Antoine Lucas <antoinelucas@libertysurf.fr>
Description:    Tools for Clustering and Principal Component Analysis
                (With robusts methods, and parallelized functions).
License:        GPL
URL:            http://mulcyber.toulouse.inra.fr/projects/amap/
Packaged:       Tue Jan 24 10:30:53 2006; lucas
Built:          R 2.4.0; i386-pc-mingw32; 2006-10-03 21:25:08; windows
```

```
library(amap)
afc(sond)
```

```
Erreur dans sign(EIG$values) : fonction complexe non implémentée
```

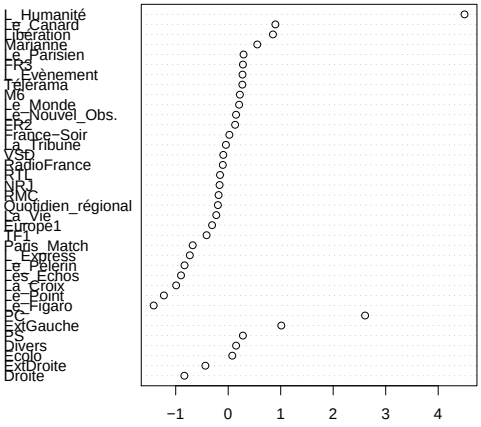
```
afc
function (x)
{
    f <- as.matrix(x/sum(x))
    fi <- apply(f, 1, sum)
    fj <- apply(f, 2, sum)
    f <- (1/fi) * t(t(f)/fj)
    acp(f, wI = fi, wV = fj, center = TRUE, reduce = FALSE)
}
<environment: namespace:amap>
```

L'intérêt pédagogique de cette fonction est indéniable. Elle a simplement le tort d'appeler une fonction `acp`, dans la même librairie, qui diagonalise des matrices non symétriques, ce qui est contradictoire avec l'essentiel de la théorie sous-jacente (les matrices symétriques réelles ont des valeurs propres et des vecteurs propres réels et le signe de valeurs propres complexes ne pose plus de problème !).

```
library(MASS)
sond.corresp <- corresp(sond)
names(sond.corresp)
```

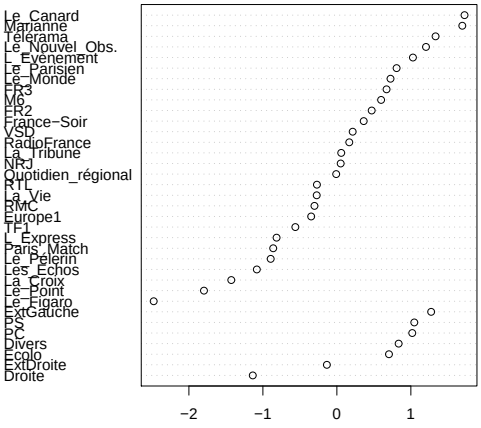
```
[1] "cor"    "rscore" "cscore" "Freq"
```

```
dotchart(c(sort(sond.corresp$cscore), sort(sond.corresp$rscore)))
```



L'association Parti Communiste/*L'Humanité* est forte et particulière. On peut l'appeler une *correspondance pointue*. Elle masque en partie une structure qui concerne l'ensemble du tableau :

```
sond <- sond[-8, ]
sond.corresp <- corresp(sond)
dotchart(c(sort(sond.corresp$cscore), sort(sond.corresp$rscore)))
```



Les opinions politiques ne sont pas des objets numériques. Leur associer une valeur, c'est faire du codage numérique (*scoring*) :

```
sond.corresp$cscore
```

```
ExtGauche         PC         PS      Ecolo     Droite  ExtDroite     Divers
1.2742747  1.0187901  1.0464341  0.7041875 -1.1366441 -0.1351406  0.8347609
```

La librairie `cocorresp` utilise l'AFC de la librairie `vegan` :

```
library(vegan)
sond.cca <- vegan::cca(X = sond)
sond.cca$CA$v[, 1]
```

```
ExtGauche         PC         PS      Ecolo     Droite  ExtDroite     Divers
1.2742747  1.0187901  1.0464341  0.7041875 -1.1366441 -0.1351406  0.8347609
```

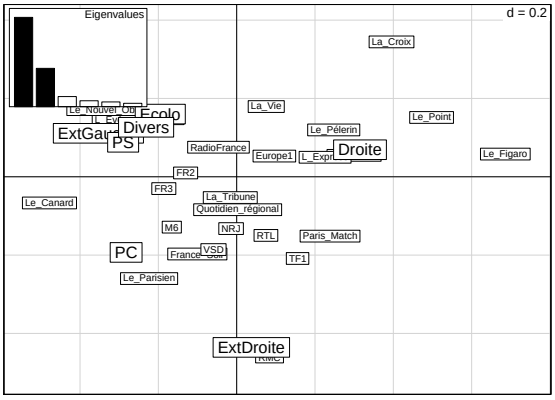
L'AFC est un cas particulier intégré à la *Canonical Correspondence Analysis*, fonction que nous aborderons avec les couplages de tableaux. Pour l'instant, les composantes `CA` de l'objet créé par `cca` sont les résultats de l'AFC et on retrouve bien le score présent dans `sond.corresp$cscore`.

```
library(ade4)
sond.coa <- dudi.coa(sond, scannf = F)
t(sond.coa$c1[, 1, drop = F])
```

```
    ExtGauche        PC        PS      Ecolo   Droite ExtDroite    Divers
CS1 -1.274275 -1.018790 -1.046434 -0.7041875 1.136644 0.1351406 -0.834761
```

L'AFC fournit donc des scores pour des modalités. On remarquera ici qu'un seul score ne permet pas d'épuiser la structure du tableau et qu'il faut en utiliser deux pour exprimer une réalité plus complexe que prévu :

```
scatter(sond.coa)
```



La fonction d'`amap` contient les éléments de compréhension des contraintes de construction des scores. Prenons l'exemple bordeaux préparé par J. van Rijckevorsel, [12, p. 32] :

```
data(bordeaux)
bordeaux
```

```
                  excellent good mediocre boring
Cru_Bourgeois            45  126       24      5
Grand_Cru_classe         87   93       19      1
Vin_de_table              0    0       52    148
Bordeaux_d_origine       36   68       74     22
Vin_de_marque             0   30      111     59
```

```
f <- as.matrix(bordeaux/sum(bordeaux))
fi <- apply(f, 1, sum)
fj <- apply(f, 2, sum)
```

Les données forment une distribution de fréquences bivariée dont on calcule les distributions marginales.

```
w <- corresp(bordeaux)
sum(fi * w$rscore)
```

```
[1] 2.081668e-17
```

```
 sum(fi * w$rscore^2)
[1] 1
 sum(fj * w$cscore)
[1] -5.551115e-17
 sum(fj * w$cscore^2)
[1] 1
```

Les scores sont de moyenne nulle et de variance unité pour les distribution marginales.

```
 matrix(w$rscore, 1) %*% f %*% matrix(w$cscore, 4)
          [,1]
[1,] 0.7685099
 w
First canonical correlation(s): 0.7685099
 Row scores:
     Cru_Bourgeois   Grand_Cru_classe       Vin_de_table Bordeaux_d_origine
         -0.9283350         -1.1010641          1.5381971         -0.2229937
     Vin_de_marque
         0.7141957

 Column scores:
 excellent       good   mediocre     boring
-1.1276880 -0.8747596  0.4626773  1.4348968
```

Les scores maximisent alors leur corrélation à travers la table de contingence, dite corrélation canonique. Ces opérations de codage numérique sous contrainte sont au cœur de l'analyse des correspondances multiples et il vaut mieux en comprendre l'existence sur ces premiers exemples.

2.2 Codage d'individus : un exemple

Utilisons le jeu de données décrit dans :

`http://pbil.univ-lyon1.fr/R/pdf/pps082.pdf`

```
 data(ours)
 summary(ours)
 altit  deniv  cloiso domain boise  hetra  favor  inexp  citat  depart
 1: 8   1:13   1:12   1: 9   1:10   1:19   1:15   1:20   1:22   AHP:5
 2:17   2:14   2: 4   2:13   2:15   2: 5   2:12   2:10   2: 7   AM :4
 3:13   3:11   3:22   3:16   3:13   3:14   3:11   3: 8   3: 4   D  :5
                                                      4: 5   HP :8
                                                             HS :4
                                                             I  :5
                                                             S  :7
```

Le tableau ne contient que des facteurs. Mais pour les 7 premières variables, le sens des modalités est commun. Elles sont notées 1,2 et 3 par ordre croissant d'avantage pour l'Ours. Nous pouvons donc les utiliser aussi comme variables quantitatives (disons comme indices semi-quantitatifs). En en faisant la somme nous mesurons la capacité d'accueil globale de la région pour l'Ours :

```
 w <- data.frame(lapply(ours[, 1:7], as.numeric))
 index <- apply(w, 1, mean)
 index
 [1] 2.571429 1.428571 2.571429 2.714286 2.285714 2.714286 2.000000 1.571429 2.285714
[10] 2.285714 1.142857 2.428571 2.571429 1.571429 2.000000 3.000000 2.714286 2.714286
[19] 2.285714 1.571429 1.571429 1.142857 1.571429 1.571429 2.000000 1.428571 1.571429
[28] 2.428571 1.428571 1.714286 2.428571 2.285714 2.142857 2.142857 1.714286 2.000000
[37] 2.142857 2.285714
 tapply(index, ours$citat, mean)
```

```
       1        2        3        4
1.961039 1.632653 2.607143 2.600000
```

L'indice est lié fortement à l'histoire de la disparition de l'espèce, mais peut-être pas d'une manière prévisible. Le calcul qui vient d'être fait peut être qualifié de *codage naïf*. Rien ne dit que ces codes disent la même chose et s'ajoutent (donc donnent une moyenne) sans précaution.

Nous pouvons souhaiter retenir des variables que leur part cohérente. Nous savons que l'ACP est destinée à cela du moins dans la conception des psychométriciens [2].

```
w.pca <- dudi.pca(w, scannf = F)
index.pca <- w.pca$l1[, 1]
cor(index, index.pca)
```

```
[1] -0.9437613
```

```
sum(cor(index, w)^2)
```

```
[1] 2.752590
```

```
sum(cor(index.pca, w)^2)
```

```
[1] 2.960705
```

```
w.pca$c1[, 1]
```

```
[1] -0.44673215 -0.26686887 -0.21571561  0.08223118 -0.50909746 -0.50186741
[7] -0.40603582
```

`index.pca` est un score de synthèse qui maximise la somme de ses corrélations avec les variables. Il est centré et réduit pour la pondération uniforme. Il fait un peu mieux (c'est normal, puisqu'il est optimal) que le codage naïf tout en restant très proche. L'amélioration s'est faite par élimination de la variable 4, appelée `domain`, dont on verra la définition dans la fiche citée. Cette définition n'est pas très claire et très différente des autres.

Ceci nous permet de poser la question : comment ferait-on pour mesurer le lien d'un score avec une variable qualitative. Par le rapport de corrélation, bien sûr, qui est le R^2 du modèle linéaire pour une variable qualitative comme le carré de corrélation simple est le R^2 du modèle linéaire pour une variable quantitative. Pour des variables qualitatives :

```
round(unlist(lapply(w, function(x) summary(lm(index ~ x))$r.squared)),
    3)
```

```
altit  deniv cloiso domain  boise  hetra  favor
0.570  0.247  0.274  0.020  0.626  0.571  0.444
```

```
round(cor(index, w)^2, 3)
```

```
     altit deniv cloiso domain boise hetra favor
[1,]  0.57 0.247  0.274   0.02 0.626 0.571 0.444
```

Pour des facteurs :

```
a1 <- unlist(lapply(ours[, 1:7], function(x) summary(lm(index ~
    x))$r.squared))
round(a1, 3)
```

```
altit  deniv cloiso domain  boise  hetra  favor
0.570  0.258  0.320  0.134  0.628  0.578  0.445
```

```
a2 <- unlist(lapply(ours[, 1:7], function(x) summary(lm(index.pca ~
    x))$r.squared))
round(a2, 3)
```

```
altit  deniv cloiso domain  boise  hetra  favor
0.599  0.230  0.214  0.121  0.767  0.747  0.492
```

la somme du premier fait 2.933 et celle du second est 3.17. Mais ni l'une ni l'autre ne sont optimales. Le but de l'analyse des correspondances multiples, l'ACM, est de trouver un score qui optimise cette quantité :

```
ours.acm <- dudi.acm(ours[, 1:7], scannf = F)
index.acm <- ours.acm$l1[, 1]
a3 <- unlist(lapply(ours[, 1:7], function(x) summary(lm(index.acm ~
    x))$r.squared))
round(a3, 3)
```

```
altit  deniv cloiso domain  boise  hetra  favor
0.583  0.145  0.335  0.246  0.797  0.732  0.420
```

La somme vaut maintenant 3.259. On ne pourrait faire mieux. On verra l'usage qu'il convient d'en faire, mais on déjà retenir que l'ACM est aux facteurs ce que l'ACP est aux variables quantitatives, une méthode de synthèse de la redondance.

2.3 Entre les scores des individus et ceux des modalités

Un dernier élément pratique est indispensable pour comprendre l'ACM. Nous avons vu qu'on peut attribuer des scores aux modalités. Quand on a deux variables qui forment les lignes et les colonnes d'un tableau, l'AFC leur donne des valeurs numériques. Quand on a un tableau de facteurs l'ACM code les individus-lignes du tableau. Quelle relation peut-on établir entre les deux ?

Noter d'abord qu'une table de contingence est une façon commode de . . . faire disparaître les individus, apparemment bien sûr.

```
t(ours[, c("citat", "boise")])
```

```
      1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17  18  19
citat "1" "2" "2" "3" "1" "3" "1" "2" "4" "1" "1" "3" "3" "2" "1" "4" "4" "4" "4"
boise "2" "1" "2" "3" "2" "3" "2" "1" "3" "3" "1" "3" "3" "1" "2" "3" "3" "3" "3"
      20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
citat "2" "1" "2" "2" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1" "1"
boise "2" "2" "1" "1" "1" "2" "2" "1" "3" "2" "1" "2" "2" "2" "3" "1" "2" "3" "2"
```

```
table(ours$citat, ours$boise)
```

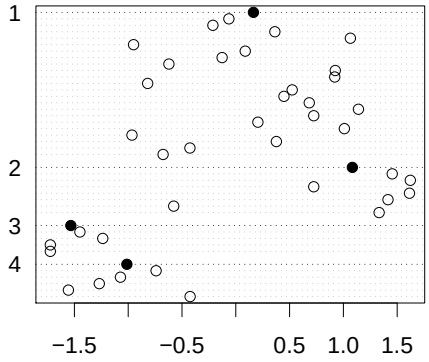
```
   1  2  3
1  5 13  4
2  5  2  0
3  0  0  4
4  0  0  5
```

Il s'agit bien de la même information : l'individu devient la case du tableau, c'est-à-dire la correspondance, ou plutôt le groupe d'individus identiques résumé par son effectif dans la table de contingence. Dans un cas (2 variables) comme dans l'autre (v variables), il y a bien des individus et des modalités donc un lien à établir entre les scores des uns et des autres.

Pour passer d'un score des individus à un score des modalités, l'opération est simple. Il suffit de prendre la moyenne des porteurs de la même modalité. Cette opération a deux significations. D'une part elle code les modalités, d'autre part elle mesure le lien :

```
dotchart(index.acm, gr = ours$citat, gdata = tapply(index.acm, ours$citat,
    mean), cex = 1.5, gpch = 19)
summary(lm(index.acm ~ ours$citat))$r.squared
```

```
[1] 0.6141054
```



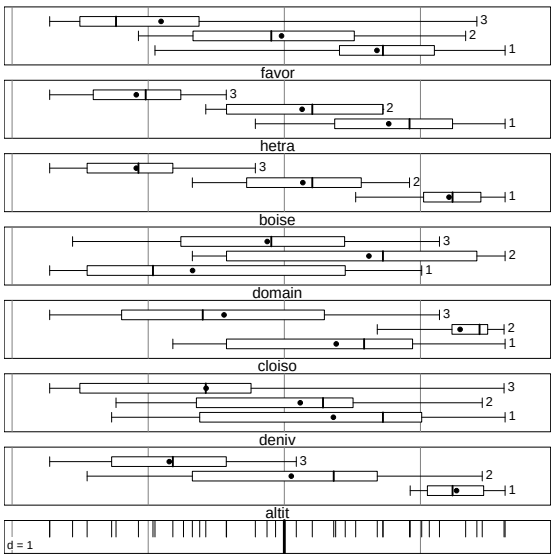
La mesure (entre 0 et 1, c'est un pourcentage de variance expliquée) croit avec la séparation des groupes. La nouveauté est ici qu'on fait l'opération sur plusieurs variables en même temps :

```
sco.boxplot(index.acm, ours[, 1:7])
round(a3, 3)
```

```
altit  deniv cloiso domain  boise  hetra  favor
0.583  0.145  0.335  0.246  0.797  0.732  0.420
```

```
round(mean(a3), 3)
```

```
[1] 0.466
```



La liste des rapports de corrélation est le complément indispensable de la figure et il est logique de mesurer le lien du score avec le tableau par la moyenne du lien avec chacun des facteurs. A retenir : un score numérique des individus donne

un score numérique des modalités par averaging (moyenne du groupe portant la même modalité). Si le score est de variance 1, les variances inter-groupes sont plus petite que 1, leur moyenne également et la pertinence du score pour le tableau est la moyenne de ces variances inter-groupes (ou moyenne de rapports de corrélation).

Dans l'autre sens, si chaque modalité de chaque variable porte un score numérique, alors on attribue à un individu la moyenne des modalités portées. C'est ce qu'on a fait dans le code naïf. Dans ce sens également, on peut désirer mesurer la pertinence, non plus du score des individus pour le tableau, mais du score des modalités des facteurs pour ce tableau. Si les modalités sont codées de manière intelligente, les modalités portée par un individu devraient être voisines, les correspondances multiples comme paquet de modalités devraient être séparées, la variance obtenue pour le score des individus devraient être grande.

L'ACM va proposer de tels codages de modalités. Le premier est :

```
round(t(ours.acm$c1[, 1, drop = F]), 3)
```

```
    altit.1 altit.2 altit.3 deniv.1 deniv.2 deniv.3 cloiso.1 cloiso.2 cloiso.3
CS1   1.854   0.075  -1.239   0.528   0.171  -0.842    0.559    1.893   -0.649
    domain.1 domain.2 domain.3 boise.1 boise.2 boise.3 hetra.1 hetra.2 hetra.3
CS1   -0.988    0.911   -0.184   1.773   0.199  -1.594   1.124   0.191  -1.594
    favor.1 favor.2 favor.3
CS1   0.997   -0.03  -1.327
```

Le premier district est noté :

```
ours[1, 1:7]
```

```
  altit deniv cloiso domain boise hetra favor
1     2     3      3      2     2     3     3
```

Le passage du code des modalités au code des individus demande de passer des numéros des modalités par variables aux numéros des modalités globalement et de faire une moyenne :

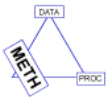
```
mean(ours.acm$c1[cumsum(rep(3, 7)) - 3 + as.numeric(ours[1, 1:7]),
    1])
```

```
[1] -0.4610403
```

```
ours.acm$li[1, 1]
```

```
[1] -0.4610403
```

Le résultat est la moyenne des modalités portées par l'individu. On l'appelle la coordonnée. Les contraintes qui pèsent sur la définition des scores des modalités demandent de passer dans le champ mathématique. On peut s'en servir sans ça.



3 Le schéma de l'ACM

L'analyse des correspondances multiples est destinée à la synthèse d'un tableau de variables qualitatives vu comme la juxtaposition de paquets d'indicatrices des classes (tableau disjonctif complet) : Chaque tableau fait une typologie simple des individus (partition). L'objectif de l'ACM est de donner un point de vue coordonné sur toutes ces partitions. On note :

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_v]$$

le tableau disjonctif complet. v est le nombre de variables qualitatives. n est le nombre d'individus (lignes) et $m_1, m_2, \cdots, m_v$ sont les nombres de modalités

de chaque variable. Le nombre de modalités total est $m = m_1 + m_2 + \cdots + m_v$. On peut considérer que l'ACM fait partie des analyses à un tableau ($\mathbf{X}$). On suppose que chaque ligne porte le poids $1/n$ (comme dans une ACP) ou le poids p_i ($\sum_{i=1}^{n} p_i = 1$). $\mathbf{D}$ est comme d'habitude la diagonale des poids des lignes (matrice $n \times n$). On compte pour chaque modalité le nombre d'individus l'utilisant (le nombre de porteurs) en faisant :

$$\mathbf{X}^t\mathbf{1}_n = [n_{11}, n_{12}, \cdots, n_{1m_1}, n_{21}, n_{22}, \cdots, n_{2m_2}, ..., n_{v1}, n_{v2}, \cdots, n_{vm_v}]$$

Il vient simplement que :

$$n_{11}+n_{12}+\cdots+n_{1m_1} = n_{21}+n_{22}+\cdots+n_{2m_2} = ... = n_{v1}+n_{v2}+\cdots+n_{vm_v} = n$$

Au lieu de compter les porteurs d'une modalité, on peut aussi sommer les poids des porteurs de cette modalité. Pour une pondération uniforme, le poids d'une modalité est la fréquence ($f_{ij} = n_{ij}/n$) du nombre de porteurs et on a :

$$f_{11} + f_{12} + \cdots + f_{1m_1} = f_{21} + f_{22} + \cdots + f_{2m_2} = ... = f_{v1} + f_{v2} + \cdots + f_{vm_v} = 1$$

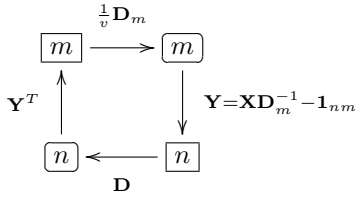
mais en général, les poids des modalités sont dans :

$$[f_{11}, f_{12}, \cdots, f_{1m_1}, f_{21}, f_{22}, \cdots, f_{2m_2}, ..., f_{v1}, f_{v2}, \cdots, f_{vm_v}] = \mathbf{X}^t\mathbf{D}\mathbf{1}_n$$

Les poids des modalités sont écrits sur la diagonale d'une matrice diagonale :

$$\mathbf{D}_m = \text{Diag}\,(f_{11}, f_{12}, \cdots, f_{1m_1}, f_{21}, f_{22}, \cdots, f_{2m_2}, ..., f_{v1}, f_{v2}, \cdots, f_{vm_v})$$

La somme de ces éléments vaut donc v. $\mathbf{D}_m$est une matrice à m lignes et m colonnes. Sa trace vaut v. On note $\mathbf{1}_{nm}$ la matrice à n lignes et m colonnes ne contenant que des 1. L'ACM (source dans [9], synthèse incontournable dans [11]) est l'analyse du schéma :



Le paramétrage de ce schéma a été choisi pour ses bonnes propriétés. En effet, on considère les vecteurs indicateurs de variables dans l'ensemble des modalités :

$$[\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_v] = \begin{bmatrix} 1 & 0 & & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & & 1 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & & 1 \end{bmatrix} \begin{matrix} \\ \\ \} m_1 \\ \\ \\ \} m_1 \\ \\ \\ \} m_v \\ \\ \end{matrix}$$

On a toujours $\mathbf{X}\mathbf{w}_k = \mathbf{1}_n$ (il y a une modalité présente et une seule pour chaque individu et chaque variable) donc

$$\left(\mathbf{X}\mathbf{D}_m^{-1} - \mathbf{1}_{nm}\right)\frac{1}{v}\mathbf{D}_m\mathbf{w}_k = \mathbf{0}_n$$

et les vecteurs $\mathbf{w}_k$ sont propres pour 0. Il y a donc dans cette analyse au moins v valeurs propres nulles. Les axes principaux sont des vecteurs à m composantes formés de $m = m_1 + m_2 + \cdots + m_v$ "morceaux". Chaque segment est un codage numérique des modalités de la variable correspondante. Par orthogonalité aux vecteurs $\mathbf{w}_k$, ces scores des modalités sont centrés (pour les poids des modalités) et $\frac{1}{v}\mathbf{D}_m$-normés, donc de variance moyenne égale à 1. Ils maximisent successivement :

$$\left\|\left(\mathbf{X}\mathbf{D}_m^{-1} - \mathbf{1}_{nm}\right)\frac{1}{v}\mathbf{D}_m\mathbf{a}\right\|_{\mathbf{D}}^2 = \left\|\left(\mathbf{X} - \mathbf{1}_{nm}\mathbf{D}_m\right)\frac{1}{v}\mathbf{a}\right\|_{\mathbf{D}}^2 = \left\|\frac{1}{v}\mathbf{X}\mathbf{a}\right\|_{\mathbf{D}}^2$$

Or le vecteur $\mathbf{1}_n$ est composante principale pour la valeur propre 0

$$\left(\mathbf{D}_m^{-1}\mathbf{X}^t - \mathbf{1}_{mn}\right)\mathbf{D}\mathbf{1}_n = 0$$

donc les autres composantes principales sont centrées et $\left\|\frac{1}{v}\mathbf{X}\mathbf{a}\right\|_{\mathbf{D}}^2$ est la variance du score des individus obtenu en attribuant la moyenne des scores des modalités portées. Inversement, une composante principale est un score des individus de moyenne 0 et variance 1 qui maximise :

$$\left\|\left(\mathbf{D}_m^{-1}\mathbf{X}^t - \mathbf{1}_{mn}\right)\mathbf{D}\mathbf{b}\right\|_{\frac{1}{v}\mathbf{D}_m}^2 = \left\|\mathbf{D}_m^{-1}\mathbf{X}^t\mathbf{D}\mathbf{b}\right\|_{\frac{1}{v}\mathbf{D}_m}^2$$

c'est-à-dire la moyenne de la variance du score des modalités obtenu en attribuant à chaque modalité la moyenne des scores des individus porteurs. L'objectif fondamental de l'ACM est de d'obtenir des scores numériques des individus qui maximisent les pourcentages de variance expliquée, en moyenne, pour toutes les variables qualitatives (rapport de corrélation) : L'ACM fait la synthèse des liens entre les variables qualitatives par l'examen des liens entre les variables qualitatives $\mathbf{q}^j$ et un code numérique de synthèse appelée variable synthétique. En ACP normée, on a trouvé une variable synthétique (la première composante principale, concrètement la première coordonnée normée des lignes) qui maximise :

$$\sum_{j=1}^{p} cor^2\left(\mathbf{x}^j, \mathbf{y}\right)$$

En ACM, on a trouvé une variable synthétique (la première composante principale, concrètement la première coordonnée normée des lignes) qui maximise :

$$\frac{1}{v}\sum_{j=1}^{v}\eta^2\left(\mathbf{q}^j, \mathbf{y}\right)$$

C'est le point de vue implanté dans la librairie `ade4` et la fonction `acm`. On peut comparer avec la version implanté dans la fonction `mca` de la librairie `MASS` du logiciel R.

4 mca et acm

C'est pour éviter les conflits que les noms sont différents. Mais les contenus ne sont pas identiques. Retrouvons d'abord les effectifs par modalités :

```
unlist(sapply(as.list(ours), summary))
```

```
  altit.1   altit.2   altit.3   deniv.1   deniv.2   deniv.3   cloiso.1
        8        17        13        13        14        11        12
 cloiso.2  cloiso.3  domain.1  domain.2  domain.3   boise.1   boise.2
        4        22         9        13        16        10        15
  boise.3   hetra.1   hetra.2   hetra.3   favor.1   favor.2   favor.3
       13        19         5        14        15        12        11
  inexp.1   inexp.2   inexp.3   citat.1   citat.2   citat.3   citat.4
       20        10         8        22         7         4         5
depart.AHP depart.AM depart.D depart.HP depart.HS depart.I depart.S
        5         4         5         8         4         5         7
```

L'analyse dans `MASS` donne :

```
library(MASS)
help(mca)
ours.mca <- mca(ours)
```

La documentation indique clairement (version 7.2-30) :

> `df` A data frame containing only factors

```
ours.mca
```

```
Call:
mca(df = ours)
Multiple correspondence analysis of 38 cases of 10 factors

Correlations 0.668 0.568  cumulative % explained  7.42 13.73
```

```
names(ours.mca)
```

```
[1] "rs"   "cs"   "fs"   "d"    "p"    "call"
```

Le résultat est une liste :

`rs`	The coordinates of the rows, in nf dimensions.
`cs`	The coordinates of the column vertices, one for each level of each factor.
`fs`	Weights for each row, used to interpolate additional factors in predict.mca.
`p`	The number of factors
`d`	The singular values for the nf dimensions.
`call`	The matched call.

L'analyse dans `ade4` est :

```
ours.acm <- dudi.acm(ours, scannf = F)
names(ours.acm)
```

```
 [1] "tab"  "cw"   "lw"   "eig"  "rank" "nf"   "c1"   "li"   "co"   "l1"   "call"
[12] "cr"
```

On conserve dans un cas les valeurs singulières et dans l'autre les valeurs propres. Les secondes sont les carrés des premières. En fait les deux fonctions calculent les secondes mais l'une des deux conservent leur carré :

```
ours.mca$d
```

```
[1] 0.6676898 0.5681736
```

```
ours.mca$d^2
```

```
[1] 0.4458097 0.3228212
```

```
ours.acm$eig
```

```
 [1] 0.445809666 0.322821219 0.250023933 0.204320821 0.183747109 0.157044406
 [7] 0.150687859 0.139059296 0.108877787 0.097752121 0.079478800 0.073567819
[13] 0.055419820 0.040440525 0.035052208 0.031079669 0.029008046 0.024858103
[19] 0.022036630 0.017134598 0.014005836 0.008237655 0.004580786 0.003515076
[25] 0.001440212
```

Seules les valeurs propres non nulles sont conservées. On en attend au plus $m - v = 35 - 10 = 25$ et on en a exactement 25. Pour les scores des lignes :

```
head(ours.mca$rs)
```

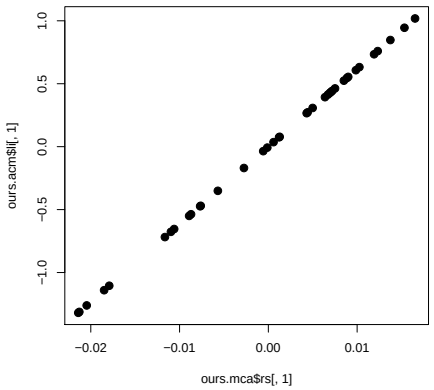
```
            1           2
1 -0.008708887 -0.006657489
2  0.007205219 -0.014715926
3 -0.007680065 -0.014097782
4 -0.021409317 -0.003557682
5 -0.005688136 -0.004361274
6 -0.021294238 -0.003341254
```

```
head(ours.acm$li)
```

```
      Axis1     Axis2
1 -0.5368518 0.4103952
2  0.4441595 0.9071506
3 -0.4734310 0.8690457
4 -1.3197589 0.2193102
5 -0.3506403 0.2688470
6 -1.3126650 0.2059687
```

La comparaison manque d'évidence. Pourtant :

```
plot(ours.mca$rs[, 1], ours.acm$li[, 1], pch = 20, cex = 2)
```



Ouf ! C'est une question de normalisation. Nous avons $v = 10$ variables, $n = 38$ individus et $m = 35$ modalités.

```
mean(ours.acm$li[, 1])
```

```
[1] -6.569766e-18
```

```
mean(ours.acm$li[, 1]^2)
```

```
[1] 0.4458097
```

Les coordonnées des lignes dans `ade4` sont de moyennes nulles et de variance (au sens de $1/n$) λ_k.

```
sum(ours.mca$rs[, 1])
```

```
[1] 5.219078e-17
```

```
sum(ours.mca$rs[, 1]^2)
```

```
[1] 0.004458097
```

Les coordonnées des lignes dans MASS sont de moyennes nulles et de somme de carrés (carré de norme canonique) λ_k/v^2 donc de variance (au sens de $1/n$) de λ_k/nv^2. A une constante près, c'est donc bien la même chose.

Pour les scores des modalités la comparaison est plus difficile.

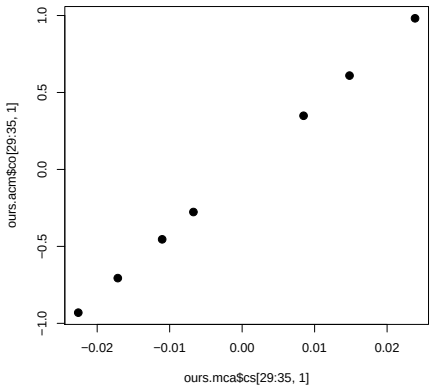
```
ours.mca$cs[29:35, ]
```

```
                      1           2
depart.AHP  0.008476450  0.038674833
depart.AM   0.014815156  0.026730595
depart.D   -0.006721478  0.009831842
depart.HP   0.023856883 -0.001535458
depart.HS  -0.017155892 -0.030224840
depart.I   -0.022607067 -0.015410554
depart.S   -0.011033094 -0.019888565
```

```
ours.acm$co[29:35, ]
```

```
               Comp1       Comp2
depart.AHP  0.3488836 -1.35456947
depart.AM   0.6097794 -0.93622763
depart.D   -0.2766504 -0.34435605
depart.HP   0.9819294  0.05377877
depart.HS  -0.7061222  1.05861207
depart.I   -0.9304879  0.53974807
depart.S   -0.4541129  0.69658848
```

```
plot(ours.mca$cs[29:35, 1], ours.acm$co[29:35, 1], pch = 20, cex = 2)
```



Encore une affaire d'échelle.

```
ours.acm$co[29:35, 1]/10/sqrt(38)/ours.mca$d[1]
```

```
[1]  0.008476450  0.014815156 -0.006721478  0.023856883 -0.017155892 -0.022607067
[7] -0.011033094
```

```
ours.acm$c1[29:35, 1]/10/sqrt(38)
```

```
[1]  0.008476450  0.014815156 -0.006721478  0.023856883 -0.017155892 -0.022607067
[7] -0.011033094
```

```
as.numeric(ours.mca$cs[29:35, 1, drop = F])
```

```
[1]  0.008476450  0.014815156 -0.006721478  0.023856883 -0.017155892 -0.022607067
[7] -0.011033094
```

Donc les scores des colonnes de `mca` sont celles de `acm` multipliées par $v/sqrtn\lambda_k$ et les coordonnées des lignes de `mca` sont celles de `acm` multipliées par $v/\sqrt{n}$

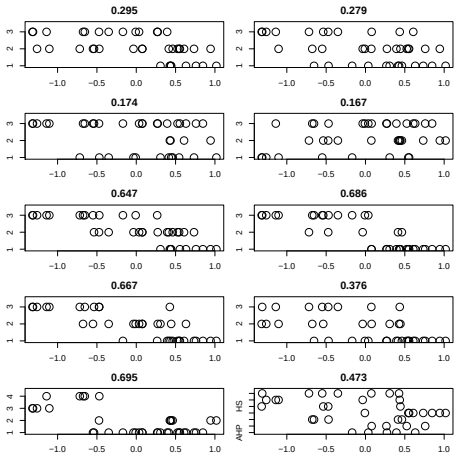
On peut se demander pourquoi ces écarts. On se reportera à Tenenhaus & Young *op. cit.* pour approfondir cette question. nv^2 est l'effectif total du tableau

de Burt $\mathbf{X}^t\mathbf{X}$, juxtaposition des v^2 tables de contingence définies par couple de variables. Chacune d'entre elles contient n individus. C'est un des points de vue. Les mises à l'échelle ont de toute manière peu d'importance dans les biplot.

Pour interpréter l'analyse, on utilise les rapports de corrélation. On peut examiner leur définition dans la petite fonction :

```
stripchart.mca <- function(mca, numfac = 1) {
    if (!(class(mca) == "mca"))
        stop("mca class expected")
    nomfic <- mca$call[2]
    tab <- eval(parse(text = nomfic))
    nvar <- mca$p
    nind <- nrow(tab)
    par(mfrow = c((nvar + 1)/2, 2))
    par(mai = c(0.3, 0.3, 0.3, 0.1))
    score <- mca$rs[, numfac] * nvar * sqrt(nind)
    f1 <- function(x) {
        lm0 <- lm(score ~ as.factor(x))
        r2 <- var(predict(lm0))/var(score)
        a0 <- split(score, x)
        stripchart(a0, pch = 1, main = paste(round(r2, digits = 3)),
            method = "stack", cex = 2)
        r2
    }
    apply(tab, 2, f1)
}
w1 <- stripchart.mca(ours.mca)
mean(w1)
```

```
[1] 0.4458097
```



Identifier le principe de la construction de la figures. Expliquer le sens de la valeur numérique placée au dessus de chaque fenêtre. Quelle est la moyenne de ces valeurs ? Interpréter l'information obtenue. Cette manière de dépouiller l'ACM est due à Saporta [9].

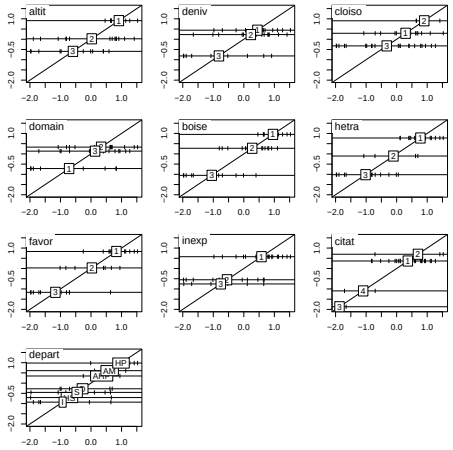
Retrouver l'information numérique dans :

```
ours.acm$cr[, 1]
```

```
[1] 0.2951241 0.2793041 0.1739039 0.1670369 0.6467664 0.6858709 0.6674468 0.3755331
[9] 0.6945030 0.4726075
```

et l'information graphique dans :

```
score(ours.acm)
```



Cette pratique s'étend à deux dimensions de manière simple (figure 1) avec :

```
par(mfrow = n2mfrow(ncol(ours)))
par(mar = rep(0, 4))
for (i in 1:(ncol(ours))) {
    s.class(ours.acm$l1, ours[, i], clab = 1.5, sub = names(ours)[i],
        csub = 3, possub = "topleft", cgrid = 0, csta = 1, cpoi = 2,
        cell = 0)
    text(2.2, -0.3, paste(round(100 * ours.acm$cr[i, 1], 0), "%",
        sep = ""), cex = 2)
    text(-0.8, -2.2, paste(round(100 * ours.acm$cr[i, 2], 0), "%",
        sep = ""), cex = 2)
}
s.label(ours.acm$l1, clab = 0, cpoi = 2, sub = "Common background",
    csub = 2.5)
par(mfrow = c(1, 1))
```

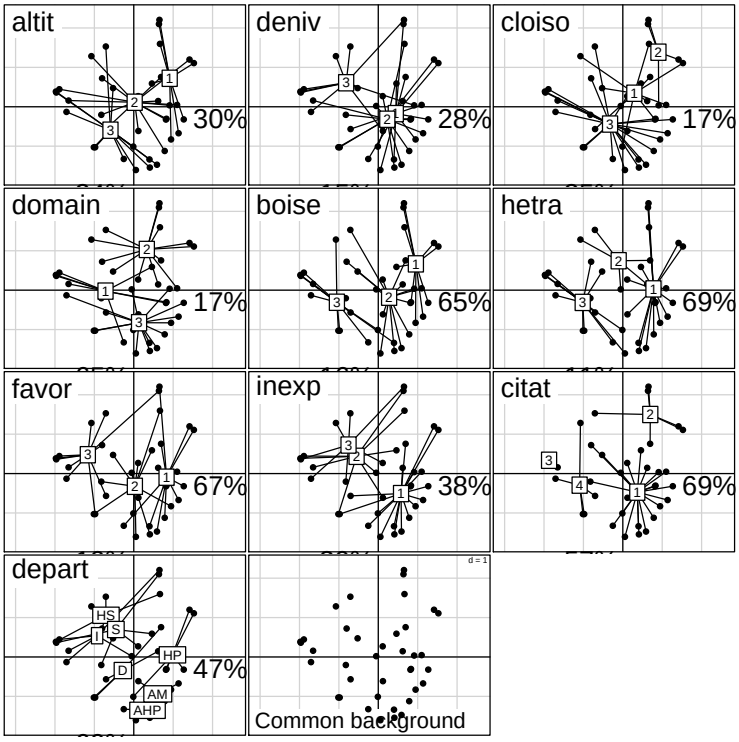


FIG. 1 – Plan 1-2 d'une analyse des correspondances multiples. Dans chaque fenêtre les individus (régions) sont placés avec le même système de deux scores (moyenne nulle, variance unité, covariance nulle). Dans chaque fenêtre, une des variables du tableau (facteur) divise le nuage en groupes des porteurs de la même modalité. Sur chacun des axes est donné le pourcentage de variance du score expliquée par le facteur (rapport de corrélation). L'ACM donne les scores qui maximise la moyenne de ces rapports de corrélation.

5 Exercice : ACM et AFC

Utiliser la fonction acm.disjonctif :

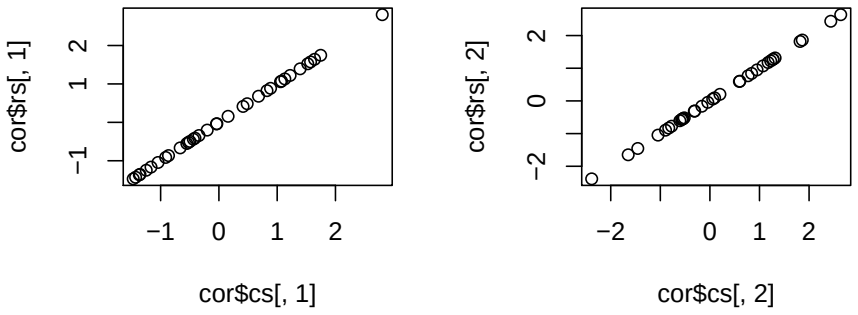
```
ours01 <- acm.disjonctif(ours)
```

Identifier le contenu de `ours01`.

```
oursburt <- t(ours01) %*% as.matrix(ours01)
```

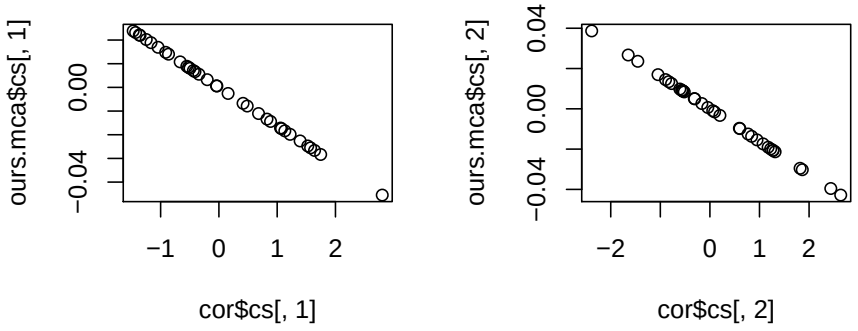
Identifier le contenu de `oursburt` (on l'appelle le tableau de Burt). Faire l'analyse des correspondances simples (AFC) du tableau de Burt.

```
cor <- corresp(as.matrix(oursburt), nf = 2)
par(mfrow = c(1, 2))
plot(cor$cs[, 1], cor$rs[, 1])
plot(cor$cs[, 2], cor$rs[, 2])
```



Expliquer la nature et l'origine de la relation entre les scores des lignes et des colonnes de l'AFC du tableau de Burt.

```
par(mfrow = c(1, 2))
plot(cor$cs[, 1], ours.mca$cs[, 1])
plot(cor$cs[, 2], ours.mca$cs[, 2])
```



Expliciter la nature de la relation entre les scores des lignes (ou des colonnes) de l'AFC du tableau de Burt et les scores des modalités de l'ACM du tableau.

Expliciter la nature de la relation entre les valeurs propres de l'AFC du tableau de Burt et celles de l'ACM du tableau (démonstrations dans [10, p. 224]).

Faire l'analyse des correspondances simples (AFC) du tableau disjonctif complet.

```
cordi <- corresp(ours01, nf = 2)
```

Expliciter la nature de la relation entre les valeurs propres de l'AFC du tableau disjonctif et celles de l'ACM du tableau.

Expliciter la nature de la relation entre les scores des lignes (respectivement des colonnes) de l'AFC du tableau disjonctif et les scores des lignes (respectivement des modalités) de l'ACM du tableau. L'ACM peut être définie comme l'AFC du tableau disjonctif complet.

L'équivalence de l'AFC du tableau de Burt et celle de l'AFC du tableau disjonctif complet est complètement détaillée dans [4, p. 126].

Il est surprenant qu'une même procédure appliquée à plusieurs formes de tableaux donnent des résultats cohérents. C'est surprenant au plan expérimental, puisqu'on associe souvent méthode et type de données. Ce ne l'est pas au plan mathématique tant la structure des facteurs (paquet complet d'indicatrices de classes) et la définition de l'AFC sont en accord.

Mais cela a pu conduire les pionniers à essayer une telle procédure sur toute sorte de tableaux de nombres positifs ou nuls avec parfois un succès pragmatique réel qu'on a justifié ensuite. Comme quoi la relation entre la nature et la fonction des routines statistiques peut être plus compliquée que ne le laisse croire la notion simple d'application.

6 ACM pondérée

La routine d'ade4 admet des pondérations quelconques pour les individus qui peuvent être des groupes d'individus identiques :

```
data(worksurv)
```

Les données sont proposées dans [8, p. 283 et suivantes] et reproduites dans [3]. Le tableau est formé de 4 facteurs mais a un attribut qui donne le nombre de personnes dans chacune des configurations de réponses à un sous-questionnaire de 4 questions (enquête de 1970).

```
lapply(worksurv, levels)
```

```
$pro
[1] "CGT"     "CFDT"    "FO"      "CFTC"    "Auton"   "Abst"    "Nonaffi" "NR"
$una
[1] "CGT"     "CFDT"    "FO"      "CFTC"    "Auton"   "CGC"     "Notaffi" "NR"

$pre
[1] "Duclos"   "Deferre"  "Krivine"  "Rocard"   "Poher"    "Ducatel"  "Pompidou"
[8] "NRAbs"

$pol
[1] "Communist" "Socialist" "Left"      "Center"    "RI"        "Right"
[7] "Gaullist"  "NR"
```

Chaque question a 8 modalités de réponse.

pro La première concerne les élections professionnelles avec les modalités de réponse `CGT`, `CFDT`, `FO`, `CFTC`, `Auton` (Syndicats autonomes), `Abst` (abstention), `Nonaffi` (liste non affiliée) et `NR` (non réponse).

una La seconde concerne l'appartenance syndicale avec les modalités de réponse `CGT`, `CFDT`, `FO`, `CFTC`, `Auton` (Syndicats autonomes), `CGC`, `Nonaffi` (non affilié) et `NR` (non réponse).

pre La troisième concerne le vote aux élections présidentielle de 1969 pour l'un des candidats `Duclos`, `Deferre`, `Krivine`, `Rocard`, `Poher`, `Ducatel`, `Pompidou` et `NRAbs` (non réponse ou abstention).

pol La dernière donne la sympathie politique de la personne avec les modalités `Communist` (PCF), `Socialist` (SFIO+PSU+FGDS), `Left` (extrême gauche), `Center` (MRP+RAD.), `RI` (républicains indépendants), `Right` (INDEP.+CNI), `Gaullist` (UNR) et NR (Non réponse).

```
counts <- attr(worksurv, "counts")
sum(counts)
```

```
[1] 1049
```

```
dim(worksurv)
```

```
[1] 319   4
```

L'opinion de 1049 personnes a fourni 319 types de réponse distincts sur les 4096 possibles.

```
worksurv.acm <- dudi.acm(worksurv, row.w = counts, scan = F)
par(mar = rep(0, 4))
par(mfrow = c(2, 2))
for (k in 1:4) {
    s.value(worksurv.acm$li, counts, cleg = 0)
    s.class(worksurv.acm$li, factor(worksurv[, k]), counts, add.plot = T,
        cell = 0, cstar = 0.5)
    text(2.3, 0.2, round(worksurv.acm$cr[k, 1], 2), cex = 1.5)
    text(0.5, 3, round(worksurv.acm$cr[k, 2], 2), cex = 1.5)
}
```

On obtient une image exprimant clairement les clivages politiques associés à la syndicalisation de cette époque (figure 2).

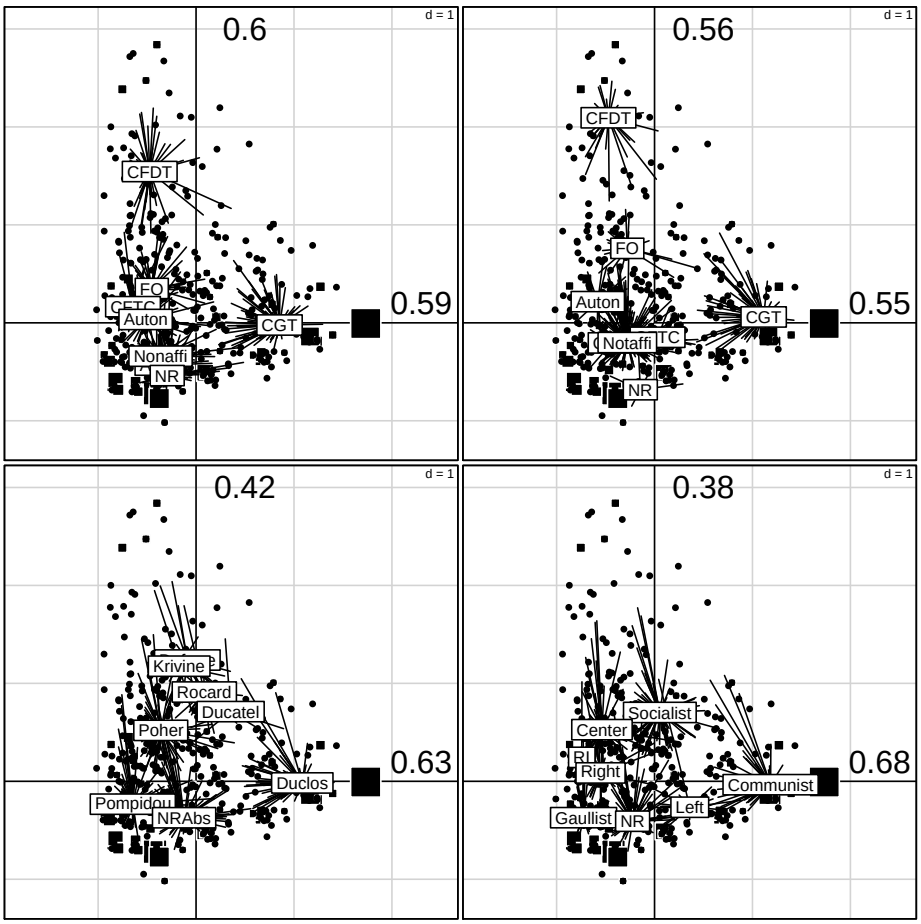


FIG. 2 – Analyse des correspondances multiples pondérée à 4 facteurs. Chaque point est une combinaison de réponse distincte et les carrés noirs donnent le nombre de personnes utilisant ce pattern de réponse. Les rapports de corrélation indique la part de variance du score expliquée par le facteur.

7 Extension : AFC et différentiateurs sémantiques

Pour l'utilisateur l'analyse des correspondances multiples est donc l'analyse des tableaux dont les colonnes sont des modalités de facteurs. Le cas le plus simple est celui d'un tableau dont toutes les variables (colonnes) sont des facteurs (par exemple les réponses à un questionnaire) :

http://pbil.univ-lyon1.fr/R/pdf/pps004.pdf

La difficulté est dans la constitution initiale du questionnaire ou dans le recodage des réponses qui peut être un gros travail. On trouvera un autre exemple remarquable dans la fiche proposée par Mickaël Séon :

http://pbil.univ-lyon1.fr/R/pdf/dssb07.pdf

Un dernier exemple permet d'illustrer une extension de cette pratique. Les données sont décrites dans :

http://pbil.univ-lyon1.fr/R/pdf/pps002.pdf

On a demandé à 11 étudiants du DEA (Diplôme d'Etudes Approfondies) de didactique des disciplines scientifiques (enseignants scientifiques) ce qu'ils pensaient de 15 disciplines scientifiques à l'aide de 5 différenciateurs sémantiques d'Osgood. L'exemple est choisi pour sa forme initiale : 11 étudiants, 15 disciplines, 6 différentiateurs et pour chacune des combinaisons possibles une modalité de réponse notée de 1 à 5. Le bordereau de réponse est un tableau disciplines-différenciateurs, il y en a un par étudiant.

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/pps002.rda"))
pps002[[1]]
```

```
      prec dur subj faux faib fant
algeb    1   1    5    5    5    5
astro    3   5    1    3    3    3
biomo    3   1    4    4    4    4
didac    3   4    4    3    1    5
ecolo    3   3    4    5    3    5
ethol    4   1    5    3    3    3
infor    1   1    5    5    5    5
lingu    3   3    4    4    4    4
medec    3   5    3    2    3    3
neuro    2   1    4    5    4    5
phynu    3   2    3    3    5    3
psych    4   5    3    3    3    3
scien    1   2    4    4    4    5
socio    3   4    4    3    3    3
techn    1   1    5    5    5    5
```

Chaque fois qu'on voit un tableau, on pense analyse multivariée. Ici, on a un cube (disciplines-différentiateurs-étudiants) avec des modalités de réponse. Il y a 3 manières de découper le cube en tranches et donc 16 tableaux différentiateurs-étudiants, 11 tableaux disciplines-différentiateurs et 6 tableaux disciplines-étudiants. Il y a encore deux manières pour un tableau de dire où sont les individus (lignes ou colonnes) et les variables (colonnes ou lignes). Il faut donc abandonner définitivement l'idée que le traitement des données doit s'adapter à leur forme. C'est exactement le contraire qui a un sens. Il faut trouver la forme de présentation des données qui va permettre de résoudre une question.

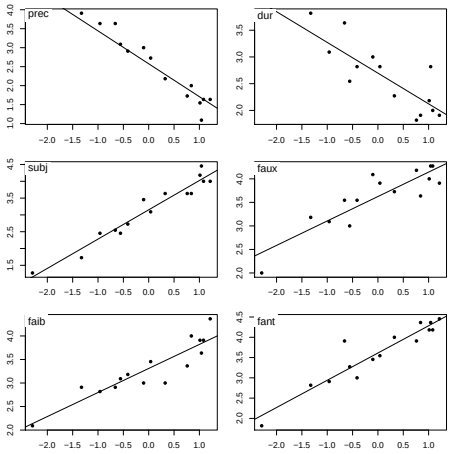
Disons : c'est la structure collective de l'opinion exprimée qui nous intéresse. L'originalité des personnes interrogées n'est pas la question, au contraire, c'est

l'opinion moyenne qui est en jeu. On peux faire la moyenne des notes pour chaque couple disciplines-différentiateurs et traiter le tableau.

```
w <- pps002[[1]]
for (k in 2:11) w <- w + pps002[[k]]
w <- w/11
round(w, 1)
```

```
      prec dur subj faux faib fant
algeb  1.1 2.8  4.5  4.3  3.6  4.4
astro  3.9 3.8  1.3  2.0  2.1  1.8
biomo  1.7 1.8  3.6  4.2  3.4  3.9
didac  3.6 3.6  2.5  3.5  2.9  3.9
ecolo  3.0 3.0  3.5  4.1  3.0  3.5
ethol  2.9 2.8  2.7  3.5  3.2  3.0
infor  1.6 1.9  4.0  3.9  4.4  4.5
lingu  2.7 2.8  3.1  3.9  3.5  3.5
medec  3.1 2.5  2.5  3.0  3.1  3.3
neuro  2.2 2.3  3.6  3.7  3.0  4.0
phynu  2.0 1.9  3.6  3.6  4.0  4.4
psych  3.9 3.8  1.7  3.2  2.9  2.8
scien  1.6 2.0  4.0  4.3  3.9  4.2
socio  3.6 3.1  2.5  3.1  2.8  2.9
techn  1.5 2.2  4.2  4.0  3.9  4.2
```

```
wpca <- dudi.pca(w, scan = F)
score(wpca)
```



Qu'est-ce qui est associé à imprécis, mou, subjectif, faux, faible et fantaisiste ? Oui, mais ! Un différentiateur n'est pas une variable quantitative simple. Gauche-Droite (par exemple !), en moyenne 3 si tout le monde s'en moque (moyenne 3 et variance nulle) n'a quand même pas la même signification que 1 pour la moitié et 5 pour l'autre (moyenne 3 et variance maximum). C'est pourquoi on remplace les moyennes (variables quantitatives) par des effectifs de réponses (variables qualitatives).

```
f1 <- function(x) {
    l0 <- lapply(x, function(y) factor(y, levels = 1:5))
    l0 <- data.frame(l0)
    row.names(l0) <- row.names(x)
    l0 <- acm.disjonctif(l0)
    l0
}
didac <- f1(pps002[[1]])
for (k in 2:11) didac <- didac + f1(pps002[[k]])
head(acm.disjonctif(pps002[[1]]))
```

```
      prec.1 prec.2 prec.3 prec.4 dur.1 dur.2 dur.3 dur.4 dur.5 subj.1 subj.3 subj.4
algeb      1      0      0      0     1     0     0     0     0      0      0      0
astro      0      0      1      0     0     0     0     0     1      1      0      0
biomo      0      0      1      0     1     0     0     0     0      0      0      1
didac      0      0      1      0     0     0     0     1     0      0      0      1
ecolo      0      0      1      0     0     0     1     0     0      0      0      1
ethol      0      0      0      1     1     0     0     0     0      0      0      0
      subj.5 faux.2 faux.3 faux.4 faux.5 faib.1 faib.3 faib.4 faib.5 fant.3 fant.4
algeb      1      0      0      0      1      0      0      0      1      0      0
astro      0      0      1      0      0      0      1      0      0      1      0
biomo      0      0      0      1      0      0      0      1      0      0      1
didac      0      0      1      0      0      1      0      0      0      0      0
ecolo      0      0      0      0      1      0      1      0      0      0      0
ethol      1      0      1      0      0      0      1      0      0      1      0
      fant.5
algeb      1
astro      0
biomo      0
didac      1
ecolo      1
ethol      0
```

```
head(didac)
```

```
      prec.1 prec.2 prec.3 prec.4 prec.5 dur.1 dur.2 dur.3 dur.4 dur.5 subj.1 subj.2
algeb     10      1      0      0      0     1     4     3     2     1      0      0
astro      0      2      2      2      5     0     2     3     1     5      9      1
biomo      5      4      2      0      0     6     1     4     0     0      0      2
didac      0      2      2      5      2     0     1     5     2     3      4      2
ecolo      0      5      2      3      1     0     2     7     2     0      1      1
ethol      0      4      4      3      0     2     1     6     1     1      3      2
      subj.3 subj.4 subj.5 faux.1 faux.2 faux.3 faux.4 faux.5 faib.1 faib.2 faib.3
algeb      1      4      6      1      0      1      2      7      1      1      3
astro      1      0      0      5      3      1      2      0      5      2      2
biomo      2      5      2      0      0      1      7      3      1      1      4
didac      1      3      1      1      0      3      6      1      3      1      4
ecolo      3      4      2      0      1      2      3      5      2      0      5
ethol      2      3      1      0      0      6      4      1      0      0      9
      faib.4 faib.5 fant.1 fant.2 fant.3 fant.4 fant.5
algeb      2      4      0      1      0      4      6
astro      2      0      6      2      2      1      0
biomo      3      2      1      1      1      3      5
didac      0      3      1      1      1      3      5
ecolo      4      0      0      4      1      3      3
ethol      2      0      1      3      3      3      1
```

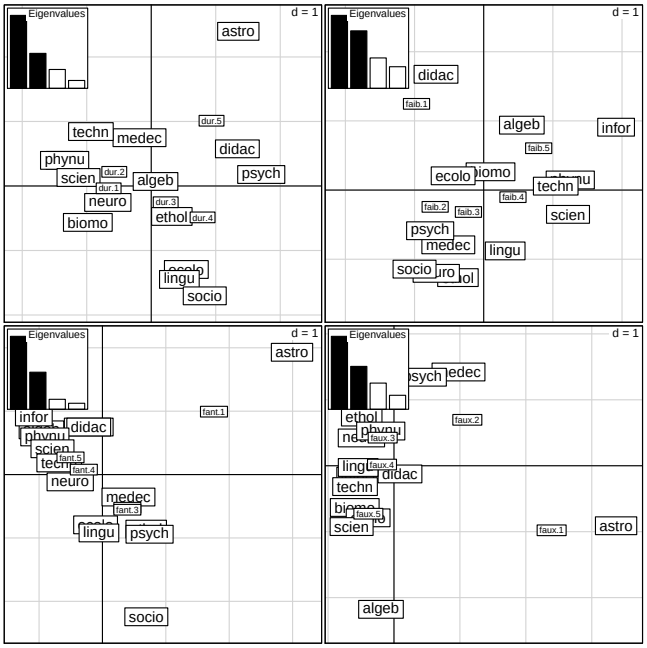
La forme des calculs indique bien qu'on est, pour les deux types de variables, dans la même logique . Compter les porteurs par modalités c'est faire une moyenne en gardant l'originalité de ce type de variables.

Ceci n'est pas une table de contingence au sens strict, dans la mesure où une des marges est connue avant l'expérience : il y a toujours 11 réponses par discipline et par couple d'adjectifs. Mais c'est un tableau d'association entre un mode de jugement (1 à 5) et une discipline. L'association précis-algèbre s'oppose clairement à l'association imprécis-psychologie. On a compté les correspondances notes-disciplines dans cette observation planifiée.

De plus c'est un tableau multiple dont on voit bien la structure en bloc. Il a la structure d'un tableau disjonctif complet mais l'indication 0,1,0,0,0 d'une modalité et d'une seule est remplacée par par une distribution de fréquence n_1, n_2, n_3, n_4, n_5.

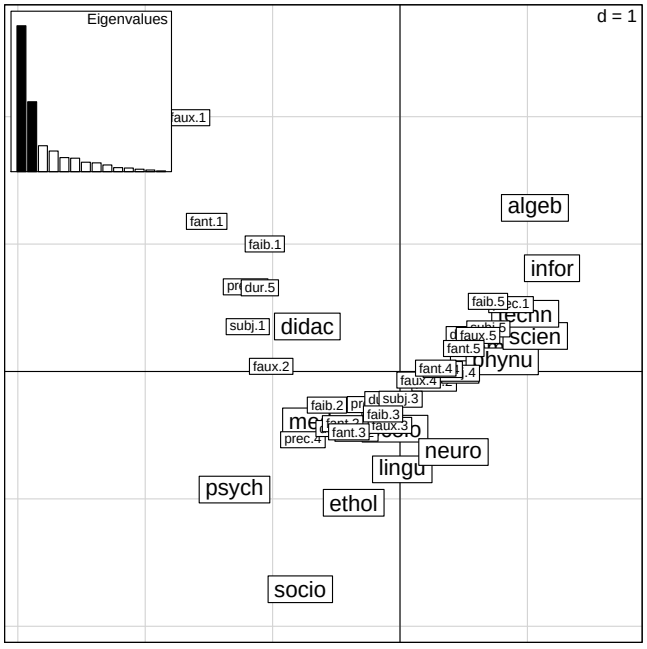
Chaque morceau relève d'une AFC :

```
par(mfrow = c(2, 2))
cla <- factor(substr(names(didac), 1, 3))
for (k in 1:4) {
    scatter(dudi.coa(didac[, cla == levels(cla)[k]], scannf = F),
        met = 3)
}
```



La coordination se fait par l'AFC du total qui est alors une extension de l'ACM au non disjonctif [5] :

```
scatter(dudi.coa(didac, scannf = F), met = 3)
```



Au total, on retrouve :

– l'ordination typique des sciences dures aux sciences molles, bien installée dans la tête des étudiants par la tradition universitaire ;

– un point aberrant, l'astrologie qui introduit une torsion comme point extrême vers le mou et point aberrant sur l'ensemble ;
– une position ambiguë de la didactique, discipline propre aux personnes interrogées.

Ceci fonctionne bien parce que l'AFC d'un tel tableau (somme par lignes et par blocs constantes) est une AFC intra-classe [1] :

```
w1 <- dudi.coa(didac, scannf = F)
w2 <- witwit.coa(w1, 15, rep(5, 6), scannf = F)
w1$eig
```

```
 [1] 0.327107554 0.156707019 0.058164980 0.046317240 0.031580904 0.030216052
 [7] 0.021198721 0.019753486 0.015045712 0.009069020 0.008184764 0.005534006
[13] 0.003736485 0.001705729
```

```
w2$eig
```

```
 [1] 0.327107554 0.156707019 0.058164980 0.046317240 0.031580904 0.030216052
 [7] 0.021198721 0.019753486 0.015045712 0.009069020 0.008184764 0.005534006
[13] 0.003736485 0.001705729
```

Ce n'est pas vrai pour les tableaux d'usage du code (séquences et acides aminés par codons) ou pour les tableaux populations et allèles par loci mais cela ne gène personne parce que *ça marche à peu près*. Pourquoi pas ? Mais à partir de l'à peu près la pente est forte, vers les abus manifestes [7] d'abord, puis les grosses bêtises.

Références

[1] J.P. Benzécri. Sur la généralisation du tableau de burt et son analyse par bandes. *Les Cahiers de l'Analyse des Données*, VII :33–43, 1982.

[2] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :498–520, 1933.

[3] B. Le Roux and H. Rouanet. Interpreting axes in multiple correspondence analysis : method of the contributions of points and deviation. In Blasius J. and M. Greenacre, editors, *Visualization of categorical data*, pages 197–220. Acamedic Press, London, 1997.

[4] L. Lebart, A. Morineau, and M. Piron. *Statistique exploratoire multidimensionnelle*. Dunod, Paris, 1995.

[5] A. Leclerc. L'analyse des correspondances sur juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, XXIII :5–16, 1975.

[6] A. Lucas and S. Jasson. Using amap and ctc packages for huge clustering. *R News*, 6(5) :58–60, 2006.

[7] G. Perrière and J. Thioulouse. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Research*, 30(20) :4548–4555, 2002.

[8] H. Rouanet and B. Le Roux. *Analyse des données multidimensionnelles*. Dunod, paris, 1993.

[9] G. Saporta. *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. PhD thesis, Thèse de 3° cycle, Université Pierre et Marie Curie, 1975.

[10] G. Saporta. *Probabilités, analyse des données et statistique*. Technip, Paris, 1990.

[11] M. Tenenhaus and F.W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. *Psychometrika*, 50 :91–119, 1985.

[12] J. van Rijckevorsel. *The application of fuzzy coding and hoerseshoes in multiple correspondence analysis*. DSWO Press, Leiden, 1987.