ESSAMADI Oussama

Solutions de devoir 1:

1- Est-il pertinent, dans le cas présent, de recourir à une Analyse en Composantes Principales (ACP) pour réduire le nombre de variables ? Justifier.

Oui, pour réduire la dimensionnalité. Cela peut faciliter l'analyse en éliminant le bruit et en mettant en évidence les structures sous-jacentes.

2- Est-il nécessaire de centrer-réduire les données avant de procéder à une ACP ? Justifier.

Oui, il est généralement recommandé de centrer et de réduire les données avant de procéder à une ACP. Cette étape, souvent appelée standardisation, est importante pour plusieurs raisons, comme la sensibilité à l'échelle et effet des unités de mesure.

3- Quelles sont les variables à exclure de l'ACP ? Pourquoi ?

Nous devrions exclure *NOM*, *ACRO*, *PAYS*, *TYPE*, et *PG* parce qu'ils ne sont pas des valeurs numériques.

4- Pour la suite du problème, la dernière variable (la masse moyenne des ions) sera considérée dans l'ACP comme " variable supplémentaire". Pouvez-vous argumenter ce choix ?

La masse ionique moyenne fournit une mesure agrégée de la composition ionique de l'eau. En incluant cette variable, nous pouvons synthétiser l'information contenue dans les différentes concentrations d'ions (*CA*, *MG*, *NA*, *K*, *SUL*, *NO3*, *HCO3*, *CL*) en une seule variable, ce qui peut simplifier l'interprétation des résultats de l'ACP.

5- Faire une ACP. Quel pourcentage de variabilité est expliqué par les deux premières composantes ?

Pourcentage de variabilité expliquée par la **1ère** composante : **53.18 %**

Pourcentage de variabilité expliquée par la **2ème** composante : **22.98 %**

Pourcentage totale variabilité expliquée par les deux première composantes est **76.15 %**

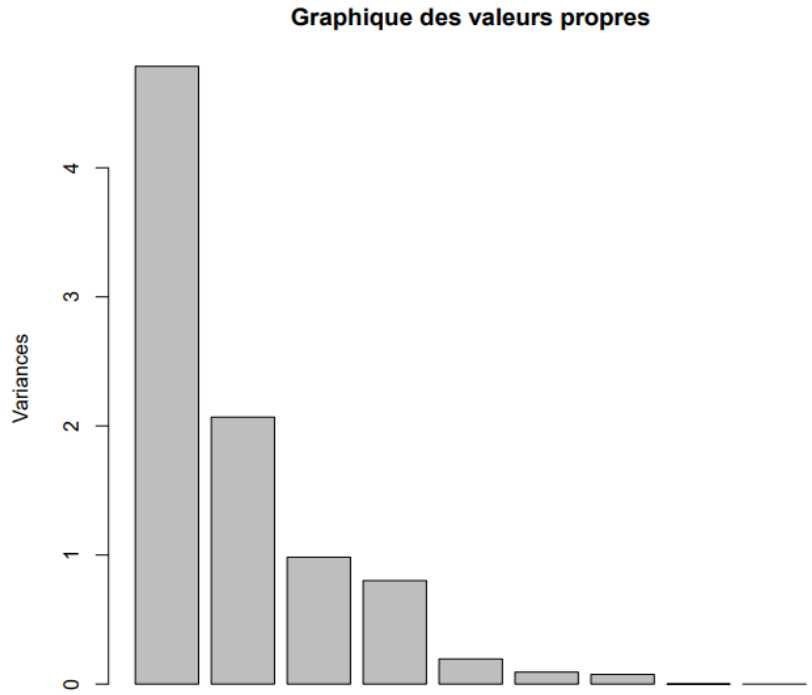
6- Combien de composantes suggériez-vous pour cette étude ? Justifiez.

Nous pouvons proposer "**trois**" éléments pour cette étude.

Selons le valeurs propres : *0.53175 0.76154 0.87073 0.95978 0.98138 0.99165 0.99996 1 1*

Avec les pourcentages de variance expliquée donnés pour chaque composante principale, nous pouvons utiliser ces informations pour décider du nombre de composantes principales à retenir dans votre analyse.

Une approche courante consiste à examiner le "scree plot" ou le graphique de la variance expliquée cumulative afin d'identifier un "coude" où l'ajout de nouvelles composantes produit des rendements décroissants en termes d'explication de la variance supplémentaire.



Dans notre cas, nous avons déjà fourni les pourcentages de variance expliquée pour chaque composante, et nous pouvons constater une nette diminution des pourcentages :

* La première composante explique une part substantielle de la variance (*53,18 %*).

* La deuxième composante contribue de manière significative (*22,98 %*).

* La troisième composante apporte encore une contribution raisonnable (*10,92 %*).

Après la troisième composante, les pourcentages commencent à diminuer rapidement, ce qui suggère des rendements décroissants en termes d'explication de la variance.

Sur la base de ces informations, nous pourrions envisager de conserver les trois premières composantes, car elles expliquent collectivement une proportion substantielle de la variance totale (*53,18 % + 22,98 % + 10,92 % = 87,08 %*).

7- Quelles sont les variables que la deuxième composante oppose ?

Les variables qui s'opposent à la deuxième comparaison sont "CA" et "SUL".

En fonction de la charge de chaque variable sur cette composante :

CA	MG	NA.	K	**SUL**	NO3	HCO3	CL	MOY
0.53957	0.17631	-0.28861	-0.31936	**0.60300**	-0.06233	-0.34737	0.06560	0.00116

8- Quelle est la variable qui a la plus forte contribution dans la troisième composante.

La variable qui a la plus forte contribution dans la troisième composante est "**NO3**".

En fonction de la charge de chaque variable sur cette composante:

CA	MG	NA.	K	SUL	**NO3**	HCO3	CL	MOY
0.16247	0.01202	0.00498	-0.02944	0.02067	**0.96694**	0.12123	-0.12093	0.08856

9- Écrire un paragraphe (au plus 150 mots) pour résumer cette analyse

L'ACP a été appliquée à un jeu de données comprenant 57 marques d'eaux en bouteille, définies par des variables telles que la concentration en ions (CA, MG, NA, K, SUL, NO3, HCO3, CL), le type d'eau (minérale ou de source), la gazéification (plate ou gazeuse), et la masse moyenne des ions. Les données ont été préalablement centrées et réduites pour éviter les biais liés à l'échelle des variables, facilitant ainsi l'interprétation des résultats. La masse moyenne des ions a été ajoutée comme variable supplémentaire pour fournir une mesure synthétique de la composition ionique. Les deux premières composantes principales expliquent conjointement 76.15% de la variance totale, avec la première composante dominante à 53.18%, suivie de près par la deuxième (22.98%) et la troisième (10.92%). Trois composantes principales ont été recommandées pour cette étude en se basant sur les valeurs propres et les pourcentages de variance expliquée, soutenues par l'analyse du "scree plot" montrant une diminution significative des pourcentages après la troisième composante. Les charges des variables pour la deuxième composante ont identifié "CA" et "SUL" comme opposées à cette composante. Enfin, la troisième composante est fortement influencée par la variable "NO3" (ions nitrates), offrant une contribution significative à la variance expliquée. Cette information est précieuse pour comprendre l'impact de la concentration d'ions nitrates sur la structure des données.