Master 2 BPMA
Examen 2023/2024
Durée : 1 heure
UE *Biologie Computationnelle 2*

Elie Maza

2 octobre 2023

Pour répondre aux questions des exercices ci-dessous, vous écrirez un script R. Pour les réponses nécessitant une phrase, vous écrirez celle-ci dans ce même script sous la forme d'un commentaire. A la fin de votre composition, vous enverrez votre script aux deux adresses suivantes : *Elie.Maza@toulouse-inp.fr* et *Elodie.Gaulin@univ-tlse3.fr*.

Exercice 1 (15 points)

Les données du fichier `airquality111` correspondent à des mesures quotidiennes de la qualité de l'air de New York mesurées entre mai et septembre 1973. Nous avons 111 jours de mesures et 4 variables quantitatives mesurées : la quantité d'ozone dans l'air (variable `Ozone`), la radiation solaire (variable `Solar.R`), la vitesse du vent (variable `Wind`), et la température (variable `Temp`). Nous avons également une variable supplémentaire correspondant au mois de la mesure (variable `Mois`, à valeurs dans {5,6,7,8,9} pour les mois de mai à septembre). On se propose ici d'effectuer une Analyse en Composantes Principales (ACP) de ces données avec le logiciel R, ou RStudio.

1. Importer le tableau de données "airquality111.csv" dans votre session R, ou RStudio, et effectuer un bref résumé numérique de l'ensemble des variables.
2. Effectuer les diagrammes en boîtes parallèles des variables qui seront utilisées pour notre ACP. Des individus particuliers sont-ils à noter ?
3. Effectuer les nuages de points de ces mêmes variables prises 2 à 2. Calculer la matrice des corrélations. Quelles sont les deux variables les plus corrélées ?
4. Effectuer l'ACP du tableau `airquality111` en utilisant la variable `Mois` comme variable supplémentaire.
5. Tracer l'éboulis des valeurs propres de notre ACP. Quel est le pourcentage d'inertie expliquée par le premier plan factoriel ?
6. D'après le critère de Kaiser, combien d'axes doit-on retenir pour notre analyse ?
7. Quelles variables expliquent le 1er axe factoriel de notre ACP ? Le 2ème axe factoriel ?
8. Quelle variable contribue le plus à l'axe 1 ? Quelle variable est la mieux représentée sur l'axe 1 ?
9. Quelles sont les valeurs du 3ème individu sur les deux premières composantes principales ?
10. Quel individu contribue le plus à l'axe 1 ? A l'axe 2 ?
11. Quel individu est le mieux représenté sur le premier plan factoriel ?
12. Tracer le graphe des individus sur le premier plan factoriel en utilisant une couleur différente pour chaque `Mois`. Cette variable `Mois` vous semble-t-elle être expliquée par le premier plan factoriel ? Commenter brièvement.
13. Effectuer un partitionnement des individus du tableau `airquality111` avec la méthode des k-means, en centrant et en réduisant vos données au préalable. Combien de groupes vous semblent nécessaires ?
14. Tracer les groupes obtenus à la question précédente sur le premier plan factoriel de notre ACP.
15. Comparer les groupes que vous avez obtenus avec la méthode k-means et les groupes formés a priori par la variable `Mois`. Commenter brièvement.

Exercice 2 (5 points)

Pour chacune des questions du QCM ci-dessous, une seule réponse est possible. Chaque question vaut un point. De points négatifs seront appliqués dans le cas d'une réponse fausse : -1/(Nbre de questions-1).

Question 1. Lors de la mise en œuvre d'une ACP ou d'une méthode de classification automatique sur des données quantitatives, pour éviter qu'une ou plusieurs variables ayant des variances trop importantes n'aient un poids trop important, il est usuel :

☐ d'enlever ces variables de l'analyse.
☐ de centrer les données.
☐ de centrer et réduire les données.

Question 2. De manière générale, quel est l'objectif des méthodes de classification automatique ?

☐ Maximiser la variabilité inter-classes.
☐ Maximiser la variabilité intra-classes.
☐ Maximiser à la fois la variabilité inter-classes et la variabilité intra-classes.
☐ Maximiser la somme des variabilités inter-classes et intra-classes.

Question 3. La méthode de classification automatique dites des k-means converge toujours vers la même classification quels que soient les individus choisis comme initialisation des centres des classes ?

☐ Vrai ☐ Faux

Question 4. Quelle fonction **R** du package *stats* permet de réaliser une CAH ?

☐ `scale` ☐ `dist` ☐ `hclust` ☐ `cutree`

Question 5. Le graphe ci-dessous représente le dendrogramme de la CAH effectuée sur une matrice de $n = 30$ lignes/individus et $p = 2$ colonnes/variables avec la distance euclidienne et la méthode du saut minimum (ou lien simple). D'après ce dendrogramme, l'individu 1 est plus proche de l'individu 3 que de l'individu 2 ?

☐ Vrai ☐ Faux