TP 5 : Analyse Factorielle des Correspondances Multiples (AFCM)

AFCM sur des données bancaires

Présentation des données

Le jeu de données contient 68 clients d'un organisme de crédit ayant souscrit un crédit à la consommation. Les 11 variables qualitatives et les modalités associées à cet exemple sont les suivantes:

1. Variables bancaires

   - `Marché` : rénovation d'un bien (appartement, maison), voiture, scooter, moto, mobilier-ameublement, île. Cette variable indique le bien pour lequel les clients ont réalisé un emprunt.
   - `Apport` : oui, non. Cette variable indique si les clients possèdent un apport personnel avant de réaliser l'emprunt. Un apport personnel représente une garantie pour l'organisme de crédit.
   - `Impayé` : 0, 1 ou 2, 3 et plus. Cette variable indique le nombre d'échéances impayées par le client, soit le nombre de fois où il n'a pas réussi à rembourser son emprunt.
   - `Taux d'endettement` : 1 (faible), 2, 3, 4 (fort). Cette variable indique le niveau d'endettement du client. Le taux d'endettement est calculé comme le rapport entre les charges (ensemble des dépenses) et le revenu. Ce taux a été discrétisé en 4 classes.
   - `Assurance` : sans assurance, AID (assurance invalidité et décès), AID + Chômage, Senior (pour les plus de 60 ans). Cette variable indique le type d'assurance à laquelle le client a souscrit.

2. Variables sociologiques

   - `Famille` : union libre (concubinage), marié, veuf, célibataire, divorcé, pacsé.
   - `Enfants à charge` : 0, 1, 2, 3, 4 et plus, 8.
   - `Logement` : propriétaire, accédant à la propriété (personne qui n'a pas encore fini de rembourser son emprunt immobilier), locataire, logé par la famille, logé par l'employeur.
   - `Profession` : ouvrier non qualifié, ouvrier qualifié, retraité, cadre moyen, cadre supérieur, ministre.
   - `Intitulé` : M, Mme, Melle.
   - `Age` : 20 (18 à 29 ans), 30 (30 à 39), 40 (40 à 49), 50 (50 à 59), 60 et plus. Le but de cette étude est de caractériser la clientèle de l'organisme de crédit. Nous voulons dans un premier temps mettre en évidence différents profils de comportements bancaires, c'est-à-dire effectuer une typologie des individus. Nous voulons ensuite étudier la liaison entre la signalétique et les principaux facteurs de variabilité des profils de comportements bancaires (i.e. caractériser les clients aux comportements particuliers).

Analyse

1. Importer le fichier `credit.csv`.

```
library(FactoMineR)
credit0 <- read.csv(file.choose(),header=TRUE,sep=";",row.names=1)
```

2. Avant de commencer l'analyse, on peut remarquer qu'il y a une modalité `mqt` correspondant à une valeur manquante. Cette modalité correspond à l'individu 68. Nous décidons donc de supprimer cet individu, qui correspond à une erreur dans le fichier. On en profite pour réordonner les colonnes du fichier pour avoir les variables dans le même ordre que lors de la présentation des données ci-dessus.

```
credit=credit0[-68,c(1,2,3,10,11,4,5,6,7,8,9)]
summary(credit)
```

   Attention la modalité manquante `mqt` correspond toujours à un niveau pour chacune des variables, il faut alors la supprimer:

```
for (i in 1:ncol(credit)) credit[,i] <- factor(as.character(credit[,i]))
attach(credit)
```

3. Que faut-il regarder avant de commencer une AFCM?

4. Réaliser l'analyse permettant de répondre à la problématique.

```
help(MCA)
# res.MCA=MCA(credit,quali.sup=6:11)
res.MCA=MCA(credit[-67,],quali.sup=6:11,graph=FALSE)
```

5. Commenter les résultats concernant les valeurs propres, le nombre d'axes retenus...

```
res.MCA$eig
barplot(res.MCA$eig[,1])
```

6. Commenter globalement l'analyse, quelles sont les grandes tendances qui se dégagent? Effectuer des représentations séparées des individus et des variables.

```
help(plot.MCA)
plot.MCA(res.MCA, choix="ind", col.ind="black")
plot.MCA(res.MCA, choix="ind", invisible=c("quali.sup","ind.sup"),
                                                        col.ind="black")
plot.MCA(res.MCA, choix="ind", invisible=c("var","quali.sup","ind.sup"),
                                                        col.ind="black")


plot.MCA(res.MCA, choix="var", col.var="black")
```

7. Etude des variables: quelles sont les variables les plus liées à l'axe 1? à l'axe 2?

```
nbvar <- 5

variable <- NULL
for (i in 1:nbvar) {variable=c(variable,rep(names(credit)[i],
                          length(unique(factor(as.character(credit[-67,i]))))))}
```

```
ctr <- res.MCA$var$contrib

# Somme des contributions par variable sur l'axe 1 #
tapply(ctr[,1],variable,sum)
# Rapport de corrélation par rapport à l'axe 1 #
tapply(ctr[,1],variable,sum)*nbvar*res.MCA$eig[1,1]
```

8. Etude des modalités: quelles sont les modalités qui contribuent le plus à la création du premier axe? du deuxième? Ces modalités sont-elles situées forcément aux extrémités du graphique? Commenter la qualité de représentation de ces modalités: les résultats obtenus vous semblent-ils surprenants?

```
# Contributions au axes #
round(res.MCA$var$contrib[rev(order(res.MCA$var$contrib[,1])),1],2)
round(res.MCA$var$contrib[rev(order(res.MCA$var$contrib[,2])),2],2)

# res.MCA$var #
```

9. Interpréter la proximité entre la modalité `Senior` de la variable `assurance` et `Rénovation` de la variable `marché`. Revenir aux données brutes pour confirmer votre interprétation.

```
plot(assurance,marché)
conting <- table(assurance,marché)
```

10. Commenter le tableau des contributions aux Chi2.

```
Test <- chisq.test(conting,correct=FALSE)
conting
Test$expected # Expected Counts
Test
round(Test$residuals^2, 2)
```

11. Interpréter la proximité entre les modalités `Impayé 3 et plus` et `AID+chomage`.
12. Interpréter la position des modalités de la variable `Apport`.
13. Variables supplémentaires:
    - On ne dispose pas de la contribution, est-ce normal?
    - A quoi correspond la valeur `test` ?

      ```
      res.MCA$quali.sup
      ```

    - Représenter les individus sur le premier plan factoriel et colorier les individus par âge.

      ```
      age <- credit[,11]
      cat <- levels(age)
      col <- c("blue","green","yellow","red","black")
      colAge <- rep("blue",length(age))
      for (i in 2:length(cat)) {
      id <- credit[,11]==cat[i]
      colAge[id] <- col[i]
      }

      plot.MCA(res.MCA, choix="ind", invisible=c("var","quali.sup","ind.sup"),
                                                          col.ind=colAge)
      ```

- Interpréter la position de la modalité `logé par la famille`.

14. Remarque: il est intéressant de revenir aux données brutes pour analyser encore plus finement la proximité entre deux modalités qui vous intéressent particulièrement.

15. Décrire de manière automatique les axes:

    ```
    dimdesc(res.MCA)
    ```

16. Réaliser la classification sur les composantes principales de l'AFCM :

    ```
    HCPC(res.MCA)
    ```