Analyse des Correspondances Multiples

GHAZI BEL MUFTI

ghazi.belmufti@gmail.com

ESSAI-1 / ANALYSE DES DONNÉES

Plan

Introduction

L'AFC introduite dans le chapitre précédent peut se généraliser de plusieurs façon au cas où plus de deux ensembles sont mis en correspondance. Une des généralisations la plus simple et la plus utilisée est *l' analyse des correspondances multiples* qui permet de décrire de vastes tableaux binaires, dont les fiches d'enquêtes socio-économiques constituent un exemple privilégié :

- les lignes de ces tableaux sont en général des individus ou observations ;
- les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponse à des questions.

Il s'agit, en fait d'une simple extension du domaine d'application de l'AFC, avec cependant des procédures de calcul et des règles d'interprétation spécifiques.

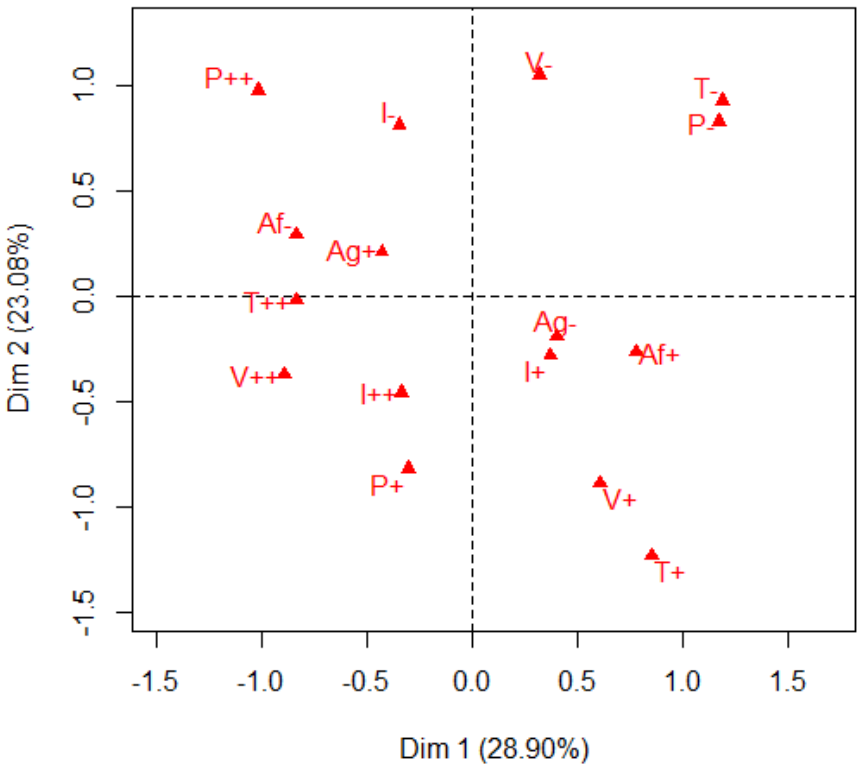
Exemple I

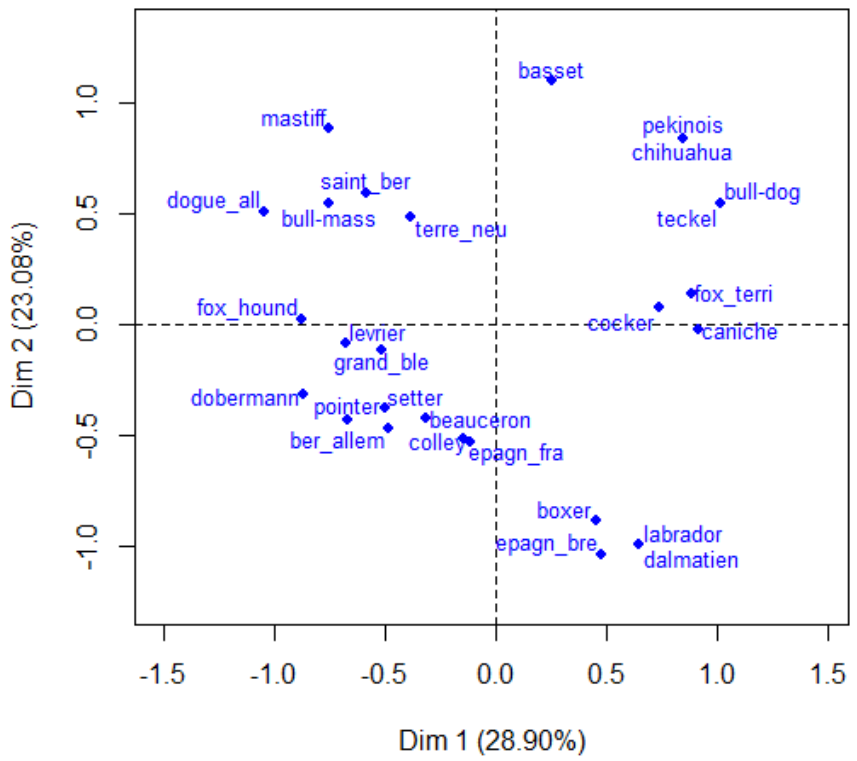
Les données `Canines` décrivent les caractéristiques de 27 races de chiens au moyens de 6 variables qualitatives. Ci-dessous les données des 6 premières races :

```
> head(Base_canines)
           taille poids velocite intellig affect agress
beauceron     T++    P+      V++       I+     Af+     Ag+
basset         T-    P-       V-       I-     Af-     Ag+
ber_allem     T++    P+      V++      I++     Af+     Ag+
boxer          T+    P+       V+       I+     Af+     Ag+
bull-dog       T-    P-       V-       I+     Af+     Ag-
bull-mass     T++   P++       V-      I++     Af-     Ag+
```

Cet exemple nous accompagnera tout au long de ce chapitre en illustrant les différentes étapes de l'ACM.

Exemple II


P++
I-
V-
T-
P-
Af-
Ag+
T++
Ag-
I+
Af+
V++
I++
P+
V+
T+
Dim 2 (23.08%)
Dim 1 (28.90%)
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


basset
mastiff
pekinois
chihuahua
saint_ber
dogue_all
bull-mass
terre_neu
bull-dog
teckel
fox_hound
fox_terri
cocker
caniche
levrier
grand_ble
dobermann
pointer
setter
beauceron
ber_allem
colley
epagn_fra
boxer
epagn_bre
labrador
dalmatien
Dim 2 (23.08%)
Dim 1 (28.90%)
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Notations

- Q : ensemble des questions ou de variables qualitatives, avec $m = |Q|$. Exemple : Le nombre de variable décrivant chaque race est $m = |Q| = 6$.
- I : ensemble des individus qui ont répondu aux questions, avec $n = |I|$. Exemple : Le nombre de race de chiens $n = 27$.
- J_q : ensemble de toutes les modalités de réponse à la question $q, 1 \leq q \leq m$, avec $|J_q|$ le nombre de modalités de la question q. Exemple : La première variable qui est `Taille` ayant 3 modalités qui sont T- (petite taille), T+ (taille moyenne) et T++ (grande taille), on a alors $|J_1| = 3$
- J : ensemble de toutes les modalités de réponse à toutes les question, avec $p = |J| = |J_1| + \cdots + |J_m|$. Exemple : Chacune des 4 premières variables ayant 3 modalités, les 2 dernières ayant chacune 2 modalités, on a $p = |J| = 3 + 3 + 3 + 3 + 2 + 2 = 16$.

Tableau disjonctif complet I

Le tableau k_{IJ} de taille $n \times p$ est appelé tableau disjonctif complet est défini par :

$$k(i,j) = \begin{cases} 1 & \text{si l'individu } i \text{ a adopté la modalité } j \text{ de } J, \\ 0 & \textit{sinon}. \end{cases}$$

	J_1	J_2	...	J_q	...	J_m	total
1							
⋮							
i	1 0 0	0 1		... k(i,j) ...			$k(i) = m$
⋮							
n							
total	...	...	...	... $k(j)$ ...	...	...	$k = nm$

Tableau disjonctif complet II

- **disjonctif :** car deux modalités j et j' d'une même question s'excluent mutuellement,
- **complet :** car à tout individu correspond une modalité de réponse à toute question q.

Exemple :

```
> head(tab.disjonctif(Base_canines))
           T- T+ T++ P- P+ P++ V- V+ V++ I- I+ I++ Af- Af+ Ag- Ag+
beauceron   0  0   1  0  1   0  0  0   1  0  1   0   0   1   0   1
basset      1  0   0  1  0   0  1  0   0  1  0   0   1   0   0   1
ber_allem   0  0   1  0  1   0  0  0   1  0  0   1   0   1   0   1
boxer       0  1   0  0  1   0  0  1   0  0  1   0   0   1   0   1
bull-dog    1  0   0  1  0   0  1  0   0  0  1   0   0   1   1   0
bull-mass   0  0   1  0  0   1  1  0   0  0  0   1   1   0   0   1
```

Propriétés des TDC

Pour tout individu $i \in I$, toute modalité $j \in J$ et toute question $q \in Q$, on a :

- $k(i) = \sum_{j \in J} k(i,j) = \sum_{q \in Q} \sum_{j \in J_q} k(i,j) = card(Q) = m$,
- $k(j) = \sum_{i \in I} k(i,j) =$ nombre d'individu ayant choisi la modalité j,
- $\sum_{j \in J_q} k(j) = n$,
- $k = \sum_{i \in I} \sum_{j \in J} k(i,j) = \sum_{i \in I} k(i) = nm$.

Principes de l'analyse des correspondances multiples

L'ACM est l'analyse des correspondances d'un tableau disjonctif complet.

Ses principes sont donc ceux de l'AFC à savoir :

- même transformation du tableau des données en profils-lignes (profils des individus) et en profils-colonnes (profils des modalités) ;
- même critère d'ajustement avec pondération des points par leurs profils marginaux ;
- même distance du χ^2.

L'ACM présente cependant des propriétés particulières dues à la nature même du TDC.

Matrice des poids

- Profil-lignes des Individus
  - Les individus sont tous affectés d'une masse identique égale à $\frac{m}{nm} = \frac{1}{n}$
  - On note $D_n = \frac{1}{n} I_n$ la matrice des poids des individus.

- Profil-colonnes des Modalités
  - Une modalité j est pondérée par sa fréquence $\frac{k(j)}{k} = \frac{k(j)}{nm}$
  - On note $D_p = Diag(\frac{k(j)}{nm})_{1 \leq j \leq p}$ la matrice des poids des modalités.

La distance du χ^2 appliquée à un TDC

- PL des individus : dans $\mathbb{R}^p$, la distance entre 2 individus i et i' s'exprime par

$$d^2(i,i') = \frac{1}{m}\sum_{j=1}^{p}\frac{n}{k(j)}(k(i,j)-k(i',j))^2$$

Cette distance conserve un sens dans la mesure où deux individus sont proches s'ils ont choisi les mêmes modalités et sont éloignés s'ils n'ont pas répondu de la même manière.

- PC des modalités : dans $\mathbb{R}^n$, la distance entre 2 modalités s'écrit

$$d^2(j,j') = \sum_{i=1}^{n} n(\frac{k(i,j)}{k(j)} - \frac{k(i,j')}{k(j')})^2$$

Ainsi deux modalités choisies par les mêmes individus coincident. Par ailleurs les modalités de faibles effectifs sont éloignées des autres modalités.

Composantes principales et relations quasi-barycentriques

- Les composantes principales ψ_α^I et ψ_α^J représentent respectivement les coordonnées des points-lignes (i.e. individus) et les coordonnées des points-colonnes (i.e. modalités) sur l'axe factoriel α. Elles on été obtenues de la même manière que pour l'AFC (voir chapitre précédent).
- Les relations de transition entre elles sont données par :

$$\begin{cases} \psi_\alpha^i = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{j=1}^{p} \dfrac{k(i,j)}{k(i)} \psi_\alpha^j \\ \\ \psi_\alpha^j = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{i=1}^{n} \dfrac{k(i,j)}{k(j)} \psi_\alpha^i \end{cases}$$

Inertie I

- On considère un tableau disjonctif complet. L'inertie totale est donnée par :

$$I_T = \sum_{j \in J} f_{.j} \rho^2(j) = \sum_{q \in Q} \sum_{j \in J_q} f_{.j} \rho^2(j),$$

où $\rho^2(j)$ est la distance du χ^2 entre une modalité (un profil colonne calculé à partir du TDC) et le profil moyen en colonne.

- On note p_j la proportion des individus ayant adopté la modalité j, on a

$$p_j = \frac{k(j)}{n}$$

- On alors $\rho^2(j) = \dfrac{1 - p_j}{p_j}$.

Inertie II

En effet

$$
\begin{aligned}
\rho^2(j) &= \sum_i \frac{1}{1/n}\left(\frac{k(i,j)}{k(j)} - \frac{k(i)}{k}\right)^2 \\
&= \sum_i n\left(\frac{k(i,j)}{k(j)} - \frac{1}{n}\right)^2 \\
&= \sum_i n\frac{k^2(i,j)}{k(j)^2} - 2\sum_i \frac{k(i,j)}{k(j)} + \sum_i \frac{1}{n} \\
&= \sum_i n\frac{k(i,j)}{k(j)^2} - 2 + 1 \\
&= \frac{n}{k(j)} - 2 + 1 \\
&= \frac{1}{p_j} - 1
\end{aligned}
$$

Inertie III

Comme $f_{.j} = \frac{k(j)}{k} = \frac{k(j)}{nm} = \frac{p_j}{m}$, où $m = Card(Q)$, on a :

$$\begin{aligned} I_T &= \sum_{j \in J} f_{.j} \rho^2(j) \\ &= \sum_{j \in J} \frac{p_j}{m} \frac{1 - p_j}{p_j} \\ &= \sum_{j \in J} \frac{1 - p_j}{m} \\ &= \frac{1}{m} [\sum_{j \in J} 1 - \sum_{q \in Q} \sum_{j \in J_q} p_j] \\ &= \frac{p}{m} - \frac{1}{m} \sum_{q \in Q} 1 = \frac{p}{m} - 1 \end{aligned}$$

Choix des axes à retenir

Ainsi

$$I_T = \frac{Card(J)}{Card(Q)} - 1$$

- L'inertie est donc égale au nombre moyen de catégories diminué d'une unité.
- Sachant que si $n \geq p - m$, on a au plus $p - m$ valeurs propres non nuls et non triviales, la moyenne de ces valeurs propres vaut $\frac{\frac{p}{m} - 1}{p - m} = 1/m$.
- La quantité $1/m$ peut donc jouer le rôle de seuil d'élimination pour les valeurs propres (exactement comme la valeur 1 du critère de Kaiser pour l'ACP normée) : on retiendra les axes dont les valeurs propres sont supérieures à ce seuil.

Contributions en ACM

En plus des cos2 et des contributions, par axe, aussi bien pour les individus que pour les modalités, l'ACM permet d'avoir les contributions suivantes à l'inertie totale :

- La contribution de la modalité j à l'inertie totale est donnée par :

$$CR(j) = f_{.j}\rho^2(j) = \frac{1 - p_j}{Card(Q)},$$

- La contribution de la question J_q à l'inertie totale est donnée par :

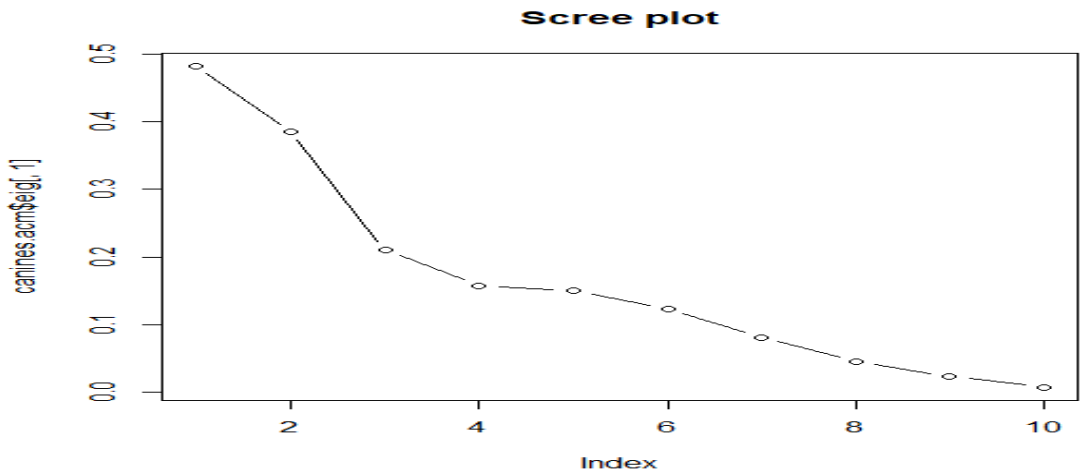
$$CR(J_q) = \sum_{j \in J_q} f_{.j}\rho^2(j) = \frac{Card(J_q) - 1}{Card(Q)},$$

- On retrouve bien

$$I_T = \sum_{q \in Q} CR(J_q) = \frac{Card(J)}{Card(Q)} - 1.$$

Exemple : ACM sur les données Canines

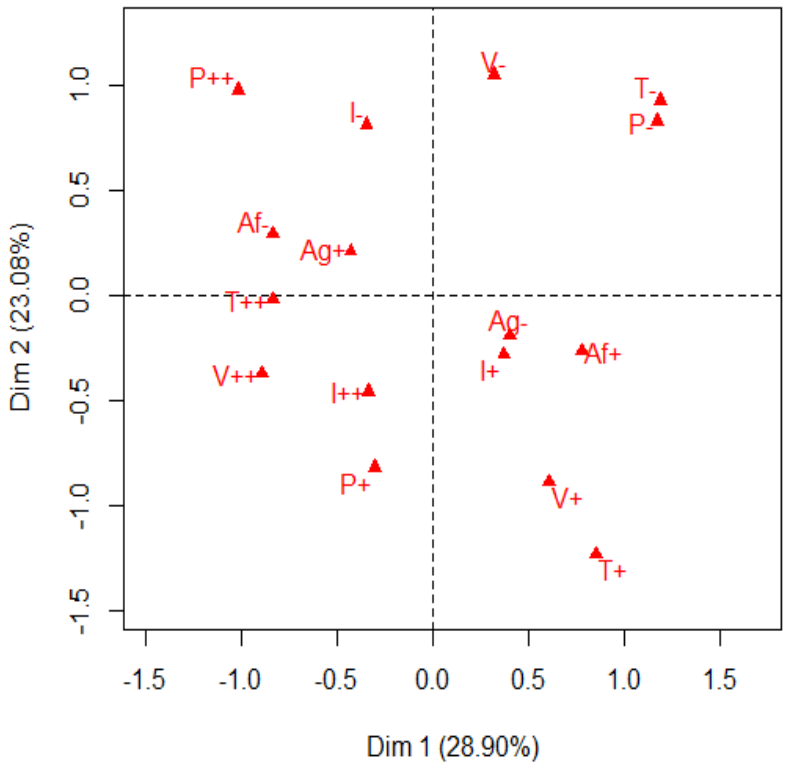
```
> library(FactoMineR)
> canines.acm <- MCA(Base_canines,ncp=2,graph=T)
> print(canines.acm$eig) #valeurs propres
        eigenvalue percentage of variance cumulative percentage of variance
dim 1   0.481606165                28.896370                          28.89637
dim 2   0.384737288                23.084237                          51.98061
dim 3   0.210954049                12.657243                          64.63785
dim 4   0.157554025                 9.453242                          74.09109
dim 5   0.150132670                 9.007960                          83.09905
dim 6   0.123295308                 7.397718                          90.49677
dim 7   0.081462460                 4.887748                          95.38452
dim 8   0.045669757                 2.740185                          98.12470
dim 9   0.023541911                 1.412515                          99.53722
dim 10  0.007713034                 0.462782                         100.00000
```



Choix du nombre d'axes à retenir

- Le nombre d'axes de l'ACM : le nombre de modalités étant $p = 16$ et le nombre de question est $m = 6$, ceci qui conduit à $p - m = 16 - 6 = 10$ axes.
- L'inertie totale vaut $p/m - 1 = 16/6 - 1 = 5/3 = 1.66$.
- Le critère qui consiste à retenir les axes dont la valeur propre est supérieure au seuil $1/m = 1.66/10 = 0.16$, conduit à retenir trois axes.
- D'autre part, le diagramme montre une chute au niveau du deuxième axe : on interprètera les 2 premiers axes.

Interprétation de la carte des modalités

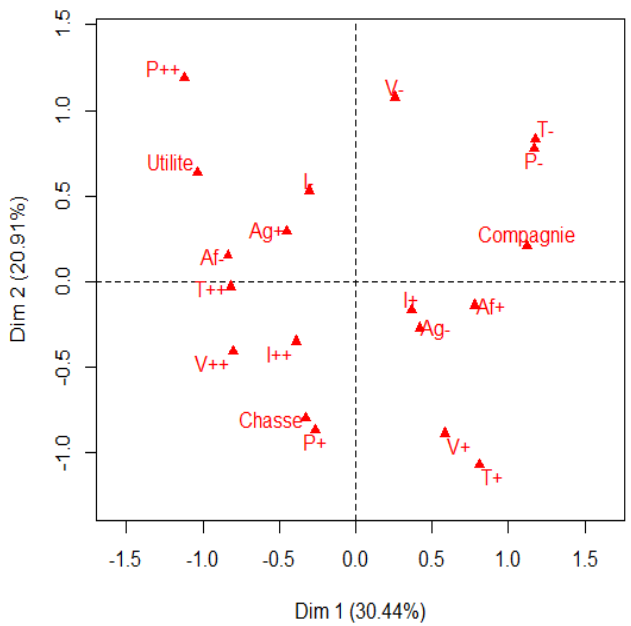


- L'axe 1 oppose (à droite) les chiens de petite taille (T-), affectueux (Af+), aux chiens de grande taille (T++), très rapide (V++) et agressifs (Ag+).
- L'axe 2 oppose (en bas) les chiens de taille moyenne (T+), très intelligents (I++) à des chiens lents (V-) et peu intelligents (I-).

Aide à l'interprétation : variable supplémentaire

Nous avons rajouté la variable "`fonction`" (à 3 modalités : chien de "compagnie", chien de "chasse" et chien d' "utilité") comme variable supplémentaire à l'ACM.

- L'axe 1 oppose les chiens de petite taille, affectueux, qui coincident avec les chiens de compagnie, aux chiens de grande taille, très rapide et agressifs qui sont les chiens d' utilité.
- L'axe 2 oppose les chiens de chasse, de taille moyenne, très intelligents à des chiens lents et peu intelligents.



Interprétation de la carte des individus

- L'axe 1 oppose les chiens de compagnie comme les Caniches et les Pékinois aux chiens d' utilité comme les Dogs Allemands et les Dobermans mais aussi les Saint-Bernards souvent dressés comme chiens de recherche en avalanche.
- L'axe 2 oppose les chiens de chasse comme les Dalmasiens et les Labradors à des chiens lents et peu intelligents comme les Bassets.