14 février 2018

# EXERCICE D'ANALYSE EN COMPOSANTES PRINCIPALES

## 1. INTRODUCTION

Le but de l'exercice est de mettre en place une analyse en composante principale pour les données des nutriments présents dans différents échantillons d'eaux.

## 2. RESOLUTION

Le code a été réalisé en Python en utilisant les librairies *numpy* pour les matrices et *matplotlib* pour le tracé des graphes.

### A. Matrice centrée réduite et matrice des corrélations

On calcule la matrice centrée réduite M en enlevant aux valeurs des variables aléatoires leur moyenne observée et en divisant par leur écart-type. Les valeurs des moyennes, des variances et de la matrice peuvent être observées en exécutant le code joint.

On obtient ensuite la matrice des corrélations selon la formule :

$$R = \frac{1}{k} \, . M^T \, . M$$

La valeur de la matrice de corrélation peut être observée en exécutant le code.

### B. Valeurs propres, vecteurs propres et coordonnées projetées

Les valeurs propres et vecteurs propres sont obtenus facilement avec les fonctions de la librairies *numpy.* On conserve pour une représentation en 2D les 2 plus grandes valeurs propres et leur vecteur propres normés associés :

$$\begin{cases} lambda_1 = 3.8163 \\ lambda_2 = 2.0681 \end{cases}$$

Les vecteurs propres peuvent êtres observés en exécutant le code.

On projette ensuite la matrice centrée réduite selon les 2 composantes principales C1 et C2 qui correspondent aux 2 valeurs propres ci-dessus.

En représentant les valeurs dans le plan, on obtient alors le graphe suivant :

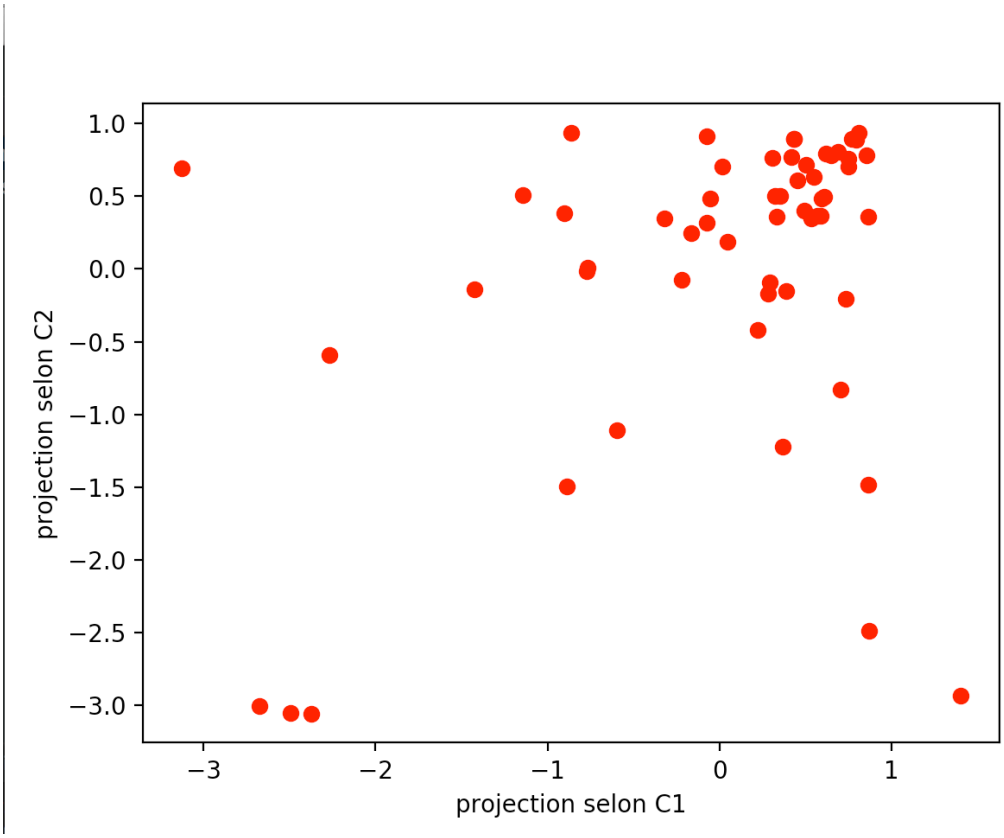


*Figure 1 : Représentation selon les 2 composantes principales*

On peut remarquer que certains points sont plutôt bien isolés.

L'inertie exprimée par (C1,C2) est de 73,5%. Ce chiffre est moyen et explique peut-être que l'on ait encore un grand groupe de points rapprochés.

### C. Corrélations

En calculant ensuite les corrélations par rapport à C1 et C2, on obtient le cercle des corrélations suivants :

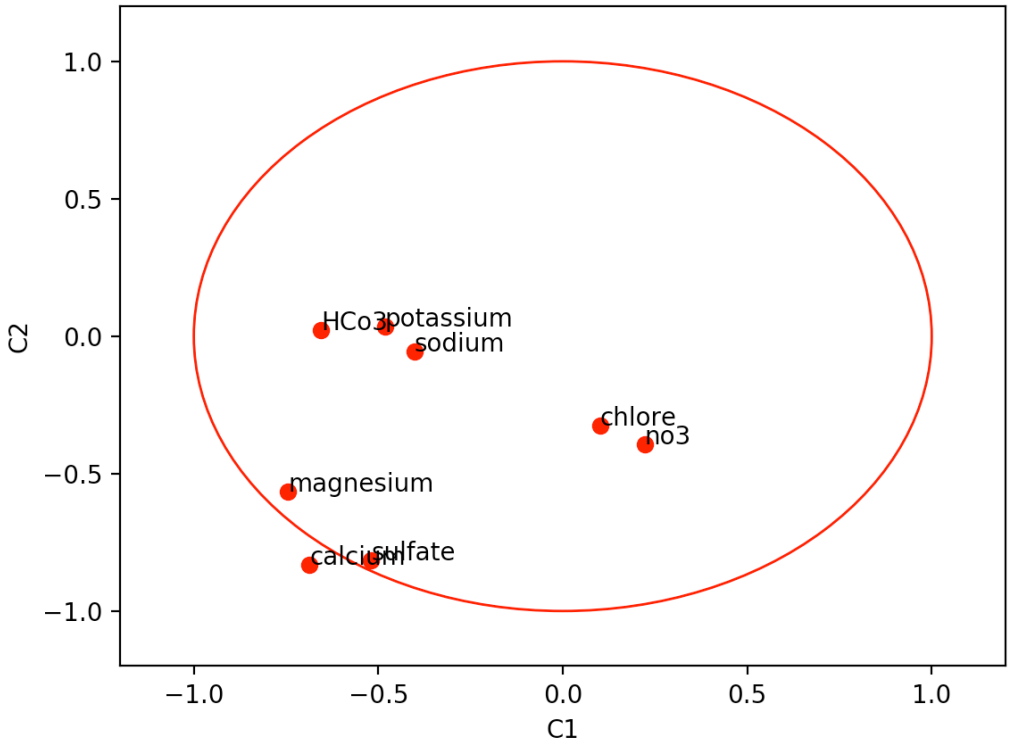


*Figure 2 : Cercle des corrélations*

On peut remarquer que les variables du magnesium, du calcium et du sulfate sont bien représentés selon C1 et C2 car elles sont proches du cercle unitaire. Ces variables ont un argument similaire dans le cercle.

La matrice centrée réduite comportait des valeurs très petites. Python a réalisé beaucoup d'arrondis. Ces erreurs d'arrondis sont multipliées au calcul de la matrice des corrélations et aggravées par les calculs suivants. Cela explique le fait que le point du calcium sorte du cercle je pense.

Les autres variables ne sont pas très bien représentées par C1 et C2 car elles sont éloignées du cercle unité, mais puisque le groupe (Chlore, No3) fait un angle d'environ 90° avec le groupe (potassium, sodium, HC03), on peut émettre l'hypothèse que ces 2 groupes de variables sont indépendants.