Devoir Maison d'Analyse des données

Charles-Meldhine Madi Mnemoi

I - Hobbies

1) Introduction

Le dataset hobbies est issue d'une enquête de l'INSEE de 2003 intitulée "Histoire de vie". 8403 individus ont répondu à 18 questions concernant leurs hobbies. Les caractéristiques des individus ont également été enregistrées (sexe, classe d'âge, statut marital, profession). Le nombre total d'hobbies de chaque personne est disponible. Le dataset est un tableau disjonctif complet : chaque variable (à part le nombre total de hobbies) est un booléen (1 si la personne pratique le hobbie/est dans la catégorie, 0 sinon).

Dans ce rapport, nous utiliserons des techniques d'analyses multidimensionnelles afin d'effectuer une analyse exploratoire du dataset. Notre but est d'identifier des profils d'individus à travers leurs loisirs.

2) Analyses

Chargeons les libraries nécessaires :

```
library(FactoMineR)
```

Ayant affaire à plusieurs variables qualitatives, on procède à une Analyse factorielle des correspondances multiples (AFCM). L'AFCM cherche le plan qui maximise la variance multidimensionnelle du tableau disjonctif complet. On peut espérer que cela nous permette de distinguer des profils d'individus.

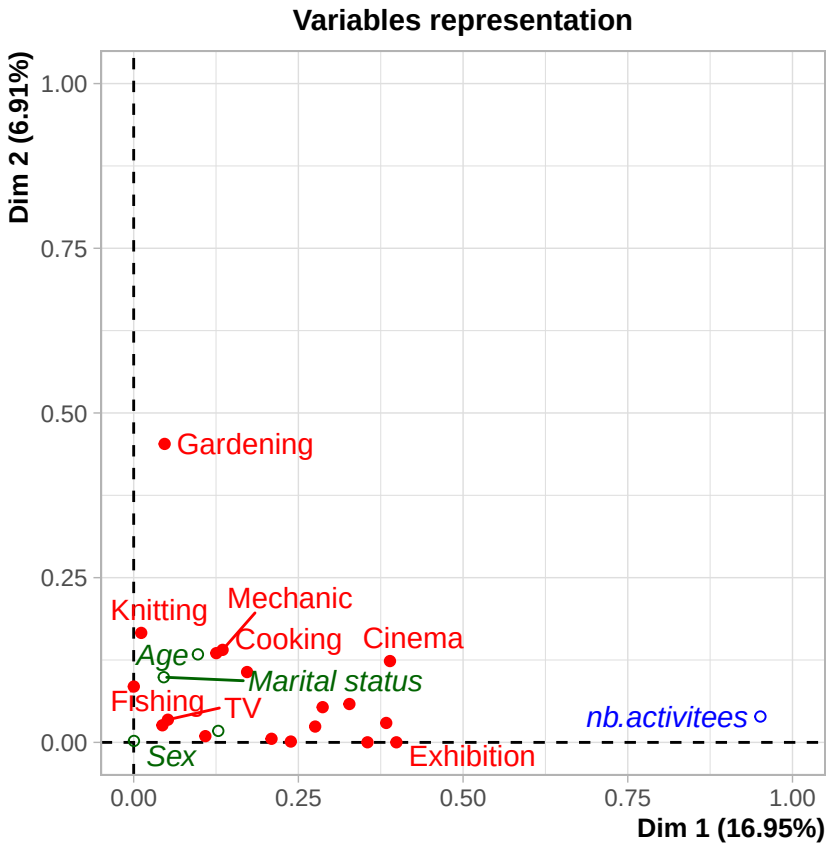
On ne projette que les variables "hobbies". Les variables donannt des informations sur les individus ne sont affichés dans l'hyperplan qu'à but d'interprétation (elles ne contribuent pas aux axes de l'hyperplan). En effet, on ne cherche qu'à étudier des profils avec des hobbies différents. Utiliser d'autres informations pourrait perturber l'analyse.

```
data(hobbies)
res.hobbies.MCA = MCA(hobbies,quali.sup = 19:22, quanti.sup = 23, graph=FALSE)
```

La représentation des individus dans le premier plan factoriel étant homogène (aucun groupe d'individu ne se distingue), nous ne la représenterons pas.

On peut étudier la représentation des variables pour aider à l'interprétation :

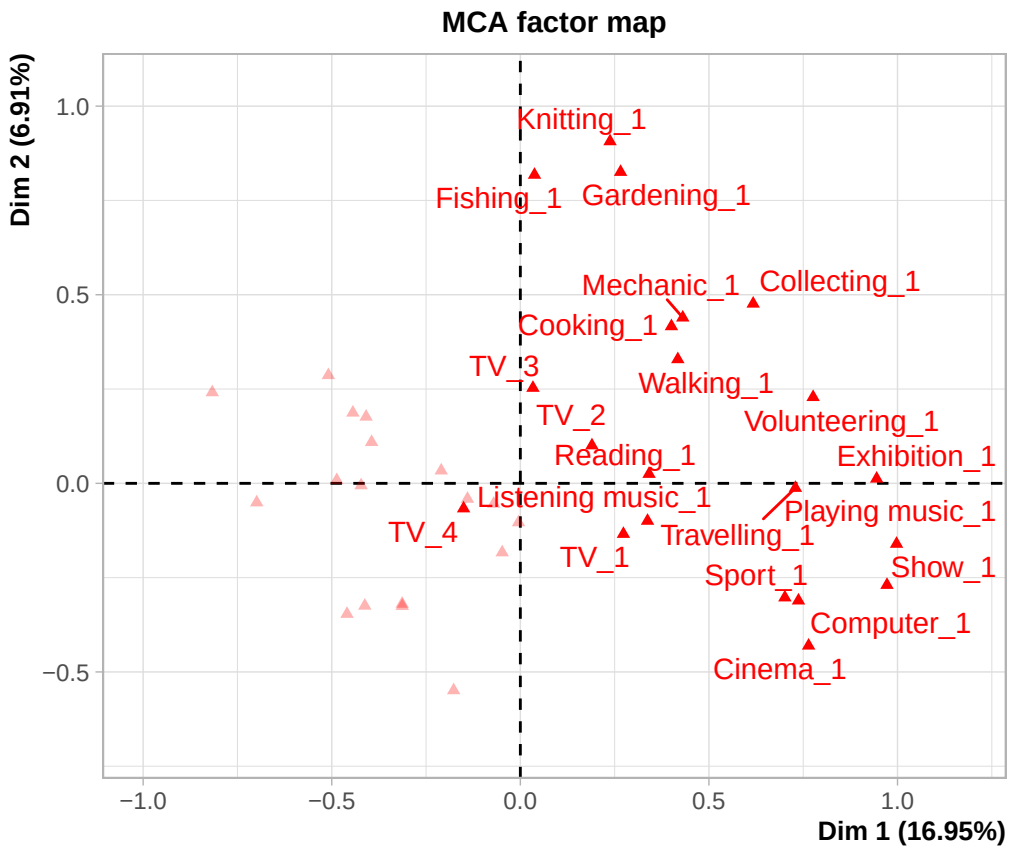
```
plot.MCA(res.hobbies.MCA,choix="var")
```



On observe que les variables qui contribuent le plus au premier axe sont Exhibition et Cinema, alors que la variable qui contribue le plus au deuxième axe est Gardening. On remarque que les variables supplémentaires semblent peu liés aux axes factoriels, spécialement Sex.

On peut également afficher les hobbies pratiqués dans le plan factoriel :

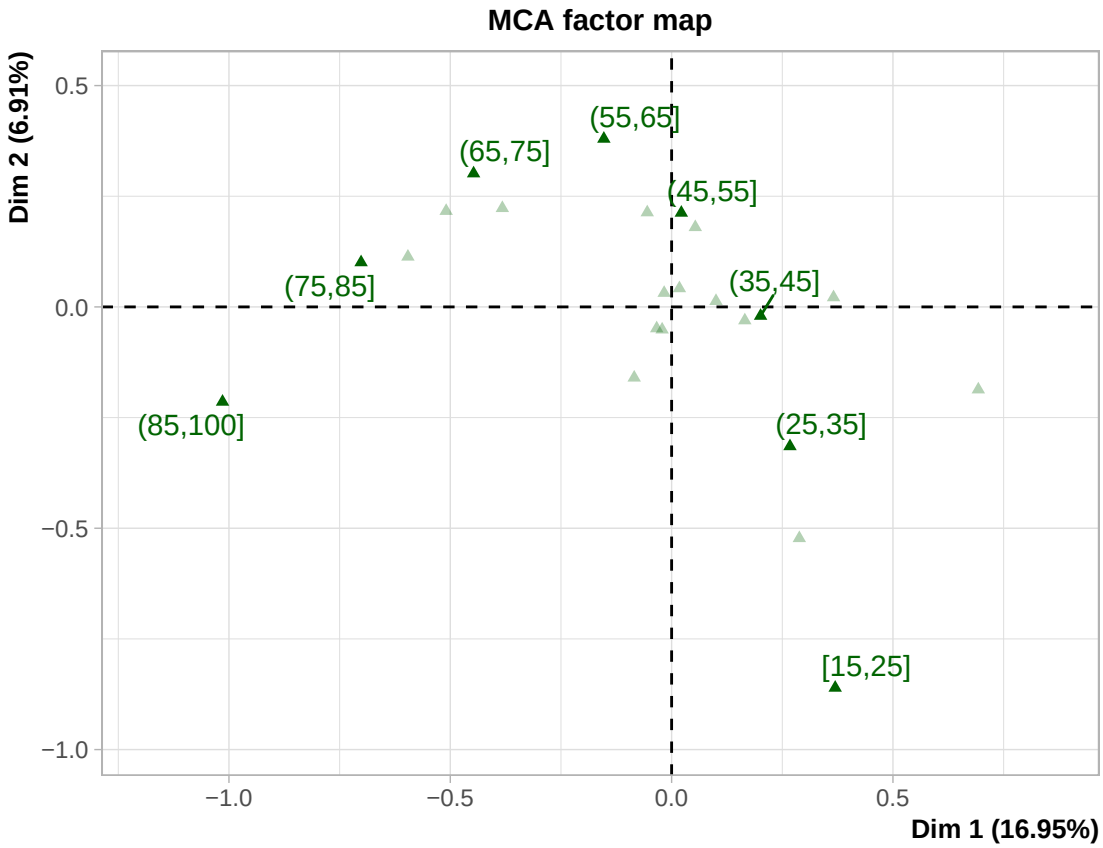
```
activities = c(seq(2,36,by=2),37:39)
plot.MCA(res.hobbies.MCA,choix = "ind",invisible = c("ind","quali.sup"),selectMod = activities)
```



on observe alors que les individus pratiquant le plus d'activités ont une grande coordonnée positive pour le premier axe factoriel (ce qui est cohérent avec la représentation de la variable quantitative supplémentaire nb.activites). Le second axe oppose quant à lui des activités "calmes typées 3ème âge" (Knitting, Fishing, Gardening) en haut et "dynamiques typés jeunes" (Cinema, Sport, Show) en bas.

On peut représenter les modalités de la variable Age pour confirmer cette hypothèse :

```
age = 3:10
plot.MCA(res.hobbies.MCA,choix = "ind",invisible = c("ind","var"),selectMod = age)
```



On observe en effet que le deuxième axe factoriel discrimine les jeunes et les personnes âgés.

II - German credit

1) Introduction

Le dataset German Credit est un dataset issu de l'Université de Hambourg. Il contient les informations concernant 1000 clients comme leur sexe, leur âge et leurs capacités à rembourser un crédit.

Dans ce rapport, nous utiliserons des techniques d'analyses multidimensionnelles afin d'effectuer une analyse exploratoire du dataset. Notre but est d'identifier des profils d'individus typiques et leur comportement face au crédit.

Pour cela nous étudierons les variables quantitatives du dataset :

- "age" : l'âge du client (en années)
- "amount" : la quantité empruntée par le client
- "duration: la durée du crédit (en mois)
- "install_rate" : la proportion du crédit souscrit en fonction des revenus du client (en %)
- "num_credits" : le nombre de crédits en cours souscrits par le client
- "num_dependants" : le nombre de personnes à charge du client

On commence par charger le dataset German Credit :

```
load("C:/Users/madic/Desktop/Charles/Ecole/Analyse de données/germanCredit.RData")
```

Puis on extrait les variables quantitatives :

```
sub_quanti = germanCredit.data[,which(names(germanCredit.data) %in% liste_var_quanti)]
sub_quanti.scale = data.frame(scale(sub_quanti, scale=TRUE, center=TRUE)) #données centrées et réduites
```

2) Analyses

Chargeons les libraries nécessaires :

```
library(FactoMineR)
library(plotly)
```

Ayant affaire à des variables quantitatives, commençons par effectuer une Analyse par Composantes Principales (ACP) du jeu de données. L'ACP est une analyse factorielle cherchant à trouver l'hyperplan qui maximise la variance multidimensionnelle de nos données. On peut donc espérer qu'elle permette de bien distinguer les points dans l'hyperplan trouvé.

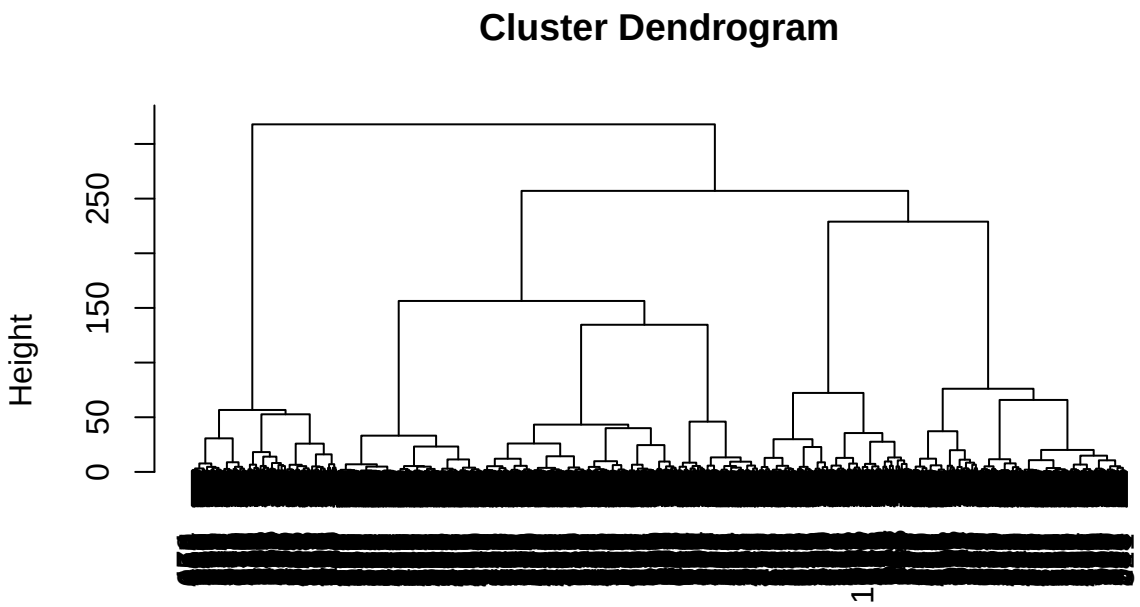
```
res.credit.pca=PCA(sub_quanti, scale=TRUE, graph=FALSE)
plot.PCA(res.credit.pca,choix="ind")
```

Malheureusement l'ACP ne donne de bons résultats (on n'arrive pas suffisamment à distinguer les profils d'individus).

Essayons donc la classification ascendante hiérarchique (CAH). Cette technique consiste à assigner chaque individu dans la classe composée uniquement d'eux mêmes puis de les regrouper en tentant de maximiser un certain critère. Dans notre cas on choisira de maximiser la distance de Ward, ce qui revient à maximiser l'inertie inter-classes, c'est à dire la distance entre les différents profils d'individus.

On commence par étudier le dendrogramme de la CAH :

```
res.credit.hclust=hclust(dist(sub_quanti.scale), method="ward.D")
plot(res.credit.hclust, sub="", xlab="")
```



En coupant l'arbre sur les branches longues, on observe ainsi trois grands clusters se dessiner. C'est ce nombre que l'on retiendra par la suite.

Pour encore mieux observer les profils d'individus, on peut appliquer la CAH à l'hyperplan factoriel trouvé par l'ACP produite plus haut :

```
res.credit.pca=PCA(sub_quanti, scale=TRUE, graph=FALSE)
res.credit.hcpc=HCPC(res.credit.pca, graph=FALSE, nb.clust=3)
```

On peut récupérer les coordonnées des individus dans l'hyperplan factoriel à l'aide de la variable `res.credit.pca$ind$coord` ainsi que le cluster auquels ils appartiennent dans `res.credit.hcpc$data.clust$clust`.

On peut ainsi représenter les individus dans certains axes de l'hyperplan et les colorer selon leurs clusters :

```
res.credit.pca=PCA(sub_quanti, scale=TRUE, graph=FALSE)
res.credit.hcpc=HCPC(res.credit.pca, graph=FALSE, nb.clust=3)
individus_dans_l_hyperplan_factoriel = res.credit.pca$ind$coord
individus_dans_l_hyperplan_factoriel = data.frame(individus_dans_l_hyperplan_factoriel)
clusters = res.credit.hcpc$data.clust$clust


fig = plot_ly(individus_dans_l_hyperplan_factoriel,x=~Dim.1,y=~Dim.2,color=clusters, mode="markers", ty
layout(title = "Représentation des individus dans le premier plan factoriel",legend=list(title=list(tex

fig
```

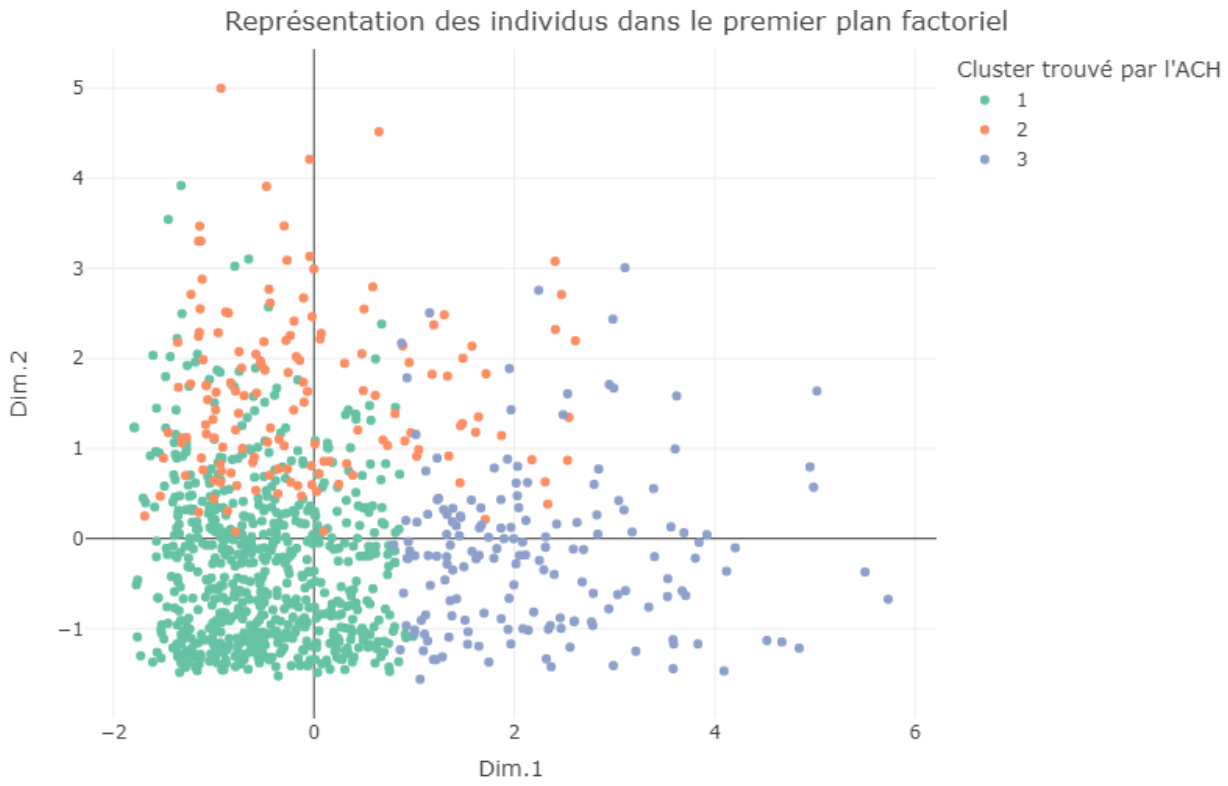


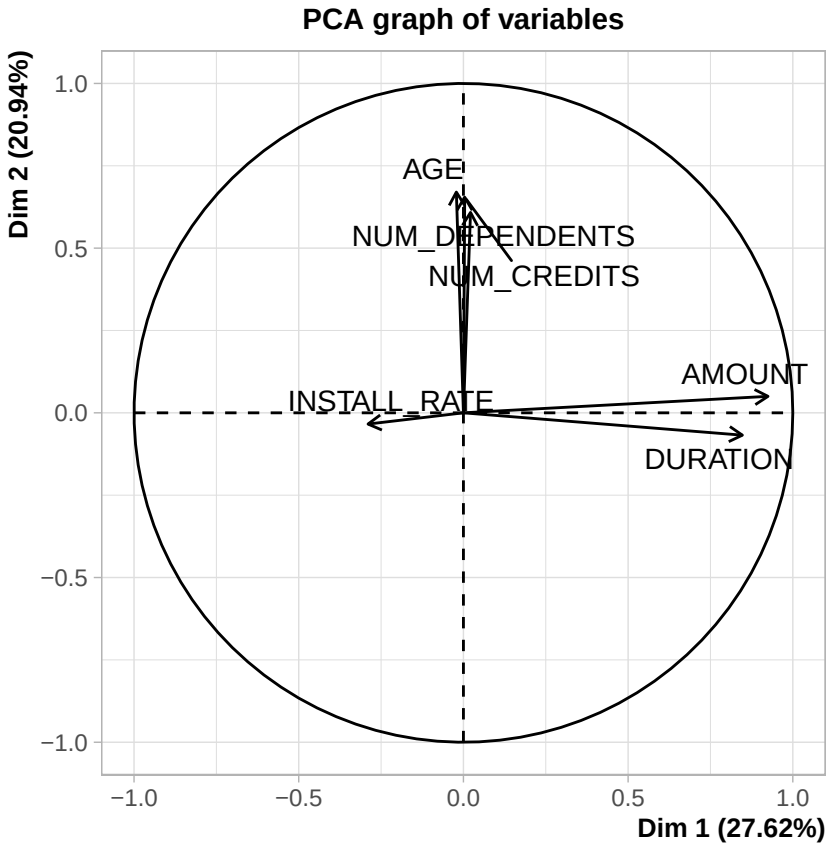
Figure 1: Représentation des individus dans le premier plan factoriel

On observe maintenant trois classes bien distinctes. On peut procéder aux interprétations.

3) Interprétation

Commençons par dessiner le cercle de corrélations de notre ACP qui nous aidera à interpréter la signification des axes factoriels :

```
plot.PCA(res.credit.pca,choix = "var")
```



On observe que les individus avec une grande valeur positive dans la première dimension du plan sont âgés, ont plus de personnes à charge et ont plus de crédits et que les individus avec une grande valeur positive dans la deuxième dimension du plan prennent des crédits d'un montant plus important et de plus longue durée. INSTALL_RATE ayant un faible module dans le cercle de corrélations, elle est mal representée sur ce plan, on ne l'interprétera donc pas.

En revenant à la figure 1, on constate alors que la classe 1 est composée d'individus relativement jeunes qui prend des crédits de faibles montants pour peu longtemps. La classe 2 est composée d'individus âgés voir très âgés qui prennent de petits crédits, étant déjà très endettés. La classe 3 est composé de personnes relativement jeunes qui prennent des crédits d'un montant important pour une longue durée.