Exercices en analyse de données

Cours de M2 - F. Husson - Agrocampus

Exercice 1 : Significativité des pourcentages d'inertie en ACP

1. Construire un tableau de données avec 5 individus et 200 variables. Pour ce faire, simuler 200 variables indépendantes avec la fonction `rnorm`*. Faire l'ACP sur ce jeu de données. Comment interpréteriez-vous le graphe des variables ?*

2. On propose de prendre le jeu de données decathlon disponible dans FactoMineR, mais de ne conserver que les 7 premières lignes et 10 colonnes quantitatives.

```
library(FactoMineR)
data(decathlon)
don <- decathlon[1:7,1:10]
```

Faire l'ACP sur ce jeu de données. Le pourcentage d'inertie expliqué par le plan est-il suffisant pour interpréter les résultats ?

3. Proposer une fonction qui permet de simuler `nbsimul` *jeux de données avec un nombre d'individus* `nind` *et un nombre de variable* `nvar` *puis qui calcule le pourcentage d'inertie expliqué par le plan principal de chacune des* `nbsimul` *ACP. La fonction doit retourner le quantile à 95 % des pourcentages d'inertie.*

4. Proposer une seconde fonction qui permute les valeurs de chacune des variables, mais en conservant les valeurs présentes dans le tableau de données. Vous pourrez utiliser les lignes de code suivantes :

```
permuteLigne <- function(v) {return(v[sample(1:length(v),replace=FALSE)])}
Xnew <- apply(X,2,permuteLigne)
```

Quel est l'intérêt d'une telle méthode par rapport à la précédente ? Est-elle généralisable ?

Exercice 2 : Introduction aux tableaux multiples

On considère un jeu de données où 6 jus d'orange sont décrits par 8 variables de chimie et 7 variables sensorielles.

1. Importer les données avec la ligne de commande suivante :

```
orange <- read.table("https://husson.github.io/img/orange_chimie_senso.csv",
   header=TRUE,sep=";",row.names=1)
```

2. Comment caractériseriez-vous les jus d'oranges du point de vue des variables de chimie seules ? Même question avec les variables sensorielles seules.

3. On aimerait relier ces deux analyses pour pouvoir comparer la perception sensorielle des jus d'orange et la caractérisation chimique. Proposez plusieurs façon de faire. Indiquer les différences d'objectif pour chacune de ces méthodes.

4. Comment feriez-vous si les variables sensorielles étaient qualitatives et non quantitatives ?

Exercices en analyse de données

Cours de M2 - F. Husson - Agrocampus

Exercice 1 : Analyse des cancers

Quarante-cinq patients atteints d'une tumeur au cerveau sont classés selon le type de tumeur dont ils sont atteints : oligodendrogliome (O), astrocytome (A), mixed oligo-astrocytome (OA) et glioblastome (GBM), ce dernier étant le cancer de grade le plus élevé. Chaque tumeur a été analysée à deux niveaux différents : au niveau du transcriptome (CGH) et au niveau du génome (génome). Nous ne détaillons pas ici comment sont réalisées ces deux analyses du point de vue biologique car cela est assez complexe et ce n'est pas utile pour répondre aux questions de l'exercice. Le tableau correspondant aux mesures de CGH contient 68 variables tandis que le tableau le génome en contient 356. La variable qualitative correspondant au type de tumeur est également notée. Les données sont disponibles à l'adresse suivante suivante.

1. *Importer les données avec la ligne de commande suivante :*

```
comp <- read.table("http://factominer.free.fr/more/gene.csv",sep=";",header=T,
    row.names=1, stringsAsFactors=TRUE)
```

Puis analyser ces données pour mettre en évidence les différences entre patient du point de vue génome et CGH simultanément.

2. *Quel groupe de variables est le plus multidimensionnel ? Quel groupe de variables est le plus lié à la représentation globale de l'AFM ? Analyser la représentation des groupes.*

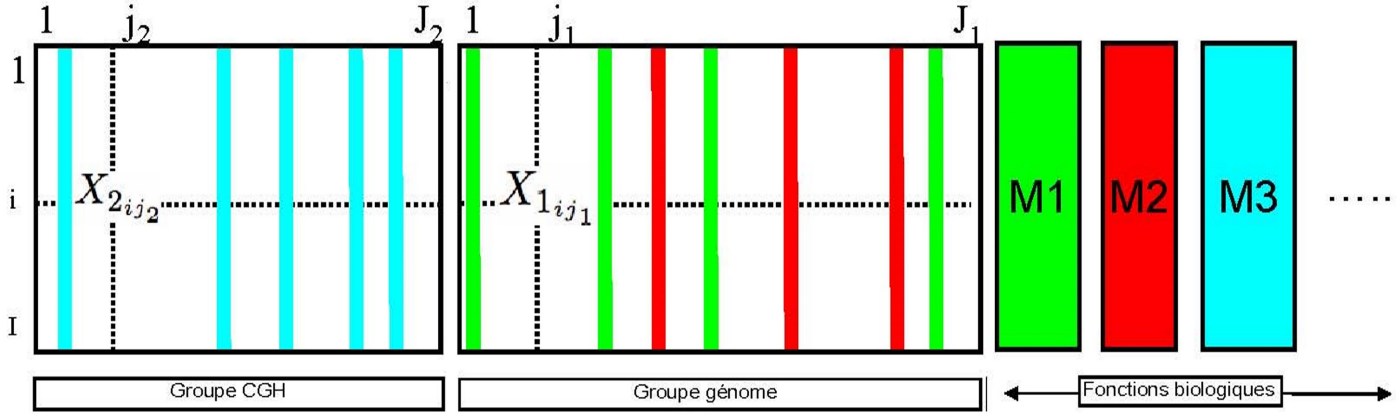
3. *Analyser la représentation des individus (prendre en compte le type de tumeur).*

4. *Construire le graphe des types de tumeur et de ses points partiels. Et analyser ce graphe.*

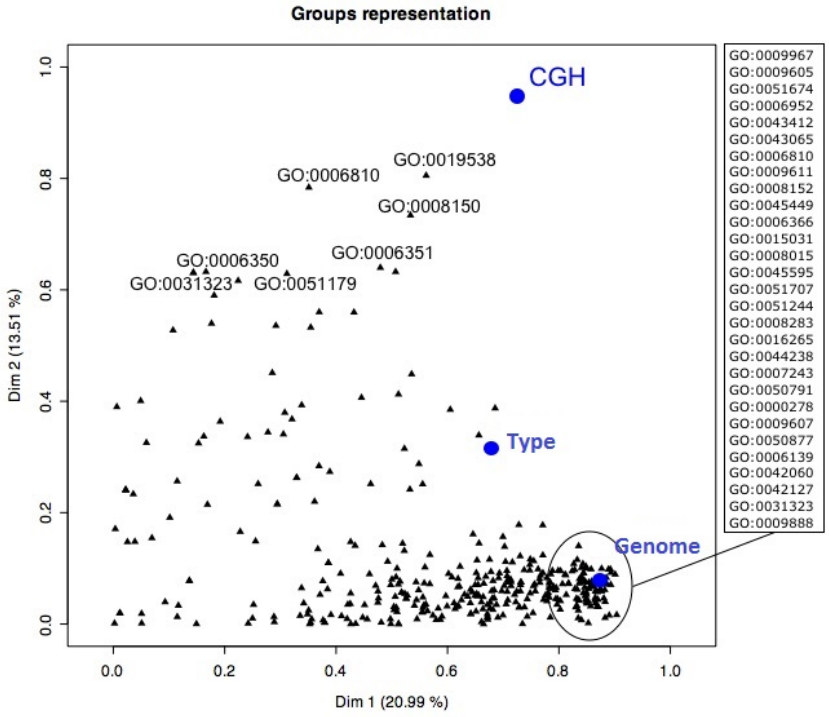
5. *Que pouvez-vous dire à partir du graphe des variables ? Puis du graphe des axes partiels ?*

6. *Interprétez.*

7. *Une information supplémentaire peut être apportée ici : elle concerne les fonctions biologiques dans lesquelles les gènes sont impliqués. Une fonction biologique peut être vue comme un groupe de gènes (i.e. un groupe de variables). Un même gène peut appartenir à différentes fonctions biologiques. On considère chaque fonction biologique comme un groupe de variables supplémentaire. On reprend donc le tableau précédent auquel on ajoute les gènes impliqués dans chaque fonction biologique (voir le tableau ci-dessous).*



Chaque fonction biologique est un groupe de variables qui sera pris comme élément supplémentaire dans l'AFM (ceci évite qu'un gène impliqué dans une fonction contribue plusieurs fois à la construction des dimensions de l'AFM).
Interpréter le graphe des fonctions biologiques :


Groups representation
CGH
GO:0019538
GO:0006810
GO:0008150
GO:0006350
GO:0006351
GO:0031323
GO:0051179
Type
Genome
GO:0009967
GO:0009605
GO:0051674
GO:0006952
GO:0043412
GO:0043065
GO:0006810
GO:0009611
GO:0008152
GO:0045449
GO:0006366
GO:0015031
GO:0008015
GO:0045595
GO:0051707
GO:0051244
GO:0008283
GO:0016265
GO:0044238
GO:0007243
GO:0050791
GO:0000278
GO:0009607
GO:0050877
GO:0006139
GO:0042060
GO:0042127
GO:0031323
GO:0009888
Dim 2 (13.51 %)
Dim 1 (20.99 %)
0.0
0.2
0.4
0.6
0.8
1.0

Exercices en analyse de données

Cours de M2 - F. Husson - Agrocampus

Exercice 1 : Analyse de données de paléoclimatologie

On s'intéresse ici à des données de paléoclimatologie, i.e. la science qui étudie les climats passés et leurs variations. Le jeu de données (voir la structure du jeu de données Fig 1) croise 700 prélevés qui mesurent le pourcentage de pollens de 31 espèces d'arbres. Ces relevés ont été effectués récemment (lors de ce siècle). A l'endroit où les prélèvements ont été effectués (latitude, longitude et altitude du lieu sont connus), nous disposons des relevés de variables climatiques : MTCO, température moyenne du mois le plus froid (mean temperature of the coldest month) ; MTWA, température moyenne du mois le plus chaud (mean temperature of the warmest month) ; GDD5, the growing degree-days (i.e. the sum of daily temperatures) above 5°C ; E_PE, the ratio of actual evapotranspiration to potential evapotranspiration ; PANN, précipitation annuelle ; TANN, température moyenne annuelle.

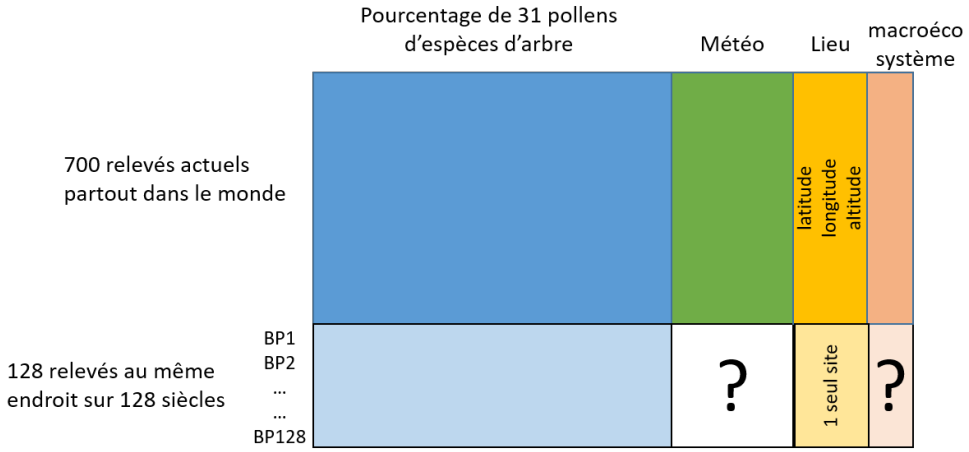


Figure 1 – Organisation du jeu de données de paléoclimatologie.

Les 700 relevés proviennent de 9 macroécosystèmes (on parle aussi de biomes) différents : COCO (cool conifer forest), COMX (cool mixed forest), COST (cool steppes), HODE (hot desert), TEDE (temperate deciduous forest), TUND (tundra), WAMX (warm mixed broad-leaved forest), WAST (warm steppes), XERO (xerophytic scrubs).
Le jeu de données comporte également les relevés d'une même carotte située au Lac de Rotsee (en Suisse, latitude 47.07 et longitude 8.3147 et à 419m d'altitude). Sur cette carotte, on peut différencier, siècle par siècle, le pourcentage de chacun des 31 pollens. Ces échantillons remontent à 128 siècles et sont notés BPxx pour Before Present xx siècles : BP15 il y a 15 siècles (ceci est approximatif, la datation avant le présent est donnée dans la colonne age). Pour ces données, on ne dispose pas du macroécosystème, ni bien entendu du climat. L'objectif est justement d'essayer de prédire le climat au cours des siècles passés à partir de la composition en les différents pollens.
Les données sont disponibles à l'adresse suivante. Vous pouvez les importer via les lignes suivantes, ainsi que visualiser la carte des températures moyennes annuelles.

```
paleo <- read.table("https://husson.github.io/img/paleo_climato.csv", header=TRUE,
                    sep=";", row.names=1)
paleo <- cbind.data.frame(paleo,present=as.factor(c(rep("Present",700),rep("Passe",128))))

library(leaflet)
pal <- colorNumeric(palette=c(low="blue",high="red"),domain=paleo[1:700,"tann"])
m <- leaflet() %>% addTiles() %>%
    addCircles(paleo[1:700,"long"],paleo[1:700,"lati"],
                     color=pal(paleo[1:700,"tann"]),fillOpacity=1,opacity=1) %>%
    addCircles(8.3147,47.07028,color="black",fillOpacity=1,opacity=1) %>%
```

```
    addPopups(8.3147,47.07028,"Lac Rotsee")
m
```

1. *En vous focalisant dans un premier temps sur les données du présent, explorer les relations entre la composition des pollens et les variables climatiques. Construire des graphes lisibles en utilisant les informations sur les macroécosystèmes, et projeter les données du passé.*

2. *Pour les données du passé, on ne dispose pas des variables climatiques. Est-ce que considérer les données du passé comme individus supplémentaires et interpréter la position de ces points a du sens dans l'analyse ? (indice : pour vous aider à répondre à la question, représenter les points partiels des individus supplémentaires uniquement).*

3. *Refaire l'analyse en sélectionnant uniquement les pollens qui étaient présents dans le passé.*

4. *Posez-vous des questions et répondez-y !*