TP2 : AFC et AFCM

1 Mise en place

Cette séance a pour objectifs de mettre en place la procédure d'analyse factorielle des correspondances (AFC) vue en cours et d'introduire l'analyse factorielle des correspondances multiples (AFCM).

Afin de faciliter la récupération des données et de disposer de certaines fonctions graphiques dans la suite de cette séance, nous commençons par charger le script *tp2.R*,

```
source("http://www.math.univ-toulouse.fr/~xgendre/ens/m2se/tp2.R")
```

2 Études et catégories socioprofessionnelles (AFC)

Nous commençons par reprendre les données utilisées en cours et issues du travail du sociologue Pierre Bourdieu,

```
T <- DataBourdieu()
```

La table de contingence `T` contient les effectifs des étudiants de l'enquête. En ligne, nous avons les différentes catégories socioprofessionnelles pour le chef de famille (`EAG` - exploitant agricole, `SAG` - salarié agricole, `PT` - patron, `PLCS` - profession libérale et cadre supérieur, `CM` - cadre moyen, `EMP` - employé, `OUV` - ouvrier et `AUT` - autre). En colonne, nous retrouvons les types d'études poursuivies (`DR` - droit, `SCE` - sciences économiques, `LET` - lettres, `SC` - sciences, `MD` - médecine ou dentaire, `PH` - pharmacie, `PD` - pluridisciplinaire et `IUT` - IUT).

La première étape consiste à calculer les profils lignes et colonnes associés à ces données. Nous spécifions également le noms des catégories associées à chaque ligne pour faciliter les représentations graphiques dans la suite.

```
P1 <- diag(1/rowSums(T)) %*% T
rownames(P1) <- rownames(T)
P2 <- diag(1/colSums(T)) %*% t(T)
rownames(P2) <- colnames(T)
```

L'ACP des profils lignes se fait comme dans la séance précédente avec la matrice des poids $W_1 = \text{diag}(n_{1\cdot}/n, \ldots, n_{p\cdot}/n)$ et la matrice $M_1 = \text{diag}(n/n_{\cdot 1}, \ldots, n/n_{\cdot q})$. Exécuter les lignes suivantes et expliquer pourquoi ce qu'elles font est équivalent à cette ACP.

```
n <- sum(rowSums(T))
M1half <- diag(sqrt(n/colSums(T)))
X <- P1 %*% M1half
ACP1 <- eigen(t(X) %*% diag(rowSums(T)/n) %*% X)
```

Que vaut `ACP1$values[1]` ? Rappeler pourquoi il est nécessaire de procéder aux modifications suivantes,

```
k <- min(nrow(T), ncol(T))
ACP1$values <- ACP1$values[2:k]
ACP1$vectors <- ACP1$vectors[, 2:k]
C1 <- X %*% ACP1$vectors
```

Grâce aux formules de transition, nous pouvons déduire directement l'ACP sur les profils colonnes et ainsi obtenir la matrice des composantes principales associée à cette seconde ACP.

```
C2 <- P2 %*% C1 %*% diag(1/sqrt(ACP1$values))
```

Calculer la part d'inertie expliquée par le plan principal et la qualité de la représentation de chaque catégorie socioprofessionnelle et de chaque type d'étude suivi. Commenter ces valeurs. La représentation graphique de l'AFC peut se faire avec la commande `PlotAFC`,

```
PlotAFC(C1, C2)
```

Commenter et interpréter le résultat obtenu.

Bibliothèque pour l'AFC

Il existe de nombreux outils pour automatiser le calcul et la représentation de l'AFC. Parmi eux, le paquet `ca` est souvent utilisé et est facile à prendre en main. Si ce paquet n'est pas présent sur votre machine, vous pouvez l'installer en local grâce à la commande `install.packages("ca")`. Ensuite, les commandes suivantes fournissent les résultats de l'AFC ainsi que sa représentation graphique,

```
library(ca)
AFC <- ca(as.table(T))
print(AFC)
summary(AFC)
plot(AFC)
```

3 Élections présidentielles de 1995 (AFC)

Le second exemple d'AFC que nous considérons est lié aux résultats du premier tour des élections présidentielles françaises de 1995. Pour chaque département (les deux départements de la Corse sont agrégés), nous disposons du nombre des inscrits, du nombre de votants, du nombre de suffrages exprimés et des nombres de voix reçues par chaque candidat. Charger les données comme suit et examiner les brièvement.

```
data <- DataElection1995()
```

La table de contingence à étudier est celle fournie par les nombres de voix de chaque candidat dans chaque département.

```
T <- as.matrix(data[, 4:12])
```

Faire l'AFC et commenter les résultats obtenus. En particulier, que dire de la Vendée (département 85) ? Examiner la ligne `data[85,]` pour vous aider à répondre. Refaire l'AFC sans ce département en utilisant la table donnée par `as.matrix(data[-85,4:12])`.

4 AFCM : un peu de théorie

L'analyse factorielle des correspondances multiples (AFCM) est la généralisation de l'AFC pour un nombre de variables qualitatives supérieur à deux. Nous supposons ici que nous disposons de $m > 2$ variables qualitatives $x^1, \ldots, x^m$ que nous observons sur n individus. Pour $k \in \{1, \ldots, m\}$, la variables x^k peut prendre c_k valeurs distinctes notées $1, \ldots, c_k$ et c désigne le nombre total de modalités,

$$c = \sum_{k=1}^{m} c_k \ .$$

Pour $i \in \{1, \ldots, n\}$, $x_i^k \in \{1, \ldots, c_k\}$ est la valeur de la variable x^k pour l'individu i.

Pour chaque modalité de chaque variable, nous pouvons considérer le vecteur indicateur associé, *i.e.* pour tout $k \in \{1, \ldots, m\}$ et tout $\ell \in \{1, \ldots, c_k\}$,

$$\mathbb{I}_{m\ell} = \begin{pmatrix} \mathbf{1}_{x_1^k = \ell} \\ \vdots \\ \mathbf{1}_{x_n^k = \ell} \end{pmatrix} \in \{0,1\}^n \ .$$

Ces vecteurs mis en colonne d'une matrice X de taille $n \times c$ donne le **tableau disjonctif complet**,

$$X = \left[\mathbb{I}_{11} \cdots \mathbb{I}_{1c_1} \mathbb{I}_{21} \cdots \mathbb{I}_{2c_2} \cdots \mathbb{I}_{m1} \cdots \mathbb{I}_{mc_m} \right] \ .$$

Enfin, nous définissons l'objet central de l'AFCM, le **tableau de Burt** $\mathcal{B}$, qui est la version symétrique de X,

$$\mathcal{B} = {}^t X X \ .$$

Il est possible de voir $\mathcal{B} = [B_{kk'}]_{1 \leqslant k,k' \leqslant m}$ comme une matrice de $m \times m$ blocs avec

- si $k \neq k'$, $B_{kk'}$ est la table de contingence associée aux variables x^k et $x^{k'}$,
- si $k = k'$, B_{kk} est une matrice diagonale contenant les effectifs marginaux de la variable x^k, notés $n_1^k, \ldots, n_{c_k}^k$.

Pour réaliser l'AFCM, il faut considérer l'ACP des profils lignes du tableau de Burt,

$$P = \frac{1}{m} D\mathcal{B}$$

avec $D = \mathrm{diag}(1/n_1^1, \ldots, 1/n_{c_1}^1, 1/n_1^2, \ldots, 1/n_{c_2}^2, \ldots, 1/n_1^m, \ldots, 1/n_{c_m}^m)$. La matrice des poids est donnée par

$$W = \frac{1}{nm} D^{-1}$$

et la métrique est donnée par

$$M = nmD \ .$$

Il faut noter que, par symétrie de $\mathcal{B}$, il n'est pas nécessaire de considérer les profils colonnes. De plus, avec l'AFCM, seules certaines valeurs propres sont considérées. Il n'est donc pas possible de mesurer la qualité globale de la représentation à partir des valeurs propres. Il est néanmoins toujours possible d'utiliser les contributions des modalités à l'inertie des axes pour les interpréter.

5 Cancer du sein (AFCM)

Afin d'illustrer l'AFCM, nous considérons des données issues d'une étude médicale sur le cancer du sein dans trois centres (Tokyo, Boston et Glamorgan). Les patientes sont réparties en groupes suivant leur âge ("< 50", "$50 - 69$" et "> 70"), leur survie ("Oui" et "Non"), la taille de l'inflammation ("Petite" et "Grande") et sa nature ("Maligne" et "Bénine"). Les données brutes peuvent être récupérées avec la commande suivante,

```
data <- DataCancer()
```

Pour les plus en avance, il est possible d'essayer de calculer le tableau de Burt à partir de `data`. Plus simplement, vous pouvez récupérer ce tableau grâce à la commande qui suit,

```
B <- DataCancerBurt()
```

Calculer les matrices W et M et, en utilisant `B`, procéder à l'AFCM et à sa représentation graphique. Commenter et interpréter les résultats.