ACP – Python

Yéro Diamanka

2024-02-12

Nous allons explorer le jeu de données trees.csv qui donne des mesures de diamètre (girth, en pouces), hauteur (height, en pieds) et volume (volume, en pieds cubes) de cerisiers noirs.

0.1 Importation de bibliothèques !

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

```
trees=pd.read_csv('trees.csv')
```

1 - Centrer et réduire les données

```
scaler = StandardScaler()
scaler.fit(trees)
Z = scaler.transform(trees)
Z
```

2 - Lancer l'ACP

```
pca = PCA()
pca.fit(Z)
```

3 - Examiner les ratios cumulés de variance. Combien de variabilité est expliquée par les deux premiers axes ?

```
np.cumsum(pca.explained_variance_ratio_)
```

99% de la variabilité est expliquée par les deux premiers axes, c'est très élevé.

4 - Afficher les valeurs propres

```
l = 3*pca.explained_variance_ratio_
l
```

5 - Extraire les facteurs

```
Gn=pca.components_
G=np.empty(shape=Gn.shape)
for i in range(0, Gn.shape[0]):
    G[i,:]=Gn[i,:]*np.sqrt(l[i])
G
```

6 - Tracer le cercle des corrélation dans le premier plan factoriel pour les variables

```
fig,ax=plt.subplots(figsize=(5,5))
for i in range(0, Gn.shape[1]):
    ax.arrow(0,0,  # la flèche part de l'origine
             G[0, i],  G[1, i],  # et arrive en (G_1i,G_2i)
             head_width=0.05,head_length=0.07,length_includes_head=True)
    ax.text(G[0, i] + 0.01,G[1, i]-0.02, trees.columns[i],fontsize=8)
# affichage des lignes horizontales et verticales
ax.plot([-1, 1], [0, 0], color='grey', ls='--')
ax.plot([0, 0], [-1, 1], color='grey', ls='--')
# nom des axes, avec le pourcentage d'inertie expliqué
ax.set_xlabel('G{} ({}%)'.format(1,
round(100*pca.explained_variance_ratio_[0],1)))
ax.set_ylabel('G{} ({}%)'.format(2,
round(100*pca.explained_variance_ratio_[1],1)))
ax.set_title("Cercle des corrélations (G{} et G{})".format(1, 2))
an = np.linspace(0, 2 * np.pi, 100)
ax.plot(np.cos(an), np.sin(an))
```

Examiner le cercle des corrélation dans le premier plan factoriel pour les variables.

7 - Quelles variables sont bien représentées dans ce plan ?

Toutes : les pointes des flèches sont très proches du cercle unité. On le confirme en calulant les cos carrés.

```
Gsq = G**2
print(Gsq[0,:]+Gsq[1,:])
```

8 - Que pensez vous des positions relatives des variables Girth et Volume dans ce plan ?

Elles sont proches, donc très corrélées positivement.

9 - Quelles sont les variables qui contribuent le plus au premier axe factoriel ?

```
Contrib=(Gn**2)/np.sum(Gn**2,axis=0)
print(Contrib[0,:])
```

Ce sont les variables de diamlètre et de volume.

10 - Quelles sont les variables qui contribuent le plus au deuxième axe factoriel ?

```
print(Contrib[1,:])
```

C'est la variable de hauteur.

11 - Extraire les composantes principales

```
F=pca.fit_transform(Z)
F
```

12 - Tracer le nuage de point des individus projetés dans le premier plan factoriel.

```
fig,ax=plt.subplots(figsize=(5,5))
# individus
```

```
 3 ax.scatter(F[:,0],F[:,1])
 4 for i in range(trees.shape[0]):
 5     ax.text(F[i,0]+0.1,F[i,1],'{}'.format(i),fontsize=8)
 6 ax.set_xlabel('F{} ({}%)'.format(1,
round(100*pca.explained_variance_ratio_[0],1)))
 7 ax.set_ylabel('F{} ({}%)'.format(2,
round(100*pca.explained_variance_ratio_[1],1)))
 8 ax.set_title("Individus projetés (F{} et F{})".format(1, 2))
 9 ax.grid()
10 ax.plot([min(F[:,0]), max(F[:,0])],[0,0], linestyle="--", color='C7')
11 ax.plot([0, 0],[min(F[:,1]), max(F[:,1])], linestyle="--", color='C7')
```

13 - Quels sont les individus qui sont bien représentés dans ce plan ?

```
1 cos2ind = pd.DataFrame(
2     columns=[['axe 1','axe 2','somme']],
3     index=[np.arange(trees.shape[0])])
4 for i in np.arange(trees.shape[0]):
5     for k in np.arange(2):
6         cos2ind.iloc[i,k]= F[i,k]**2/(sum(Z[i,:]**2))
7     cos2ind.iloc[i,2]=cos2ind.iloc[i,0]+cos2ind.iloc[i,1]
8 cos2ind
```

Tous les individus sont très bien représentés sur ce plan. Seul l'arbre 15 est un peu moins bien reeprésenté que les autres.

14 - Que peut-on dire des arbres qui sont le plus à droite sur premier plan factoriel ?

Ce sont ceux de plus grand diamètre/volume

15 - Que peut-on dire des arbres qui sont le plus à gauche sur premier plan factoriel ?

Ce sont ceux de plus petit diamètre/volume

16 - Que peut-on dire des arbres qui sont le plus en haut (resp. le plus en bas) sur premier plan factoriel ?

Ce sont ceux de plus petite (resp. grande) hauteur.