Analyse de données
M1 Statistique et économétrie - 2013
V. Monbet
Analyse factorielle des correspondances

A travers ce TD, nous allons apprendre à mettre en oeuvre l'analyse factorielle des correspondances. Le TD commence par une application simple sous SAS. La seconde application est traitée sous R. Il s'agit d'analyser les votes au premier tour des élections présidentielles. Cet exemple permettra d'aider à bien comprendre l'AFC notamment en la comparant à l'ACP.

1 Questions de cours

1. A quel type de questions permet de répondre l'analyse factorielle des correspondances ? Donner un exemple.
2. Peut-on l'utiliser pour analyser des données qui sont des réalisations de variables continues ?
3. Combien de variables peut-on considérer ? Quel est la dimension maximum de l'espace factoriel ?
4. Rappeler la définition du tableau de contingence ainsi que celle des profils moyens.
5. Comment mesure t'on la contribution d'une modalité à la formation des axes factoriels ?

2 CSP et vacances

Les données sont sous une forme particulière avec trois colonnes. La première contient un effectif conjoint associé aux deux modalités contenues dans les colonnes suivantes. La deuxième colonne contient un code de catégorie socio-professionnelle tandis que la 3ème contient un code correspondant à un type d'hébergement pour les vacances.

2.1 Lecture et description des variables

La lecture des données se fait par le programme :

```
vacs = read.table("vaccsp_names.dat",header=T)
```

1. Identifier les différentes modalités et représenter leurs distributions (vous pourrez utiliser la fonction `table`.
2. Tracer également les profils lignes ou colonnes et commentez ces profils.
3. Calculer le test du χ^2 d'indépendance entre les variables. Que dire de la liaison entre celles-ci ?

```
CSP = unique(vacs$CSP)
heberg = unique(vacs$heberg)
vacs.tab = matrix(0,length(CSP),length(heberg))
cnt = 0
for (i in 1:length(CSP)) {
for (j in 1:length(heberg)) {
cnt = cnt+1
vacs.tab[i,j] = vacs$effectif[cnt]
}
}

rownames(vacs.tab) <- CSP
colnames(vacs.tab) <- heberg
chi2 = chisq.test(vacs.tab)
```

2.2 AFC

L'AFC est obtenue, par exemple, à l'aide des commandes suivantes :

```
library(FactoMineR)
res.ca = CA(vacs)
```

Interpréter les graphiques obtenus : positions relatives des modalités hébergement puis des modalités CSP et enfin des modalités des deux variables.

3 Parfums

On dispose d'un tableau de contingence contenant 12 parfums décrits par 39 mots. Une valeur xij correspond au nombre de fois où le descripteurs j a été utilisé pour decrire le parfum i. Nous voulons savoir quels sont les parfums qui ont le même profil de mots ? Quels sont les mots qui se ressemblent c'est à dire qui sont associés de la même façon aux mêmes parfums ?

– Importer le jeu de données "parfums.tex".

– Analyser le tableau de données à l'aide d'une AFC.

```
library(FactoMineR)
perfume = read.table("perfume.txt",header=T,sep="\t",row.names=1)
res.ca = CA(perfume,col.sup=14:39)

plot(res.ca,invisible="row")
plot(res.ca,invisible=c("col","col.sup"))
```

– Interpréter globalement le premier plan factoriel. Quelles sont les 2 variables qui contribuent le plus ce plan ? Peut-on retrouver ce résultat sur le graphique ? Si vous regardez les données brutes, ceci vous parait-il logique ?
– Interpréter la proximité entre J'adore eau de parfum et J'adore eau de toilette.
– Comment caractériser le parfum Lolita Lempika ? Quel est l ?adjectif qui lui correspond le mieux ? Interpréter finement les positions des modalités "Lolita Lempika" de la variable "parfums" et "sugary" et "vanilla" de la variable "descripteurs".

4 Premier tour des présidentielles

Pour le premier tour des élections présidentielles de 2007, on connaît, pour chacun des 95 départements métropolitains et la Corse 1, dans l ?ordre, le nombre de voix des candidats.

4.1 Analyse descriptive

Importer les données dans R en exécutant les instructions suivantes.

```
presid <- read.csv("~\Presidentielle.CSV", row.names = 1)
```

1. Quelles sont les différentes variables reproduites dans le tableau ? Quelle est leur nature ? Quelles sont les modalités sur lesquelles on va faire porter l'analyse factorielle des correspondances ?
2. Obtenir les statistiques descriptives des deux variables observées, en utilisant la fonction `summary`.
3. Montrer, en utilisant un test du χ^2, que les deux variables sont liées (fonction : `chisq.test`)

4.2 AFC

Nous allons utiliser la routine d'AFC du package `FactoMineR`.

1. Réaliser l'AFC du tableau de contingence des votes au premier tour des élections présidentielles.

```
library(FactoMineR)
res.ca <- CA(presid, graph = FALSE)
```

2. Représenter les valeurs propres en utilisant des diagrammes en baton. Par combien d'axes l'information est-elle représentée de manière satisfaisante ?

```
par(mfrow=c(1,2))
barplot(res.ca$eig$per, ylab = "Inertie expliquée (en pourc.)",
      xlab = "Composante")
barplot(res.ca$eig$cum, ylab = "Inertie cumulée expliquée (en pourc.)",
      xlab = "Composante")
abline(h = 80, lty = 2, lwd = 2)
```

3. Analyse des profils moyens. Commenter les profils ligne et colonne. Répondez par exemple aux questions suivantes : Quelles sont les régions qui comptent significativement plus (resp. moins) d'électeurs que d'autres ? Quels sont les candidats qui ont obtenu significativement plus (resp. moins) de voix que les autres ?

```
res.ca$call$marge.row
res.ca$call$marge.col
```

4. Représentation des modalités. Représenter la projection des modalités sur le premier plan factoriel puis sur le plan formé par les facteurs 3 et 4., et discuter les graphiques obtenus.

```
plot(res.ca, cex = 0.8)
plot(res.ca, axes = c(3, 4), cex = 0.8)
```

   Commenter les contributions des différentes modalités aux premiers axes factoriels :

```
res.ca$row$contrib
res.ca$col$contrib
```

5. Modalités supplémentaires. On pourrait choisir de ne pas entrer certaines des modalités dans l'inférence des axes. On pense en particulier à l'outremer qui est très éloigné du centre de gravité de l'analyse ainsi qu'à deux "petits" candidats (Nihou et Schivari). On procède la façon suivante :

```
res1.ca = CA(presid,row.sup = 23, col.sup = c(11,12),
      presid, graph = FALSE)
dev.new()
par(mfrow=c(1,2))
plot(res1.ca, cex = 0.8)
plot(res1.ca, axes = c(3, 4), cex = 0.8)
```

6. Utiliser les graphiques et résultats précédents pour répondre aux questions :
   (a) Y a-t-il des régions qui se ressemblent, c'est-à-dire dans lesquels les résultats (en pourcentages) des différents candidats sont voisins ? Y a-t-il au contraire des régions qui s'opposent (résultats très différents) ?

(b) Y a-t-il des régions dont les résultats sont proches des résultats nationaux ? Y a-t-il des régions "à part" (dont les résultats s'écartent notablement des résultats nationaux) ?

(c) Y a-t-il des candidats dont les résultats se ressemblent : ils n'obtiennent pas nécessairement les mêmes scores, mais les régions où ils obtiennent de bons scores sont les mêmes ? Y a-t-il des candidats dont les résultats s'opposent ?

(d) Y a-t-il des candidats pour lesquels la répartition des votes est la même dans toutes les régions ? Y a-t-il des candidats pour lesquels le vote est concentré dans certaines régions ?

(e) Comment les régions "à part" et les candidats à "vote inégalement réparti" s'associent-ils ?

4.3 Comparaison avec l'analyse en composantes principales

Il peut être intéressant de comparer les résultats de l'AFC avec ceux de l'ACP afin de mettre en évidence le rôle joué par la distance du χ^2.

1. Quelles données proposez-vous d'utiliser pour l'ACP ? Pourquoi ?
2. Réaliser l'ACP en considérant les candidats comme individus et en utilisant des données normalisées par le nombre total de vote par candidat.
3. Commenter les graphiques en comparaison avec ceux de l'AFC.