

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 génome (génomique) et au niveau du transcriptome (CGH). 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 sur le génome contient 356 variables tandis que le tableau CGH en contient 68. 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)
```

Puis analyser ces données pour mettre en évidence les différences entre patient du point de vue génomique 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.

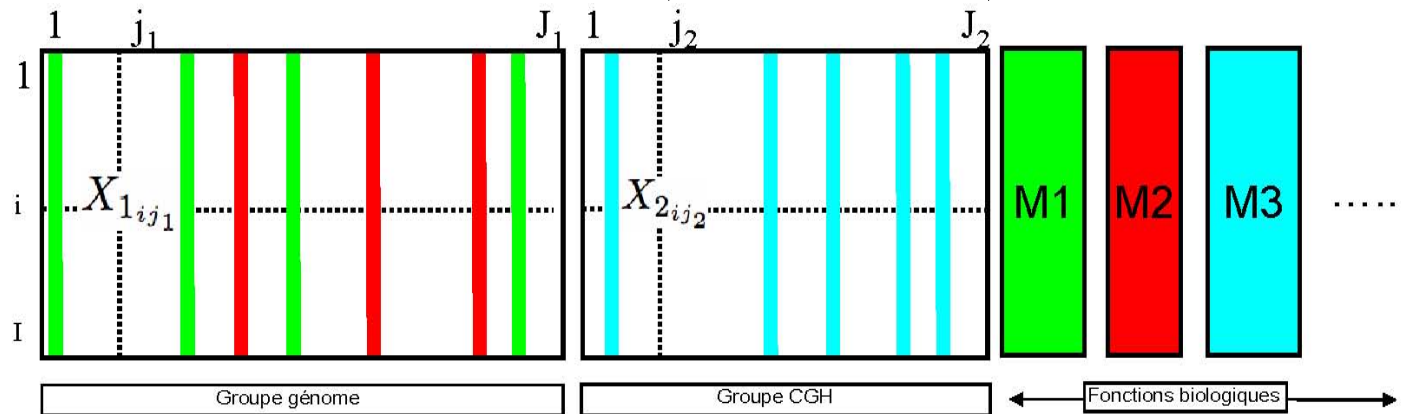
3. Analyser la représentation des individus (prendre en compte le type de tumeur).

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 :

