Résumé des méthodes de l'analyse en composantes principale :

Table des matières

1. Introduction

Le datascientist, dans son travail doit réaliser plusieurs types d'étude sur les jeux de données qui l'intéresse. Parmi elles, les analyses en composantes principale permettent de mettre en lumière les potentielles corrélations entre les différentes variables du jeu de donnée.

Nous verrons donc tout d'abord la réalisation de l'import des données puis la manière dont un data analyste réalise des analyses bivariées et plus particulièrement, des analyses bivariées mixes.

Dans le cadre de ce TP, les packages « Pandas », « sklearn », « Numpy », « seaborn », « prince » et « matplotlib » ont été utilisés. Le jeu de donnée DecathlonData.csv, représentant les résultats d'athlètes dans les 10 disciplines du décathlon ainsi que leurs points, résultats et compétition de la performance, composera la base de notre étude.

2. Import des données

Afin d'apporter les données à étudiant dans python, on utilise le package Pandas qui simplifie la création et la manipulation de dataframe.

Suite à l'import des packages,

```
# Import des librairies
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
```

Il suffit d'utiliser read_csv, une fonction de pandas, pour lire un csv et le récupérer sous forme de dataframe. Exemple :

```
# Import des données
df = pd.read_csv('Data\DecathlonData.txt', sep='\t')
df.head()
```

Il

	centm	Longjump	Shotput	Highjump	quatrecentm	centdixhurdle	Discus	Polevault	Javeline	mille500m	Rank	Points	Competition
SEBRLE	11,04	7,58	14,83	2,07	49,81	14,69	43,75	5,02	63,19	291,7	1	8217	Decastar
CLAY	10,76	7,4	14,26	1,86	49,37	14,05	50,72	4,92	60,15	301,5	2	8122	Decastar
KARPOV	11,02	7,3	14,77	2,04	48,37	14,09	48,95	4,92	50,31	300,2	3	8099	Decastar
BERNARD	11,02	7,23	14,25	1,92	48,93	14,99	40,87	5,32	62,77	280,1	4	8067	Decastar
YURKOV	11,34	7,09	15,19	2,1	50,42	15,31	46,26	4,72	63,44	276,4	5	8036	Decastar

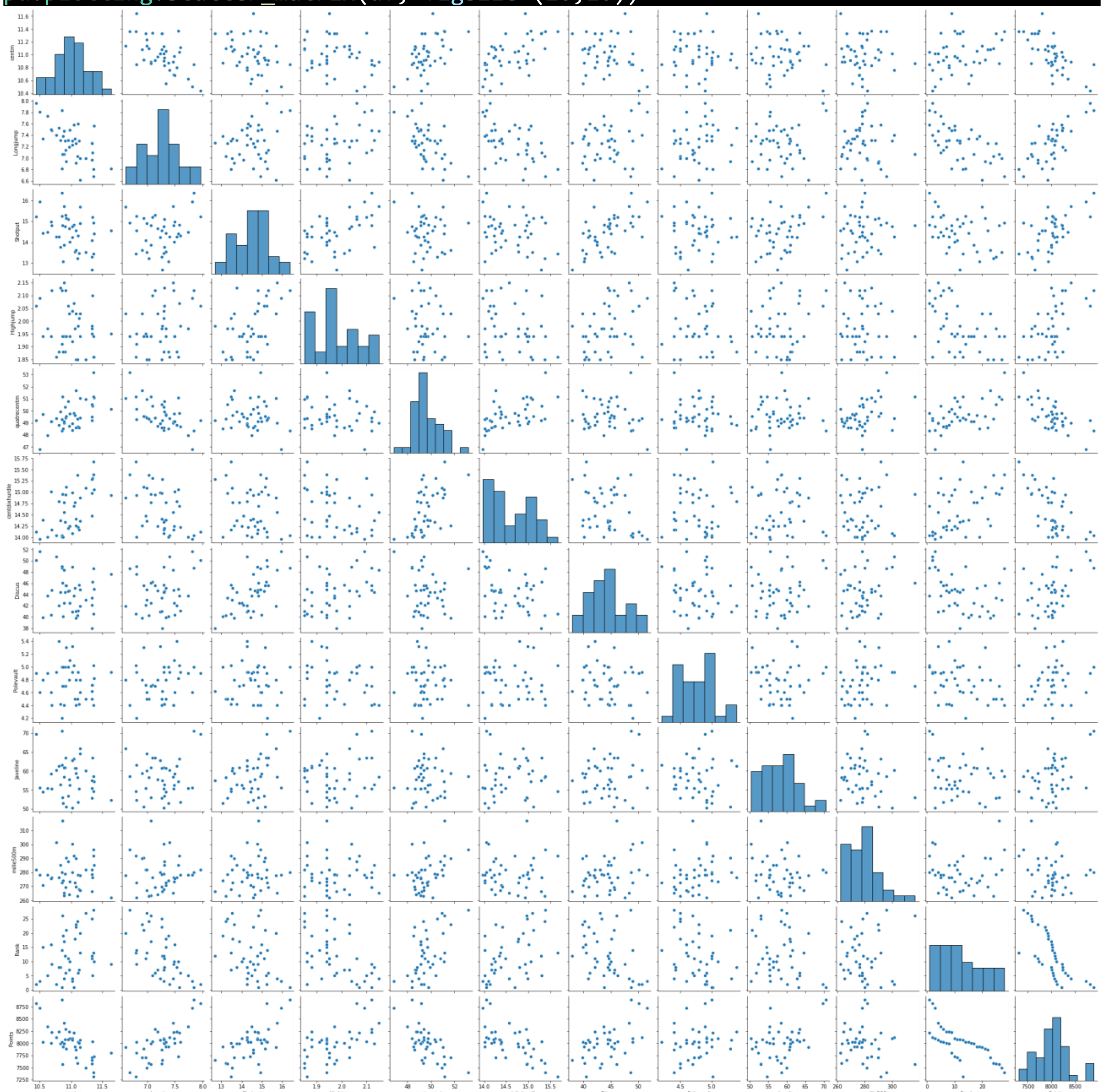
3. Étude de cas : Utilisation de l’ACP

Nous allons étudier le cas particulier de l’application de l’ACP à un jeu de donnée :

3.1. Nuage de points 2 à 2

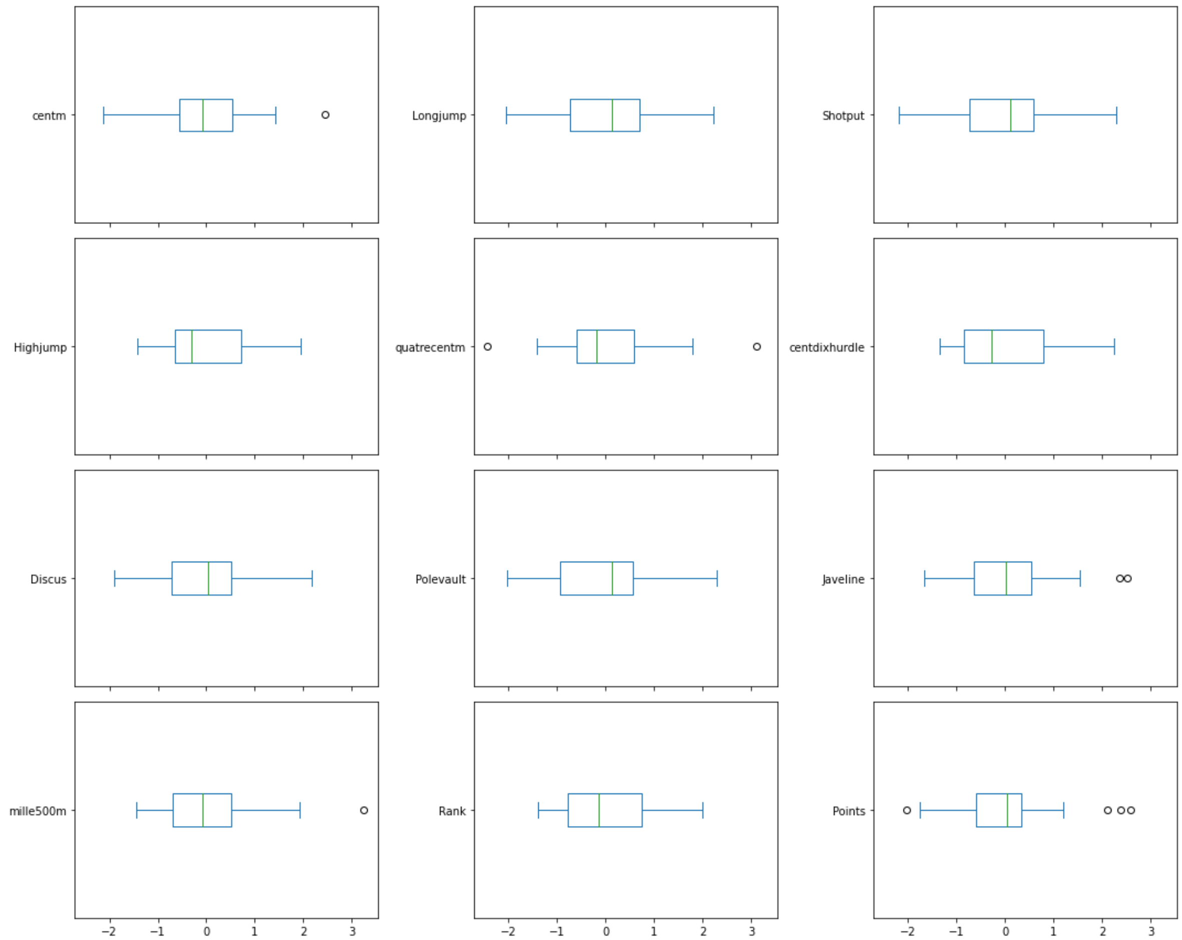
On réalise un graphique de nuages de points entres chaque variable explicative deux à deux :

```
pd.plotting.scatter_matrix(df, figsize=(10,10))
```



3.2. Diagrammes de Tukey

On réalise la même chose avec des diagrammes de Tukey sur le jeu de donnée standardisé pour étudier les valeurs atypiques :



3.3. Analyse en composantes principale

On récupère les valeurs composantes principales et les variances expliquées. Du fait que l'on possède 12 variables initiales on doit obtenir la variance expliqué de 12 composantes principales.

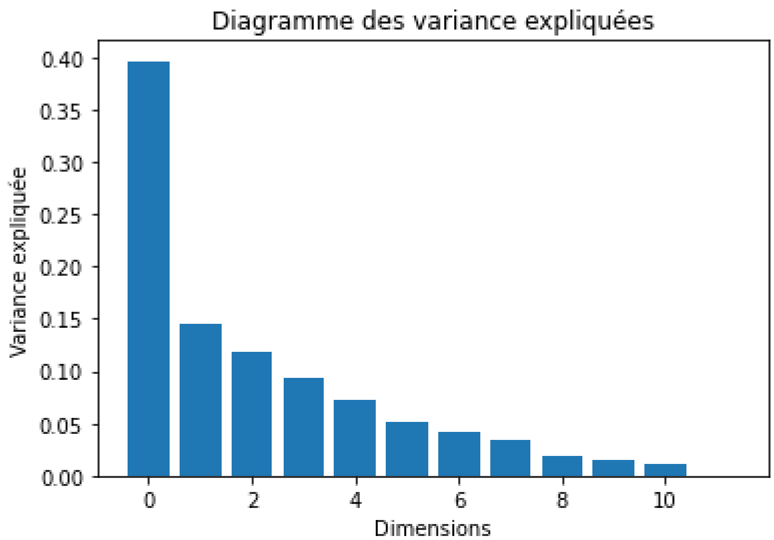
```
# Analyse en composantes principale
pca = PCA()
principal_components = pca.fit_transform(df_standardized)

# Variances expliquées
explained_variances = pca.explained_variance_ratio_
```

En étudiant les des variances en fonction du nombre de dimension on en vient à la conclusion que seul les corrélations entres deux variables suffisent à expliquer 54% des valeurs du jeu de données

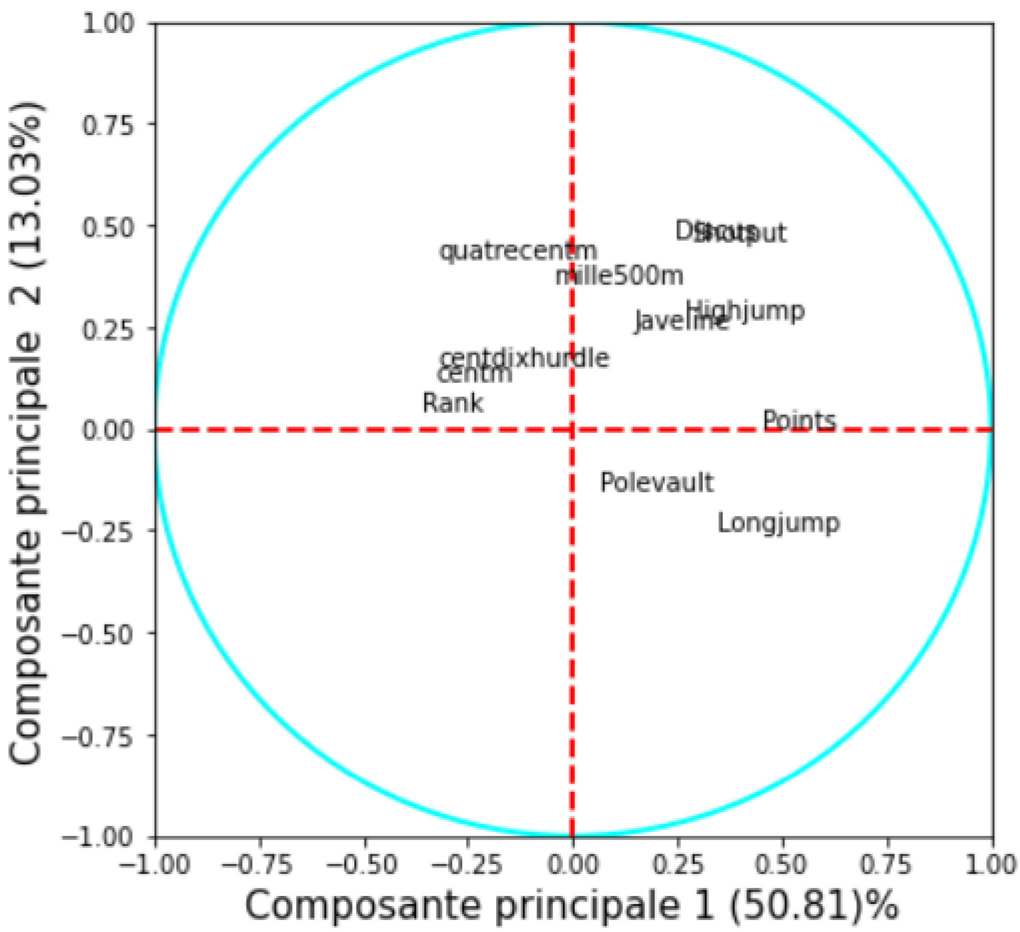
```
eig = pd.DataFrame(
    {
        "Dimension" : ["Dim" + str(x + 1) for x in range(12)],
        "Variance expliquée" : pca.explained_variance_,
        "% variance expliquée" : np.round(pca.explained_variance_ratio_ * 100),
        "% cum. var. expliquée" : np.round(np.cumsum(pca.explained_variance_ra-
tio_) * 100)
    }
)
eig
```

	Dimension	Variance expliquée	% variance expliquée	% cum. var. expliquée
0	Dim1	4.758790	40.0	40.0
1	Dim2	1.740146	15.0	54.0
2	Dim3	1.414902	12.0	66.0
3	Dim4	1.131778	9.0	75.0
4	Dim5	0.861942	7.0	83.0
5	Dim6	0.607319	5.0	88.0
6	Dim7	0.510451	4.0	92.0
7	Dim8	0.411084	3.0	95.0
8	Dim9	0.235209	2.0	97.0
9	Dim10	0.187364	2.0	99.0
10	Dim11	0.140961	1.0	100.0
11	Dim12	0.000054	0.0	100.0

Graphiquement on se rend compte que la courbe est exponentiellement descendante et qu'augmenter le nombre de dimension pour l'étude de l'ACP n'apporte qu'une quantité d'informations négligeable.

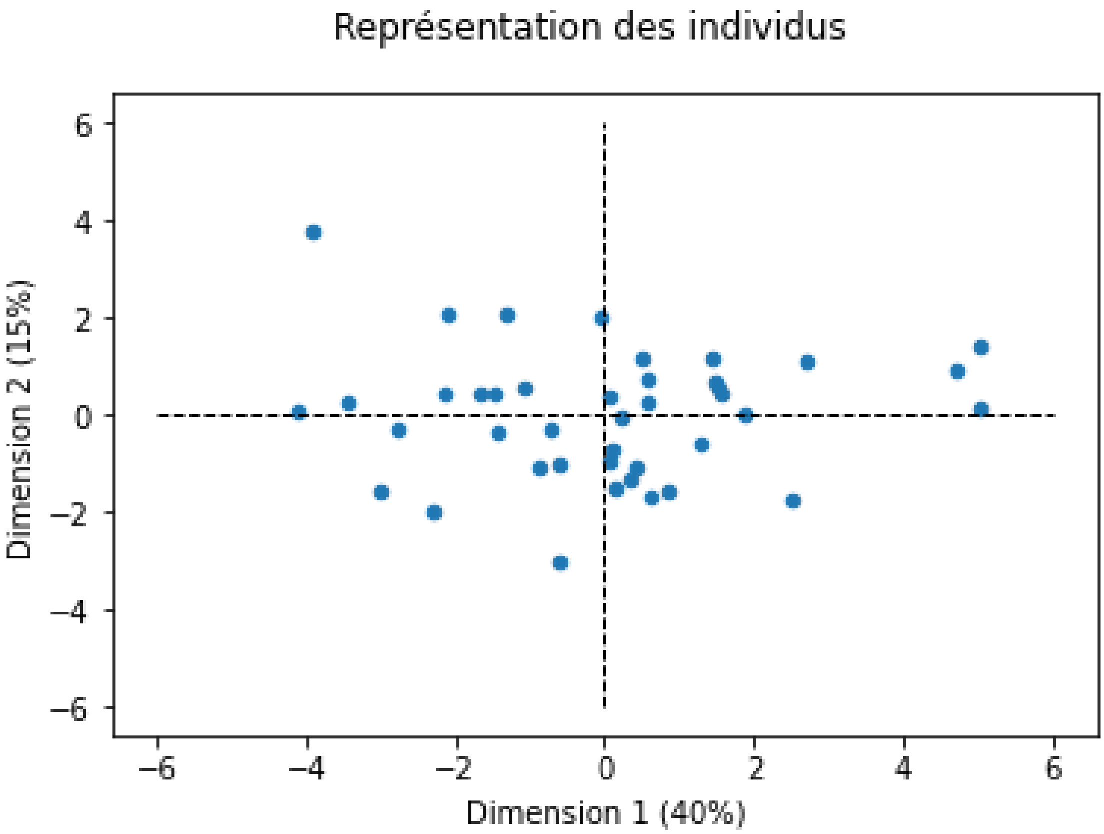


On réalise ensuite un cercle de corrélation afin de voir à quel point et de quel manière les données sont liées. Le cercle est réalisé avec seulement deux dimensions transportant 64% de l'information.



Ce cercle de corrélation nous permet de que la valeur des points et le rang du sportif sont fortement corrélé négativement. Plus les points d'un sportif sont élevés plus sont rang va diminuer (il se rapproche de la place de 1er). Le résultat du cercle de corrélation parait cohérent. On peut également remarquer que les sportifs obtiennent, à très peu de chose près, le même score entre le Shotput et le Discus (lancer de poids et lancer de disque, deux discipline très proche). Il reste cependant 35% de variance non capturé par le modèle ce qui peu indiquer que deux dimension sont quasi insuffisante pour capturer efficacement la variance du modèle.

3.4. Interprétation de l'ACP par nuage de point :



Avec deux dimensions on explique seulement 55% de l'information mais on remarque que les individus sont plus ou moins regroupé sur l'axe des abscisses et ne présente pas d'autres alignement notable. Cela implique que la valeur des données est moins impactée par la deuxième dimension que la première. Le fait que notre jeu de donnée présente de nombreuses variables initiales pourrai être la raison du peu d'informations utiles de ce graphique.