Université de Carthage
Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Examen d'Analyse des Données

1 ère année du cycle de formation d'ingénieurs

Durée de l'épreuve : 1 heure 30 - Documents non autorisés
Nombre de pages : 5 - Date de l'épreuve : 19 mai 2019

Exercice 1 : Les données `poison` proviennent d'une enquête menée auprès d'enfants de l'école primaire qui ont subi des intoxications alimentaires. Ils ont été interrogés sur leurs symptômes et sur ce qu'ils ont mangé. Les données contiennent 55 lignes (individus) et 11 colonnes (variables).
Nous avons effectué une Analyse des Correspondances Multiples sur ces données. Les résultats de cette ACM ainsi que les statistiques descriptives élémentaires sur ces données sont présentés à l'Annexe.

1. Quelle relation existe-t-il entre l'ACM et l'Analyse Factorielle des Correspondances (AFC) ?

2. Calculer le taux d'inertie cumulé des 2 premiers axes de cette ACM.

3. Discuter le nombre d'axes à retenir en vous basant sur 3 critères différents.

4. Donner une interprétation de la première carte des modalités. Indiquer la commande R qui permetterait de faciliter l'interprétation de cette carte ?

5. Donner une interprétation de la première carte des individus. On commentera uniquement les individus dont le numéro apparaît sur la carte.

6. Quelle démarche proposeriez vous pour une interprétation plus précise et plus détaillée de la carte des individus ?

Exercice 2 : On considère la base de données `donnees_pays` datant de 1991. Pour chacun des 10 pays suivants : Af. du Sud, Algérie, Allemagne, Arabie S., Egypte, Ethiopie, Finlande, France, Koweit et Tunisie on dispose des valeurs des 6 variables suivantes : le PNB/h. mesuré en $ US, les taux d'inflation (T. Inflat.) et de chômage (T. Chom.) en pourcentage, "Com." désigne la balance des échanges commerciaux mesurée en Milliards de $ US, "Popu." désigne la population mesurée en millions d'habitants et "Sup." désigne la superficie mesurée en millions de km^2. On a effectué une ACP sur les 6 variables décrites ci-dessus en exécutant le script suivant :

```
> library(FactoMineR)
> summary(donnees_pays)
      PNB            T.INFLAT          T.CHOM           COM               POPU
 Min.   :  110   Min.   : 3.20   Min.   : 0.0   Min.   :-10.100   Min.   : 1.30
 1st Qu.: 1398   1st Qu.: 3.65   1st Qu.: 0.0   1st Qu.: -0.810   1st Qu.:10.12
 Median : 5069   Median : 6.30   Median : 6.4   Median :  3.190   Median :31.46
 Mean   : 9872   Mean   :14.69   Mean   : 7.9   Mean   :  6.077   Mean   :34.15
 3rd Qu.:19273   3rd Qu.:18.52   3rd Qu.:13.6   3rd Qu.: 16.567   3rd Qu.:55.73
 Max.   :25800   Max.   :50.00   Max.   :24.3   Max.   : 23.500   Max.   :81.00
```

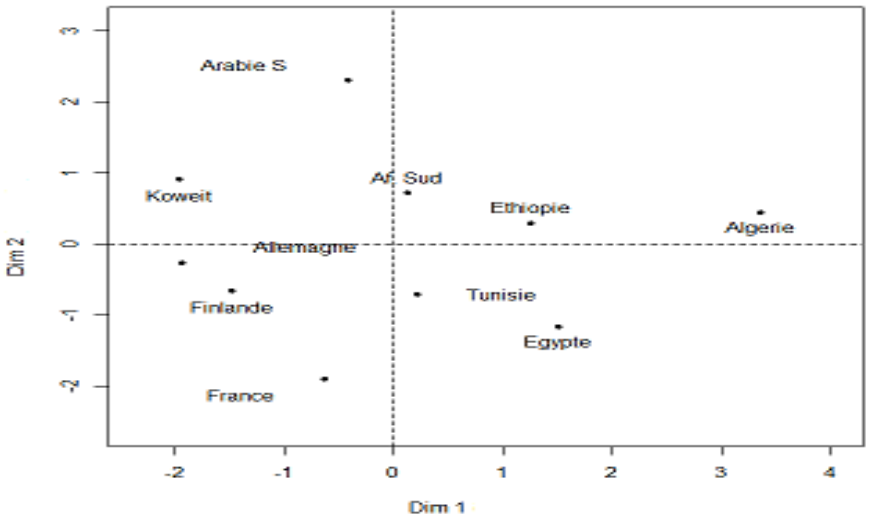

FIGURE 1 – Carte des individus (Axes1-2)

```
      SUP
 Min.   :0.020
 1st Qu.:0.335
 Median :0.775
 Mean   :0.938
 3rd Qu.:1.220
 Max.   :2.380
> res.pca <- PCA(donnees_pays, graph=TRUE)
> round(res.pca$eig[1:4,1],3)
comp 1 comp 2 comp 3 comp 4
 2.559  1.296  1.000  0.598
> round(res.pca$var$coord[,1:3],2)
       Dim.1 Dim.2 Dim.3
PNB    -0.77 -0.29  0.12
INFLAT  0.91  0.10  0.13
CHOM    0.64 -0.47 -0.26
COM    -0.48  0.74  0.19
POPU    0.09 -0.39  0.91
SUP     0.70  0.53  0.21
```

1. Combien de composantes principales devrait-on retenir ?

Dans la suite, on suppose que l'on retient 3 composantes principales.

2. Donner une interprétation des axes retenus.

3. Donner une interprétation de la première carte des individus.

On a effectué une classification automatique des 10 pays en exécutant le script suivant, la hiérarchie issue de cette classification est présentée ci-dessous :

```
> library(FactoMineR)
> library(cluster)
> classif<-agnes(donnees_pays, method="ward")
```

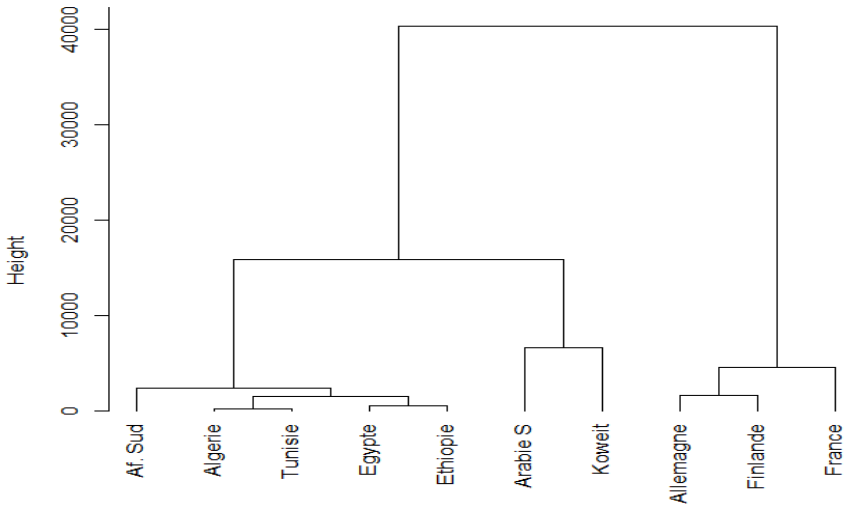

FIGURE 2 – Classification hiérarchique sur les 10 pays

```
> plot(classif,xlab="individuals",main="")
> title("Dendrogram")
> classes<-cutree(classif,k=3)
> classes
 [1] 1 1 2 3 1 1 2 2 3 1

> res_nbclust<-NbClust(donnees_pays,min.nc = 2, max.nc = 4, index="silhouette",
method = "ward.D")
> res_nbclust
$All.index
     2      3      4
0.7064 0.6753 0.7827

$Best.nc
Number_clusters     Value_Index
         4.0000          0.7827
```

4. En vous basant sur la hiérarchie déterminer la meilleure partition des 10 pays.

5. Commenter les résultats de `res_nbclust`.

6. Dans la suite on considère la partition en 3 classes. En utilisant la fonction `catdes` pour décrire les 3 classes, nous avons obtenu les résultats donnés ci-dessous. En vous basant sur ces résultats, donner une interprétation des 3 classes.

```
#Résultats de la fonction catdes

$`1`
          v.test Mean in category Overall mean sd in category Overall sd     p.value
INFLAT  2.145506            25.68        14.69       15.18254   15.36701 0.031912440
PNB    -2.604789          1286.00      9871.80      918.44651 9888.47679 0.009193073

$`2`
```

```
      v.test Mean in category Overall mean sd in category Overall sd     p.value
PNB 2.737165         23653.33       9871.8       1976.298   9888.477 0.006197117

$`3`
      v.test Mean in category Overall mean sd in category Overall sd    p.value
COM 1.992221            21.05        6.077           1.05    11.2736 0.04634683
```

Annexe

```
> library(FactoMineR)
> library(factoextra)
> summary(poison)
   Nausea      Vomiting    Abdominals     Fever          Diarrhae       Potato
 Nausea_n:43   Vomit_n:33   Abdo_n:18   Fever_n:20   Diarrhea_n:20   Potato_n: 3
 Nausea_y:12   Vomit_y:22   Abdo_y:37   Fever_y:35   Diarrhea_y:35   Potato_y:52
     Fish         Mayo       Courgette       Cheese         Icecream
 Fish_n: 1   Mayo_n:10   Courg_n: 5   Cheese_n: 7   Icecream_n: 4
 Fish_y:54   Mayo_y:45   Courg_y:50   Cheese_y:48   Icecream_y:51

> res.mca <- MCA (poison, graph = TRUE)
> round(res.mca$eig[1:5,1],3)
[1] 0.335 0.129 0.107 0.096 0.079
```

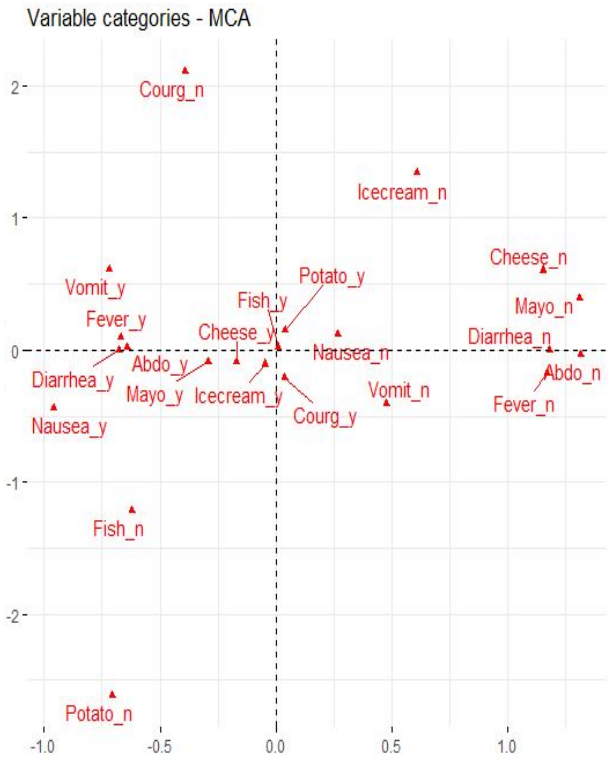

FIGURE 3 – Carte des modalités (Axes1-2)

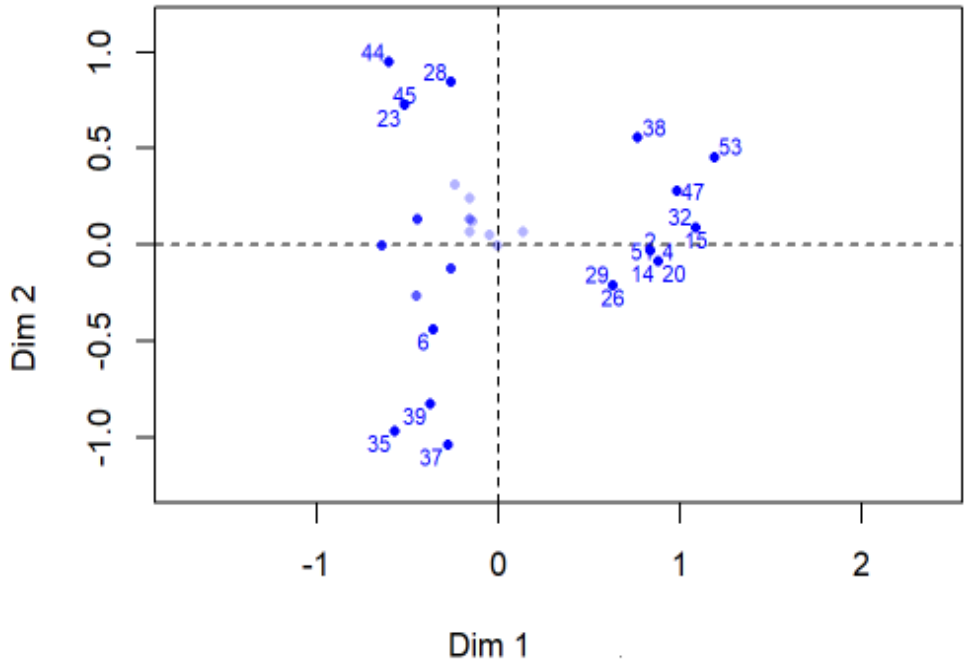

FIGURE 4 – Carte des individus (Axes1-2)