Analyse en composantes principales
- L3 Math, ISD, TP 3 (partie 1) -

Ce TP vise à initier aux différentes phases de l'analyse en composantes principales : préparation des données, mise en oeuvre de l'ACP, réalisation des graphiques, interprétation.

Partie 1 - Heptathlon, decastar 2023

Dans cette partie, nous considérons les résultats à l'épreuve d'heptathlon du decastar 2023. L'heptathlon est la version féminine du décathlon.

Préparation des données

Le fichier suivant donne accès aux scores des athlètes (source : decastar 2023).
https://perso.univ-rennes1.fr/valerie.monbet/ISD/heptathlon2023.csv
Ouvrir un nouveau notebook sous google colab [1]. charger les données avec la méthode `read_csv` de pandas. Vous pouvez utiliser un exemple du cours ou du 1er TP.

1. Charger les données avec la méthode `read_csv` de pandas.La table sera nommée `heptathlon`.
2. Utiliser la méthode `head` pour voir les premières lignes de la table.
3. Utiliser la méthode `info` pour obtenir une description synthétique de la table.
   Quelles sont les variables de la table de données ? Combien d'athlètes participent à la compétition ?
4. La table contient des données manquantes. On peut le vérifier à l'aide de la commande

   ```
   heptathlon.isnull().sum()
   ```

   Quelles variables présentent des données manquantes ?
   Deux athlètes n'ont pa terminé la compétition. Nous allons les retirer du jeu de données.

   ```
   heptathlon = heptathlon.dropna(axis=0)
   ```

5. La variable `800m` est une chaine de caractère (voir les informations sur la table). Aidez vous de l'exemple de la conversion des dates des données météo vue en cours pour convertir les temps de course en une variables numérique mesurée en secondes.

Analyse en composantes principales

6. Créer un tableau numpy nommé X qui contient les résultats des 7 épreuves de l'heptathlon.
7. Centré et réduire ce tableau à l'aide de la fonction StandardScaler et scikit learn.

1. depuis le drive, cliquer sur le bouton "+ nouveau" et dans la liste choisir "google colaboratory"

8. Réaliser l'analyse en composantes principales avec l'aide de la fonction PCA et scikit learn.
   On rappelle que la méthode `fit` permet d'obtenir un objet qui contient les axes factoriels[2] de l'ACP, la méthode `fit_transform` permet d'obtenir les composantes[3] de l'ACP.

Graphiques de l'ACP

9. Tracer un diagramme en barres pour montrer la décroissance des valeurs propres.
   Combien d'inertie est expliquée par les 2 premiers axes ?
   Combien d'axes doit-on conserver pour expliquer au moins 70% d'inertie ?
   Combien d'axes proposez vous de conserver pour la suite de l'analyse ? Pourquoi ?
10. Tracer, pour les axes factoriels conservés, des diagrammes en barres pour illustrer leur composition (en fonction des variables d'origine).
    Quelles variables expliquent le mieux le premier axe factoriel ? Le seond ?
11. Tracer la projection des observations sur le 1er plan factoriel. Vous pouvez utiliser plotly (ou matplotlib). Vous indiquerez le nom de l'athlète au dessus de son point.
    D'après les 2 dernières figures, quels sont les disciplines fortes de la néerlandaise Oosterwegel ?
    Il est intéressant de colorer les points par le score final (variable `Points`) comme dans la figure 1.

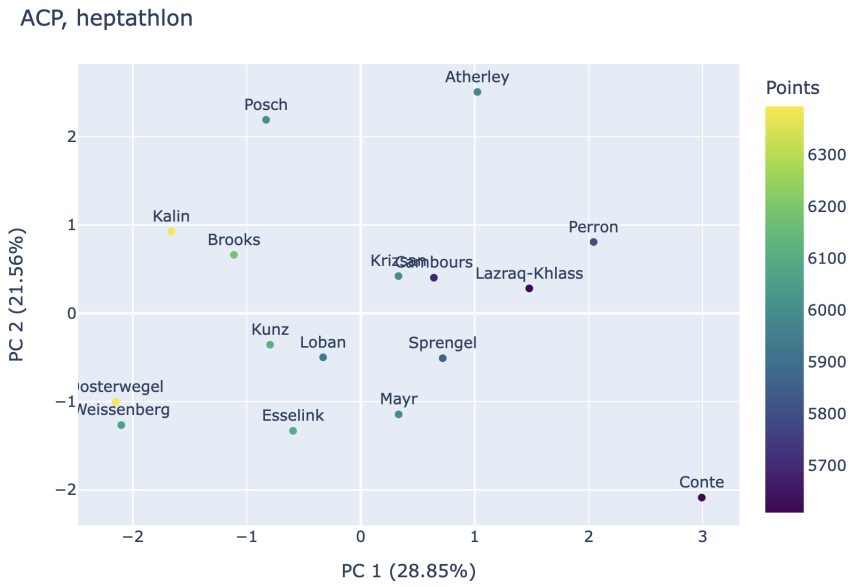

FIGURE 1 – Premier plan factoriel de l'ACP, données heptathlon 2023

2. *loadings*
3. *components*