Analyse des Correspondances Multiples

Anne B Dufour

UCB Lyon1

Décembre 2012

Analyse des Correspondances multiples

- Les données : un tableau de variables qualitatives $\mathbf{q}^j$
- La plus simple des méthodes K-tableaux

Objectif

Rechercher une variable synthétique maximisant la somme des carrés des rapports de corrélation

$$\frac{1}{\nu}\sum_{j=1}^{\nu}\eta^2(\mathbf{q}^j,\mathbf{z})$$

Présentation des données

Sous-échantillon d'une étude sur les conditions de vie et d'aspirations des Français (1981) :

- 105 individus
- 9 variables
  1. sexe - `sex` : féminin (F), masculin (M)
  2. âge - `age` exprimé en années
  3. La famille - `fam` est le seul endroit où l'on se sent bien : oui, non
  4. Les dépenses de logement - `dep` sont pour vous : negligeable (NEG0), sans gros problème (SGP1), une lourde charge (LC2), une très lourde charge (TLC3)
  5. Disposez-vous d'un magnétoscope - `mag` ? oui, non
  6. Avez-vous souffert récemment de maux de tête - `mdt` ? oui, non
  7. Avez-vous souffert récemment de mal de dos - `mdd` ? oui, non
  8. Vous imposez-vous régulièrement des restrictions - `res` ? oui, non
  9. Regardez-vous la télévision - `tve` ? tous les jours (TLJ3), assez souvent (AS2), pas très souvent (PTS1), jamais (JAM0)

Présentation des données

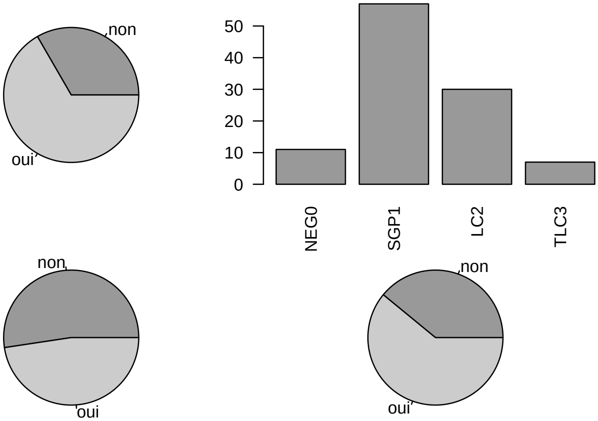
Il est extrait du livre de Lebart, Morineau, Piron (1995)

```
library(ade4)
cvaf <- read.table("http://pbil.univ-lyon1.fr/R/donnees/cvaf.txt", h=T)
cvaf[1:18,]
```

```
   sex age fam  dep mag mdt mdd res  tve
1    F  27 oui SGP1 non oui oui non TLJ3
2    F  42 oui  LC2 non non oui oui PTS1
3    M  71 oui SGP1 non non non oui TLJ3
4    M  52 oui SGP1 non oui oui non TLJ3
5    F  36 oui SGP1 non non non oui PTS1
6    M  22 non SGP1 non non oui non PTS1
7    M  26 non SGP1 non non non non  AS2
8    F  43 oui SGP1 oui oui non non TLJ3
9    F  33 oui SGP1 non non non oui TLJ3
10   F  54 non TLC3 non non oui oui PTS1
11   M  57 non  LC2 non oui oui non PTS1
12   M  33 oui SGP1 non oui oui oui TLJ3
13   M  65 oui SGP1 non non oui non TLJ3
14   F  58 oui SGP1 non non non non  AS2
15   F  33 non  LC2 non oui non oui TLJ3
16   M  37 oui TLC3 non non non oui TLJ3
17   M  46 oui  LC2 non non oui oui  AS2
18   F  30 non  LC2 non oui non oui TLJ3
```

Présentation des données

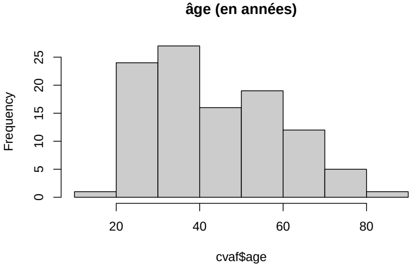
4 variables actives : famille (3), dépenses de logement (4), mal de dos (7), restrictions (8)



Présentation des données

5 variables illustratives : sexe (1), âge (2), magnétoscope (5), maux de tête (6), télévision (9)

```
sex          age            mag       mdt        tve
F:52   Min.   :19.00     non:83    non:72    AS2 :27
M:53   1st Qu.:32.00     oui:22    oui:33    JAMO: 3
       Median :41.00                         PTS1:22
       Mean   :43.89                         TLJ3:53
       3rd Qu.:54.00
       Max.   :82.00
```



Plusieurs manières de voir le tableau

Tableau Brut :

```
cvaf2 <- cvaf[,c(3,4,7,8)]
cvaf2[1:3,]
```

```
  fam  dep mdd res
1 oui SGP1 oui non
2 oui  LC2 oui oui
3 oui SGP1 non oui
```

Tableau disjonctif complet :

```
cvaf2.disj <- acm.disjonctif(cvaf[,c(3,4,7,8)])
cvaf2.disj[1:3,]
```

```
  fam.non fam.oui dep.LC2 dep.NEG0 dep.SGP1 dep.TLC3 mdd.non mdd.oui res.non res.oui
1       0       1       0        0        1        0       0       1       1
2       0       1       1        0        0        0       0       1       0
3       0       1       0        0        1        0       1       0       0
```

Partitionnement du tableau

Soit un tableau $\mathbf{X}$ constitué de ν variables observées sur n individus. Chaque variable constitue des paquets d'indicatrices. La juxtaposition de ces paquets constitue un tableau disjonctif complet.

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \ldots \ \mathbf{X}_\nu]$$

0	1	0	0	1	0	0	1	1	0
0	1	1	0	0	0	0	1	0	1
0	1	0	0	1	0	1	0	0	1
...	...	...	...	...	...	...	...	...	...
0	1	0	1	0	0	0	1	1	0
0	1	0	0	1	0	1	0	1	0
0	1	0	0	1	0	0	1	1	0

```
matX <- as.matrix(cvaf2.disj)
```

Information globale sur les modalités

- ν est le nombre de variables
- m_j est le nombre de modalités lié à la variable j
- m est le nombre total de modalités : $m = \sum_{j=1}^{\nu} m_j$

Exemple.
La variable 3 possède 2 modalités : $m_3 = 2$
La variable 4 possède 4 modalités : $m_4 = 4$
La variable 7 possède 2 modalités : $m_7 = 2$
La variable 8 possède 2 modalités : $m_8 = 2$
L'ensemble des modalités des 4 variables actives est : $m = 2 + 4 + 2 + 2$ soit $m = 10$.

Pondérations des lignes et des colonnes

- Pondération des lignes
  Soit i une des n lignes du tableau : $p_i = \frac{1}{n}$

$$\sum_{i=1}^{n} p_i = 1$$

```
round(1/105,6)
[1] 0.009524
```

- Pondération des colonnes
  On compte pour chaque modalité le nombre d'individus associé :

$$\mathbf{X}^T \mathbf{1}_n = [n_{11}\; n_{12}\; \ldots\; n_{1m_1}\; n_{21}\; n_{22}\; \ldots\; n_{2m_2}\; \ldots\; n_{\nu 1}\; n_{\nu 2}\; \ldots\; n_{\nu m_\nu}]$$

```
apply(matX,2,sum)
fam.non  fam.oui  dep.LC2 dep.NEGO dep.SGP1 dep.TLC3  mdd.non  mdd.oui  res.n
     35       70       30       11       57        7       55       50
res.oui
     64
```

Pondérations des lignes et des colonnes

On a la relation simple :

$$\begin{aligned} n &= n_{11} + n_{12} + \cdots + n_{1m_1} \\ &= n_{21} + n_{22} + \cdots + n_{2m_2} \\ &= n_{\nu 1} + n_{\nu 2} + \cdots + n_{\nu m_\nu} \end{aligned}$$

Le poids d'une modalité est la fréquence du nombre de porteurs et on obtient :

$$\begin{aligned} 1 &= f_{11} + f_{12} + \cdots + f_{1m_1} \\ &= f_{21} + f_{22} + \cdots + f_{2m_2} \\ &= f_{\nu 1} + f_{\nu 2} + \cdots + f_{\nu m_\nu} \end{aligned}$$

Ecriture matricielle

- Tableau disjonctif complet des données : $\mathbf{X}$
- Pondération des lignes : $\mathbf{D} = diag(\frac{1}{n}\ \frac{1}{n}\ \ldots\ \frac{1}{n})$
- Pondération des colonnes :

$$\mathbf{X}^T\mathbf{1}_n = [n_{11}\ n_{12}\ \ldots\ n_{1m_1}\ n_{21}\ n_{22}\ \ldots\ n_{2m_2}\ \ldots\ n_{\nu 1}\ n_{\nu 2}\ \ldots\ n_{\nu m_\nu}]$$

$$\mathbf{X}^T\mathbf{D}\mathbf{1}_n = [f_{11}\ f_{12}\ \ldots\ f_{1m_1}\ f_{21}\ f_{22}\ \ldots\ f_{2m_2}\ \ldots\ f_{\nu 1}\ f_{\nu 2}\ \ldots\ f_{\nu m_\nu}]$$

$$\mathbf{D}_m = diag(f_{11}\ f_{12}\ \ldots\ f_{1m_1}\ f_{21}\ f_{22}\ \ldots\ f_{2m_2}\ \ldots\ f_{\nu 1}\ f_{\nu 2}\ \ldots\ f_{\nu m_\nu})$$

La somme des éléments de la diagonale vaut ν.

- On note $\mathbf{1}_{nm}$ la matrice de dimensions $n \times m$ ne contenant que des 1.

Exemple de pondérations des lignes et des colonnes

- Pondération des lignes

```
matD <- diag(rep(1/105,105))
matD[1:3,1:3]
```

```
           [,1]       [,2]       [,3]
[1,] 0.00952381 0.00000000 0.00000000
[2,] 0.00000000 0.00952381 0.00000000
[3,] 0.00000000 0.00000000 0.00952381
```

- Pondération des colonnes

```
vec105 <- rep(1, 105)
t(t(matX)%*%vec105)
```

```
     fam.non fam.oui dep.LC2 dep.NEGO dep.SGP1 dep.TLC3 mdd.non mdd.oui res.no
[1,]      35      70      30       11       57        7      55      50      4
     res.oui
[1,]      64
```

```
round(t(t(matX)%*%matD%*%vec105),2)
```

```
     fam.non fam.oui dep.LC2 dep.NEGO dep.SGP1 dep.TLC3 mdd.non mdd.oui res.no
[1,]    0.33    0.67    0.29      0.1     0.54     0.07    0.52    0.48    0.3
     res.oui
[1,]    0.61
```

```
t(matX)%*%matD%*%vec105 -> vecmod
matDm <- diag(vecmod[,1])
```

Schéma de dualité associé

$$
\begin{tikzcd}
\boxed{\mathsf{m}} \arrow[r, "\frac{1}{\nu}\mathbf{D}_m"] & \boxed{\mathsf{m}} \arrow[d, "\mathbf{Y}"] \\
\boxed{\mathsf{n}} \arrow[u, "\mathbf{Y}^T"] & \boxed{\mathsf{n}} \arrow[l, "\mathbf{D}"]
\end{tikzcd}
$$

$$\mathbf{Y} = \mathbf{X}\mathbf{D}_m^{-1} - \mathbf{1}_{nm}$$

Matrice des pondérations des variables

```
sum(matDm)
```

```
[1] 4
```

```
round((1/4)*matDm,2)
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,] 0.08 0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.0  0.00
 [2,] 0.00 0.17 0.00 0.00 0.00 0.00 0.00 0.00  0.0  0.00
 [3,] 0.00 0.00 0.07 0.00 0.00 0.00 0.00 0.00  0.0  0.00
 [4,] 0.00 0.00 0.00 0.03 0.00 0.00 0.00 0.00  0.0  0.00
 [5,] 0.00 0.00 0.00 0.00 0.14 0.00 0.00 0.00  0.0  0.00
 [6,] 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00  0.0  0.00
 [7,] 0.00 0.00 0.00 0.00 0.00 0.00 0.13 0.00  0.0  0.00
 [8,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.12  0.0  0.00
 [9,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.1  0.00
[10,] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00  0.0  0.15
```

Sens de la matrice Y

```
mat1nm <- matrix(1,ncol=10,nrow=105)
matDmm <- diag(1/vecmod[,1])
matY <- matX%*%matDmm - mat1nm
round(matY[1:7,],2)
```

```
  [,1] [,2] [,3] [,4]  [,5] [,6]  [,7] [,8]  [,9] [,10]
1   -1  0.5 -1.0   -1  0.84   -1 -1.00  1.1  1.56 -1.00
2   -1  0.5  2.5   -1 -1.00   -1 -1.00  1.1 -1.00  0.64
3   -1  0.5 -1.0   -1  0.84   -1  0.91 -1.0 -1.00  0.64
4   -1  0.5 -1.0   -1  0.84   -1 -1.00  1.1  1.56 -1.00
5   -1  0.5 -1.0   -1  0.84   -1  0.91 -1.0 -1.00  0.64
6    2 -1.0 -1.0   -1  0.84   -1 -1.00  1.1  1.56 -1.00
7    2 -1.0 -1.0   -1  0.84   -1  0.91 -1.0  1.56 -1.00
```

```
round(sapply(as.data.frame(matY),mean),2)
```

```
V1  V2  V3  V4  V5  V6  V7  V8  V9 V10
 0   0   0   0   0   0   0   0   0   0
```

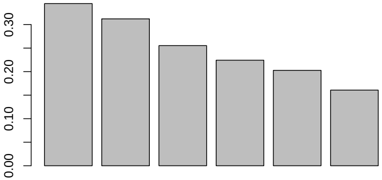
```
round(sapply(as.data.frame(matY),var)*104/105,2)
```

```
  V1    V2    V3    V4    V5    V6    V7    V8    V9   V10
2.00  0.50  2.50  8.55  0.84 14.00  0.91  1.10  1.56  0.64
```

Valeurs propres

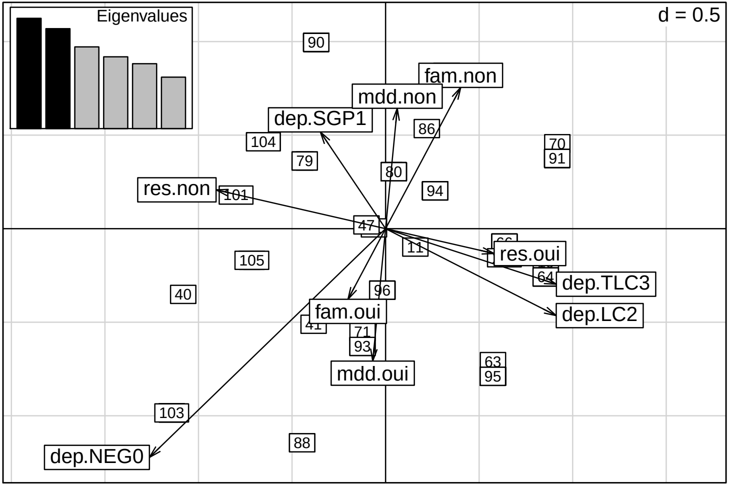
Nombre de valeurs propres : $m - \nu$ ou encore $(2-1)+(4-1)+(2-1)+(2-1) = 6$

```
acm <- dudi.acm(cvaf[,c(3,4,7,8)],scannf=F,nf=6)
barplot(acm$eig)
```



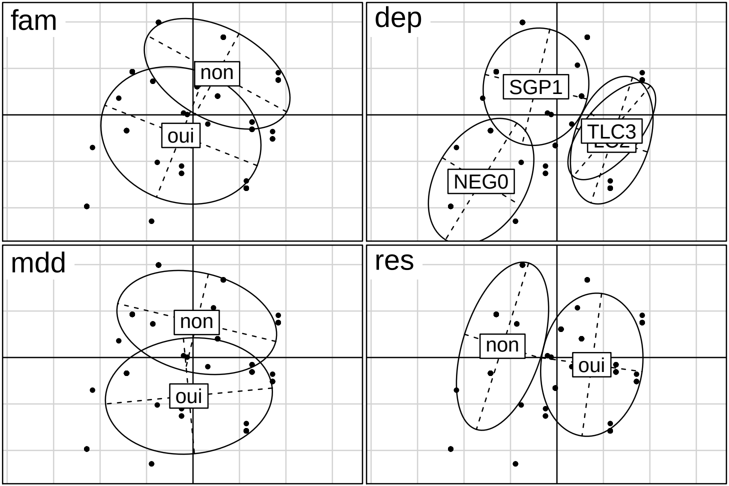
Représentation globale

```
scatter.dudi(acm)
```



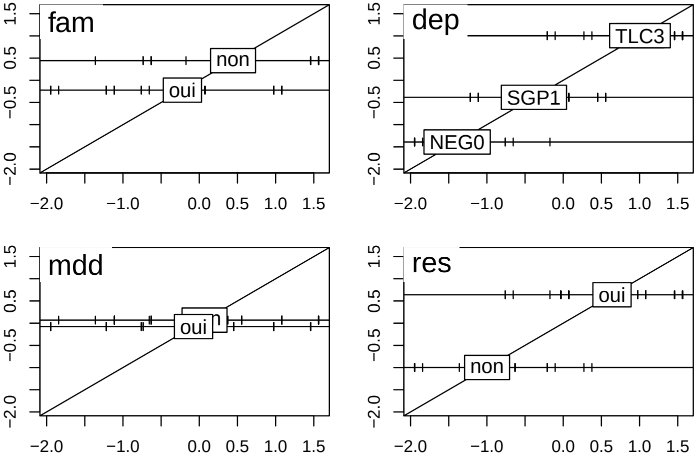
Variable par variable

```
scatter(acm)
```



Représentation sur un seul axe

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



Rapports de corrélation

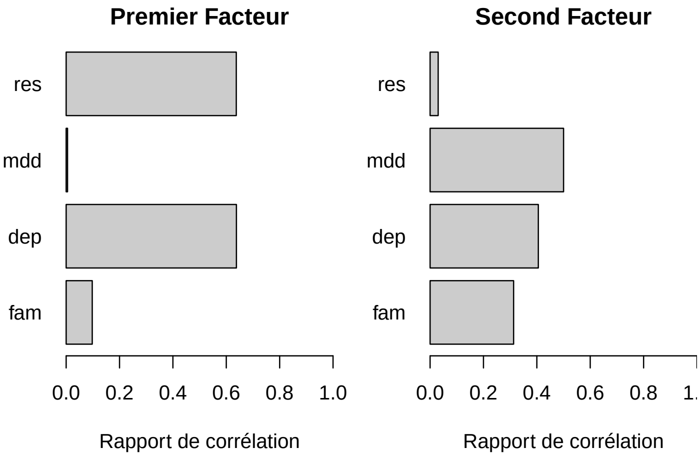


Tableau de Burt

On reprend le tableau disjonctif complet $\mathbf{X}$.

$$\mathbf{X}^T\,\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T\,\mathbf{X}_1 & \dots & \mathbf{X}_1^T\,\mathbf{X}_\nu \\ \vdots & \ddots & \vdots \\ \mathbf{X}_\nu^T\,\mathbf{X}_1 & \dots & \mathbf{X}_\nu^T\,\mathbf{X}_\nu \end{bmatrix}$$

- Le bloc diagonal $\mathbf{X}_j^T\,\mathbf{X}_j$ donne sur sa diagonale les sommes marginales de la variable j.
- Le bloc non diagonal $\mathbf{X}_j^T\,\mathbf{X}_k$ est la table de contingence entre les variables j et k.

Tableau de Burt - exemple

```
cvcv <- acm.burt(cvaf[,c(3,4,7,8)],cvaf[,c(3,4,7,8)])
colnames(cvcv) <- c("fnon","foui","dlc","dn","dsgp","dtlc","mddnon","mddoui","rnon
rownames(cvcv) <- colnames(cvcv)
cvcv[1:2,1:2]
```

```
     fnon foui
fnon   35    0
foui    0   70
```

```
cvcv[3:6,3:6]
```

```
     dlc dn dsgp dtlc
dlc   30  0    0    0
dn     0 11    0    0
dsgp   0  0   57    0
dtlc   0  0    0    7
```

```
cvcv[1:2,3:6]
```

```
     dlc dn dsgp dtlc
fnon  11  2   20    2
foui  19  9   37    5
```

AFC du tableau de Burt

```
afc <- dudi.coa(cvcv, scannf=F, nf=6)
afc$eig
```

```
[1] 0.11876533 0.09748052 0.06519261 0.05031634 0.04111429 0.02584048
```

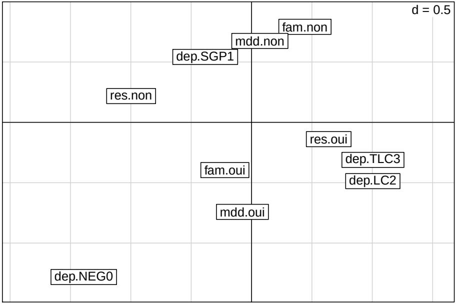
```
sqrt(afc$eig)
```

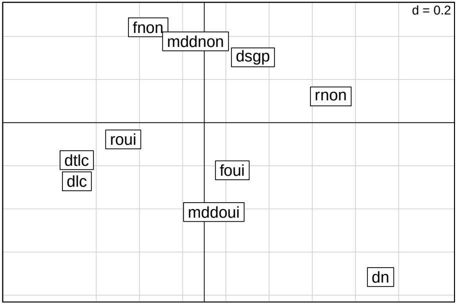
```
[1] 0.3446235 0.3122187 0.2553284 0.2243130 0.2027666 0.1607498
```

```
acm$eig
```

```
[1] 0.3446235 0.3122187 0.2553284 0.2243130 0.2027666 0.1607498
```

Lien entre ACM et AFC





Conclusion

Le tableau des données brutes est constitué des individus en lignes et des variables qualitatives en colonnes. On peut l'écrire de différentes façons :

- un tableau disjonctif complet,
- un tableau de Burt.

Ceci a des conséquences sur les liens entre AFC et ACM :
une AFC sur un tableau de Burt équivaut à une ACM sur un tableau disjonctif complet.