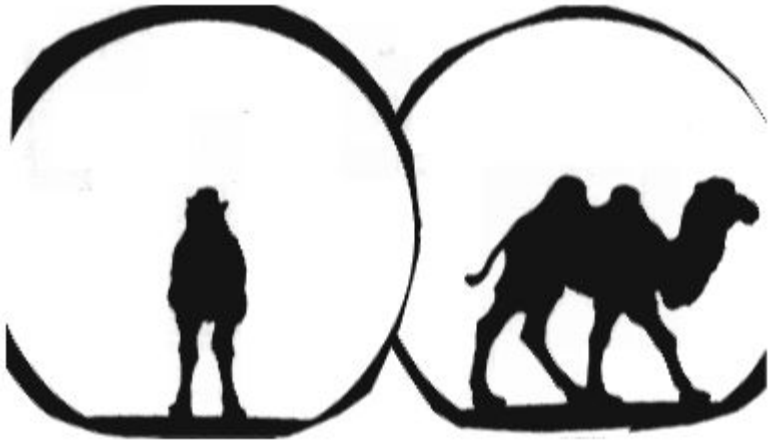
Analyse en composantes principales ACP

Master SD

- Quelle est la meilleure projection ?



Les données en ACP

En ACP les données se présentent dans un tableau X à n lignes et p colonnes où

- chaque ligne représente un individu
- chaque colonne représente une variable
- Les variables sont quantitatives : la matrice X est constituée de valeurs numériques

Les données en ACP

- *X* est une matrice $n \times p$ de valeurs numériques : *X* est une matrice $n \times p$ de valeurs numériques :

$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}$$

Les données en ACP

- Un individu est un élément de R*p*
  Le *i*ème individu :

$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{1p} \\ x_{21} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{i1} & \cdot & \cdot & \cdot & x_{ij} & \cdot & x_{ip} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{np} \end{bmatrix}$$

Les données en ACP

- Une variable est un élément de R*n*
  La *j*ème variable :

$$X = \begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1j} & \cdot & x_{1p} \\ x_{21} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & x_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{nj} & \cdot & x_{np} \end{bmatrix}$$

Données centrées

- Moyennes par colonnes :

$$\begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1j} & \cdot & x_{1p} \\ x_{21} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & x_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{nj} & \cdot & x_{np} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \bar{x}_1 & \cdot & \cdot & \cdot & \bar{x}_j & \cdot & \bar{x}_p \end{bmatrix}$$

- Centrage des données :

$$X = \begin{bmatrix} x_{11} - \bar{x}_1 & . & . & x_{1j} - \bar{x}_j & . & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & . & . & . & . & x_{2p} - \bar{x}_p \\ . & . & . & . & . & . \\ . & . & . & x_{ij} - \bar{x}_j & . & . \\ . & . & . & . & . & . \\ x_{n1} - \bar{x}_1 & . & . & x_{nj} - \bar{x}_j & . & x_{np} - \bar{x}_p \end{bmatrix}$$

Ecart-type

- On peut calculer l'écart-type pour chaque variable :

$$\begin{bmatrix} x_{11} & \cdot & \cdot & \cdot & x_{1j} & \cdot & x_{1p} \\ x_{21} & \cdot & \cdot & \cdot & \cdot & \cdot & x_{2p} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & x_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{n1} & \cdot & \cdot & \cdot & x_{nj} & \cdot & x_{np} \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \sigma_1 & \cdot & \cdot & \cdot & \sigma_j & \cdot & \sigma_p \end{bmatrix}$$

Données centrées-réduites

- Centrage puis réduction :

$$X = \begin{bmatrix} (x_{11} - \bar{x}_1)/\sigma_1 & . & . & (x_{1j} - \bar{x}_j)/\sigma_j & . & (x_{1p} - \bar{x}_p)/\sigma_p \\ (x_{21} - \bar{x}_1)/\sigma_1 & . & . & . & . & (x_{2p} - \bar{x}_p)/\sigma_p \\ . & . & . & . & . & . \\ . & . & . & (x_{ij} - \bar{x}_j)/\sigma_j & . & . \\ . & . & . & . & . & . \\ (x_{n1} - \bar{x}_1)/\sigma_1 & . & . & (x_{nj} - \bar{x}_j)/\sigma_j & . & (x_{np} - \bar{x}_p)/\sigma_p \end{bmatrix}$$

Coefficient de corrélation

- Rappel (coefficient de) corrélation de 2 variables :

$$\mathbf{cor}(X_j, X_k) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{\sigma_j}\right)\left(\frac{x_{ik} - \bar{x}_k}{\sigma_k}\right)$$

- C'est le produit scalaire des deux colonnes centrées-réduites associées (à 1=*n* près) :

$$X = \begin{bmatrix} . & (x_{1k} - \bar{x}_k)/\sigma_k & . & \leftrightarrow & . & (x_{1j} - \bar{x}_j)/\sigma_j & . \\ . & . & . & \leftrightarrow & . & . & . \\ . & . & . & \leftrightarrow & . & . & . \\ . & (x_{ik} - \bar{x}_k)/\sigma_k & . & \leftrightarrow & . & (x_{ij} - \bar{x}_j)/\sigma_j & . \\ . & . & . & \leftrightarrow & . & . & . \\ . & (x_{nk} - \bar{x}_k)/\sigma_k & . & \leftrightarrow & . & (x_{nj} - \bar{x}_j)/\sigma_j & . \end{bmatrix}$$

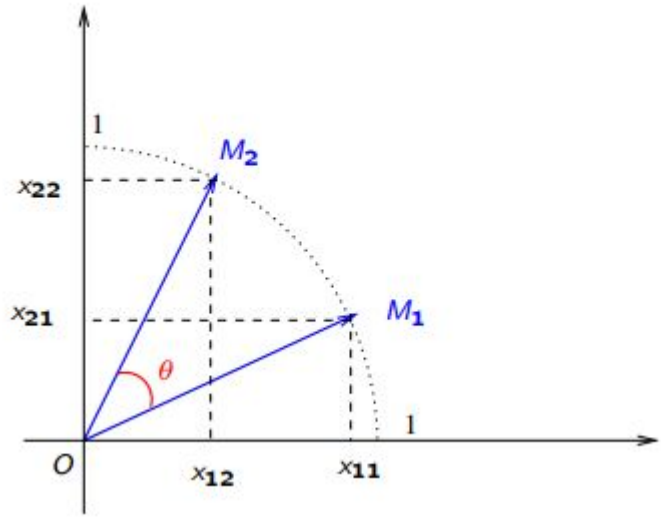
Rappel : Produit scalaire

- Pour des vecteurs de norme 1, la produit scalaire donne une mesure de l'angle (via le cos) :

$$\langle \overrightarrow{OM_1}, \overrightarrow{OM_2} \rangle = \cos(\theta) = x_{11}x_{12} + x_{21}x_{22}$$

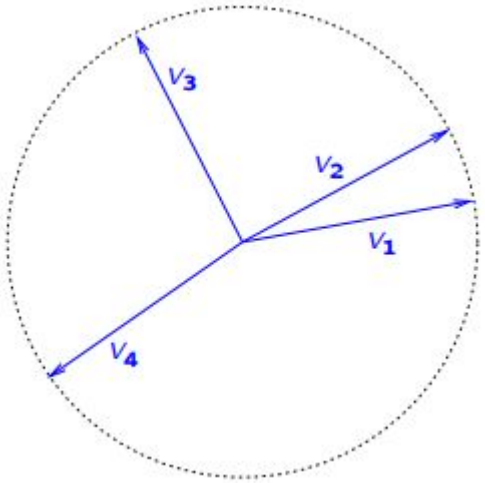
$$\overrightarrow{OM_1} = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$$

$$\overrightarrow{OM_2} = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$$



Interprétation

- *X* centrée-réduite *)* les colonnes ont même norme (≡ norme 1)
  - Les *p* colonnes sont alors dans une (hyper)sphère (de rayon 1)
  - L'angle formé par les vecteurs colonnes renseignent la corrélation sur les variables



$\mathbf{cor}(V_1, V_2) \approx 1$

$\mathbf{cor}(V_1, V_4) \approx \mathbf{cor}(V_2, V_4) \approx -1$

$\mathbf{cor}(V_1, V_3) \approx \mathbf{cor}(V_2, V_3) \approx \mathbf{cor}(V_4, V_3) \approx 0$

$$\mathrm{cov}(X_1, X_2) = \frac{\sum_{i=1}^{n} \left(X_{i,1} - \overline{X_1}\right)\left(X_{i,2} - \overline{X_2}\right)}{n-1}$$

C’est le signe de la covariance qui importe :

cov(X1 , X2) > 0 : X1 augmente quand X2 augmente

cov(X1 , X2) < 0 : X1 augmente quand X2 diminue

Matrice C de covariances

- Matrice C de covariances

$$C = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \dots & \mathrm{cov}(X_1, X_j) & \dots & \mathrm{cov}(X_1, X_p) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \dots & \mathrm{cov}(X_2, X_j) & \dots & \mathrm{cov}(X_2, X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathrm{cov}(X_j, X_1) & \mathrm{cov}(X_j, X_2) & \dots & \mathrm{var}(X_j) & \dots & \mathrm{cov}(X_j, X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \mathrm{cov}(X_p, X_1) & \mathrm{cov}(X_p, X_2) & \dots & \mathrm{cov}(X_p, X_j) & \dots & \mathrm{var}(X_p) \end{pmatrix}$$

- Propriétés :

C est une matrice carré de taille p x p

C est une matrice symétrique

ACP non-normée ou ACP normée?

S’il est recommandé de toujours « centrer » ses donnés en ACP, la question de les « réduire » (ACP normée) dépend de vos données :

- Si vos données sont toutes dans la même unité de mesure et varient dans des gammes de valeurs identiques : l’ACP non-normée suffit
- Si vos données sont dans des unités de mesure différentes et varient dans des gammes de valeurs différentes : l’ACP normée est recommandée

Notion de corrélation : ACP normée

- Si l'ACP est basée sur la matrice de covariances, l'ACP normée est basée elle sur la matrice de corrélations :

$$C = \begin{pmatrix} 1 & \rho(X_1,X_2) & \dots & \rho(X_1,X_j) & \dots & \rho(X_1,X_p) \\ \rho(X_2,X_1) & 1 & \dots & \rho(X_2,X_j) & \dots & \rho(X_2,X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho(X_j,X_1) & \rho(X_j,X_2) & \dots & 1 & \dots & \rho(X_j,X_p) \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho(X_p,X_1) & \rho(X_p,X_2) & \dots & \rho(X_p,X_j) & \dots & 1 \end{pmatrix}$$

Propriétés :

C est une matrice carré de taille p x p

C est une matrice symétrique

C possède une diagonale de 1

- Soit la matrice de covariances C de taille p x p, elle admet p valeurs propres et p vecteurs propres associés, tels que

$$CV_j = \lambda_j V_j$$

Choix du nombre d’axes à retenir

- Choix du nombre d’axes à retenir Deux critères empiriques pour sélectionner le nombre d’axes :
- Critère du coude : sur l’eboulis des valeurs propres, on observe un décrochement (coude) suivi d’une décroissance régulière. On sélectionne les axes avant le décrochement
- Critère de Kaiser: on ne retient que les axes dont l’inertie est supérieure à l’inertie moyenne I/p (un peu étroit). Kaiser en ACP normée: I/p= 1 : On ne retiendra que les axes associés à des valeurs propre supérieures à 1

- Construction des nuages de points projetés Chaque nuage de points (variables et individus) est construit en projection sur les plans factoriels : un plan factoriel est un repère du plan défini par deux des q axes factoriels retenus.
- Ex : Si l'on retient 3 axes, on tracera 3 graphique s pour chaque nuage: le nuage projeté sur le plan (axe1, axe2), celui projeté sur le plan (axe1, axe3), celui projeté sur le plan (axe2,axe3). L'examen des plans factoriels permettra de visualiser les corrélations entre les variables et d'identifier les groupes d'individus ayant pris des valeurs proches sur certaines variables.

Interprétation des axes

Pour chaque axe retenu et chaque nuage, on regarde

- Quelles sont les variables qui participent le plus à la formation de l'axe
- Quels sont les individus qui participent le plus à la formation de l'axe

Outil de mesure : contributions des points (individus si non anonymes et variables) à l'inertie de cet axe.

Ce sont les points dont la contribution est supérieure à la moyenne qui permettent de donner un sens à l'axe.

Limites

- Principale faiblesse de l’ACP: sensibilité aux points extrêmes. Ce manque de robustesse est notamment lie au rôle central qu'y joue le coefficient de corrélation : les points extrêmes, en perturbant les moyennes et corrélations, polluent fortement l'analyse - on peut cependant envisager de les déplacer en point supplémentaire. l'ACP est inadaptée aux phénomènes non linéaires qui plus est en grande dimension. Pour ce genre de problème, d'autres méthodes ont été développées, comme l'ACPN (Analyse en Composantes Principales par Noyau).

Quelles applications?

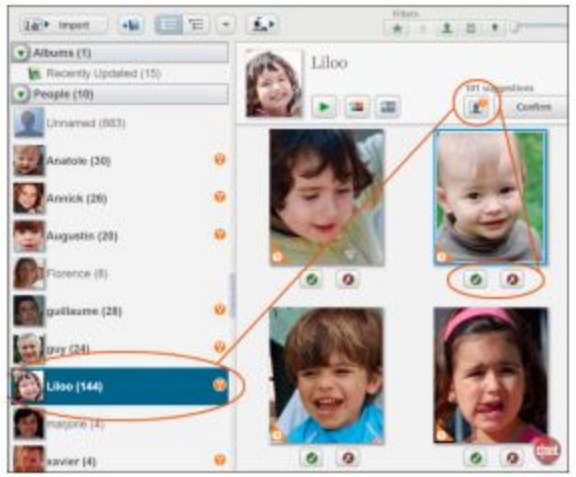
Analyses de données :

- Réduction du nombres de variables explicatives avant modélisation
- Obtention de nouvelles variables explicatives non corrélées

Imagerie :

- Compression d'image
- Reconnaissance faciale

Quelles applications?



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

X = pd.read_excel("cars.xlsx",sheet_name=0,header=0,index_col=0)



n = X.shape[0]
p= X.shape[1]

print(n)
sc = StandardScaler()
#transformation – centrage-réduction
Z = sc.fit_transform(X)
print(Z)
```

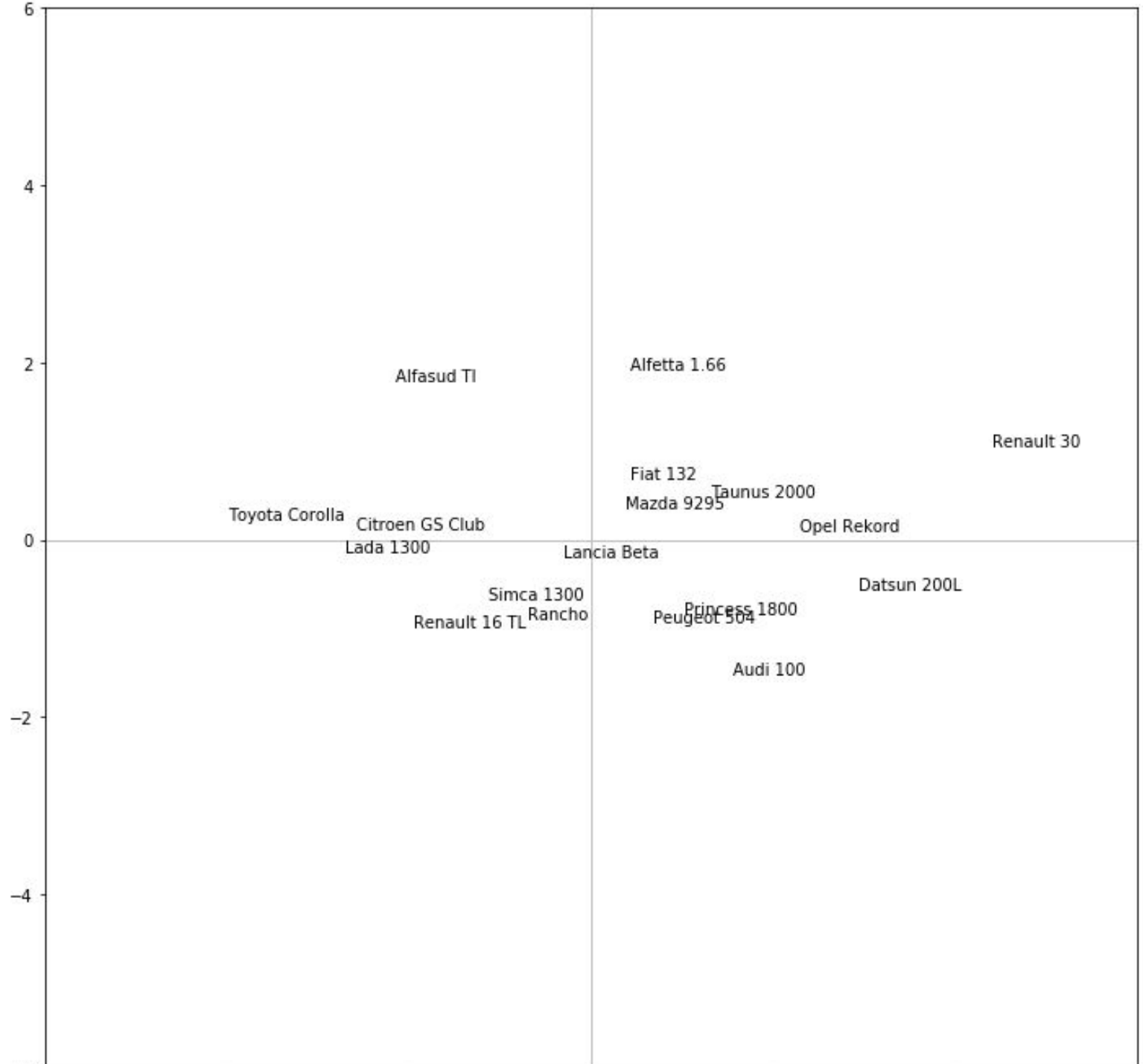
Modele	CYL	PUISS	LONG	LARG	POIDS	V_MAX
Alfasud TI	1350	79	393	161	870	165
Audi 100	1588	85	468	177	1110	160
Simca 1300	1294	68	424	168	1050	152
Citroen GS Club	1222	59	412	161	930	151
Fiat 132	1585	98	439	164	1105	165
Lancia Beta	1297	82	429	169	1080	160
Peugeot 504	1796	79	449	169	1160	154
Renault 16 TL	1565	55	424	163	1010	140
Renault 30	2664	128	452	173	1320	180
Toyota Corolla	1166	55	399	157	815	140
Alfetta-1.66	1570	109	428	162	1060	175
Princess-1800	1798	82	445	172	1160	158
Datsun-200L	1998	115	469	169	1370	160
Taunus-2000	1993	98	438	170	1080	167
Rancho	1442	80	431	166	1129	144
Mazda-9295	1769	83	440	165	1095	165
Opel-Rekord	1979	100	459	173	1120	173
Lada-1300	1294	68	404	161	955	140

```
[[-0.77509889 -0.28335818 -1.88508077 -1.09734528 -1.56900676  0.56976043]
 [-0.12016326  0.01963869  1.60580955  2.0010414   0.23416142  0.14597168]
 [-0.92920139 -0.83885242 -0.44217944  0.25819889 -0.21663062 -0.53209032]
 [-1.12733318 -1.29334771 -1.00072189 -1.09734528 -1.11821472 -0.61684807]
 [-0.12841875  0.67613189  0.25599862 -0.51639778  0.19659542  0.56976043]
 [-0.9209459  -0.13185975 -0.20945342  0.45184806  0.0087654   0.14597168]
 [ 0.45221746 -0.28335818  0.72145067  0.45184806  0.60982146 -0.36257482]
 [-0.18345536 -1.49534562 -0.44217944 -0.71004695 -0.51715865 -1.54918332]
 [ 2.84080623  2.19111619  0.86108628  1.22644473  1.81193359  1.84112668]
 [-1.28143568 -1.49534562 -1.60580955 -1.87194195 -1.98223281 -1.54918332]
 [-0.16969621  1.23162613 -0.25599862 -0.90369611 -0.14149861  1.41733793]
 [ 0.45772112 -0.13185975  0.53526985  1.03279556  0.60982146 -0.02354382]
 [ 1.0080872   1.53462299  1.65235475  0.45184806  2.18759363  0.14597168]
 [ 0.99432805  0.67613189  0.20945342  0.64549722  0.0087654   0.73927593]
 [-0.5219305  -0.2328587  -0.11636301 -0.12909944  0.37691224 -1.21015232]
 [ 0.37791804 -0.08136027  0.30254383 -0.32274861  0.12146341  0.56976043]
 [ 0.95580242  0.77713084  1.18690271  1.22644473  0.30929343  1.24782243]
 [-0.92920139 -0.83885242 -1.37308353 -1.09734528 -0.9303847  -1.54918332]]
```

```
eigval = ((n-1.0)/n)*acp.explained_variance_
print(eigval)
print(acp.explained_variance_)
```

```
[4.42085806 0.85606229 0.37306608 0.21392209 0.09280121 0.04329027]
[4.68090853 0.90641889 0.39501114 0.22650574 0.09826011 0.04583676]
```

```
#positionnement des individus dans le premier plan
fig, axes = plt.subplots(figsize=(12,12))
axes.set_xlim(-6,6) #même limites en abscisse
axes.set_ylim(-6,6) #et en ordonnée
#placement des étiquettes des observations
for i in range(n):
 plt.annotate(X.index[i],(coord[i,0],coord[i,1]))
#ajouter les axes
plt.plot([-6,6],[0,0],color='silver',linestyle='-',linewidth=1)
plt.plot([0,0],[-6,6],color='silver',linestyle='-',linewidth=1)
#affichage
plt.show()
```


6
4
2
0
−2
−4
Alfasud TI
Alfetta 1.66
Renault 30
Fiat 132
Taunus 2000
Mazda 9295
Toyota Corolla
Citroen GS Club
Opel Rekord
Lada 1300
Lancia Beta
Datsun 200L
Simca 1300
Princess 1800
Rancho
Peugeot 504
Renault 16 TL
Audi 100

Représentation des variables – Outils pour l'aide à l'interprétation

- Nous avons besoin des vecteurs propres pour l'analyse des variables. Ils sont fournis par le champ **.components_**

```
print(acp.components_)
```

```
[[ 0.42493602  0.42179441  0.42145993  0.38692224  0.43051198  0.35894427]
 [ 0.12419108  0.41577389 -0.41181773 -0.446087   -0.24267581  0.6198626 ]
 [-0.35361252 -0.18492049  0.06763394  0.60486812 -0.48439601  0.48547226]
 [ 0.80778648 -0.35779199 -0.27975231  0.21156941 -0.30171136 -0.0735743 ]
 [ 0.15158003 -0.29373465  0.73056903 -0.47819008 -0.30455842  0.18865511]
 [-0.05889517 -0.63303302 -0.19029153 -0.10956624  0.5808122   0.45852167]]
```

les facteurs sont en ligne, les variables en colonne

corrélation des variables avec les axes

```
#racine carrée des valeurs propres
sqrt_eigval = np.sqrt(eigval)

print('_____________________________________________________________')
#corrélation des variables avec les axes
corvar = np.zeros((p,p))
for k in range(p):
 corvar[:,k] = acp.components_[k,:] * sqrt_eigval[k]
#afficher la matrice des corrélations variables x facteurs
print(corvar)
```

```
[[ 0.89346354  0.1149061  -0.21598347  0.37361508  0.04617627 -0.01225391]
 [ 0.88685803  0.38468911 -0.11294784 -0.16548492 -0.08948124 -0.13171084]
 [ 0.88615477 -0.38102873  0.04131023 -0.12939024  0.22255537 -0.03959265]
 [ 0.81353638 -0.4127359   0.36944822  0.09785447 -0.14567244 -0.0227967 ]
 [ 0.90518746 -0.22453248 -0.29586489 -0.13954667 -0.09277852  0.12084561]
 [ 0.75471037  0.57351941  0.29652226 -0.03402937  0.05747056  0.09540146]]
```

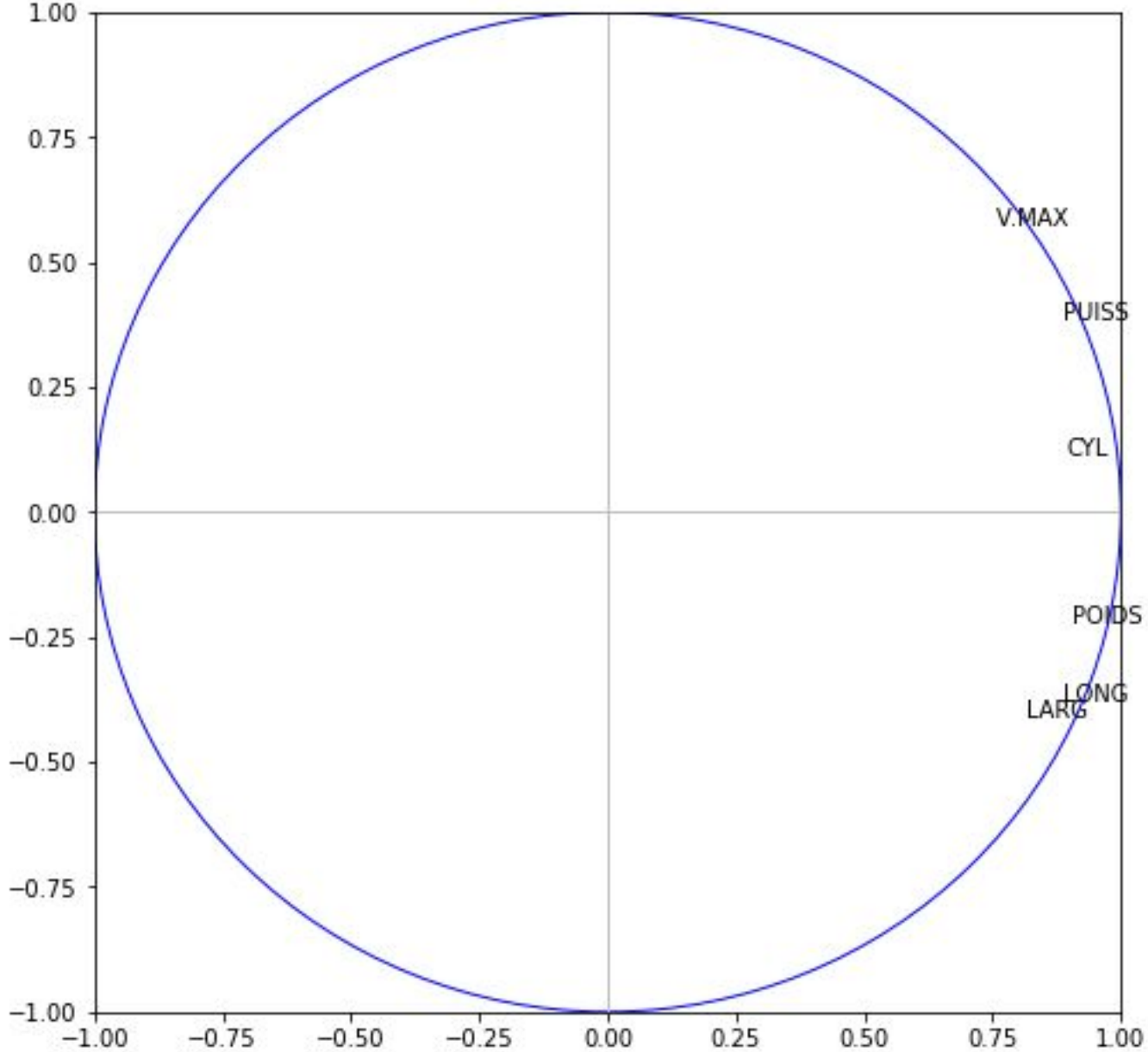
Les variables sont maintenant en ligne, les facteurs en colonne :

```
#on affiche pour les deux premiers axes
print(pandas.DataFrame({'id':X.columns,'COR_1':corvar[:,0],'COR_2':corvar[:,1]}))
```

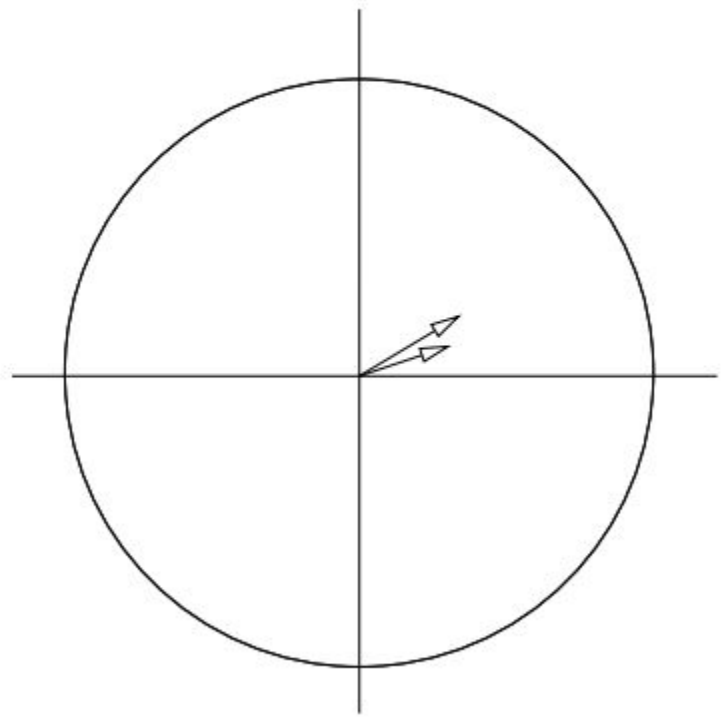
	COR_1	COR_2	id
0	0.893464	0.114906	CYL
1	0.886858	0.384689	PUISS
2	0.886155	-0.381029	LONG
3	0.813536	-0.412736	LARG
4	0.905187	-0.224532	POIDS
5	0.754710	0.573519	V.MAX

cercle des corrélations

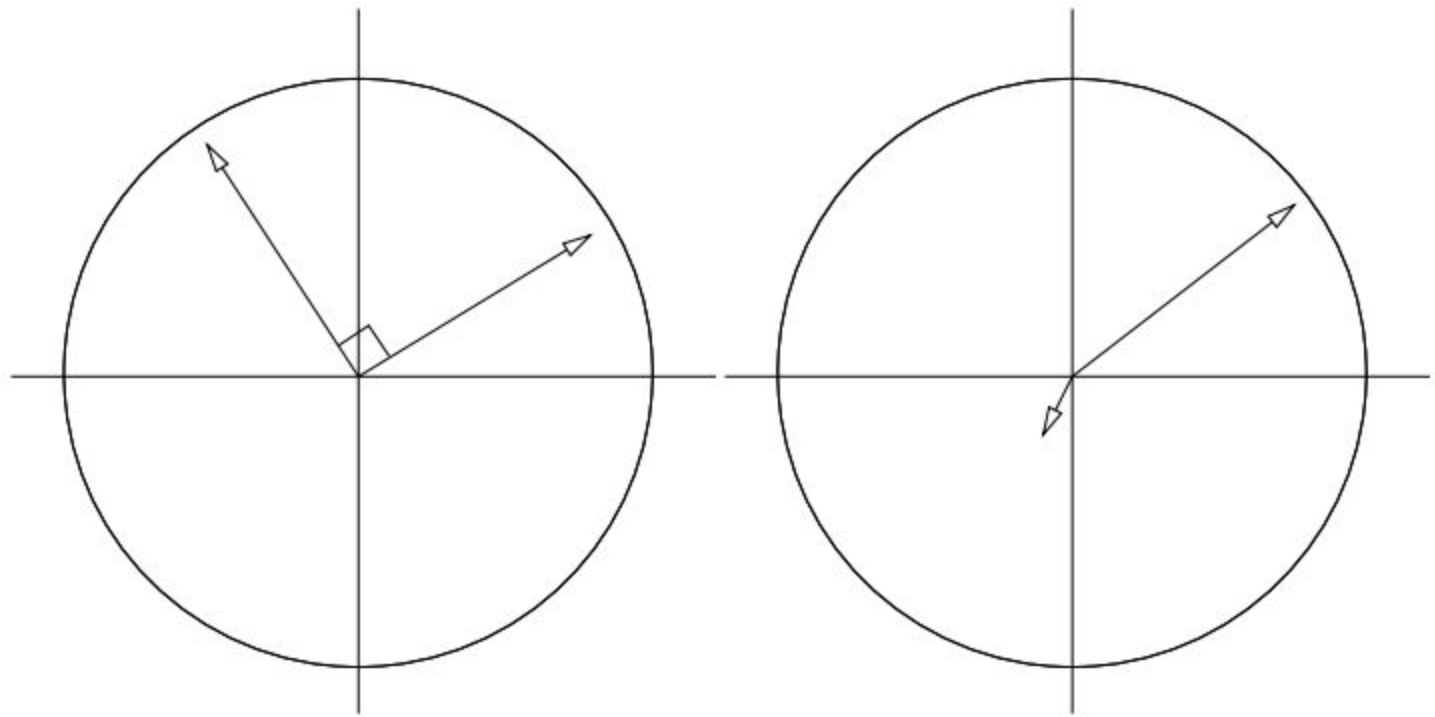
```
#cercle des corrélations
fig, axes = plt.subplots(figsize=(8,8))
axes.set_xlim(-1,1)
axes.set_ylim(-1,1)
#affichage des étiquettes (noms des variables)
for j in range(p):
 plt.annotate(X.columns[j],(corvar[j,0],corvar[j,1]))
#ajouter les axes
plt.plot([-1,1],[0,0],color='silver',linestyle='-',linewidth=1)
plt.plot([0,0],[-1,1],color='silver',linestyle='-',linewidth=1)
#ajouter un cercle
cercle =
plt.Circle((0,0),1,color='blue',fill=False)
axes.add_artist(cercle)
#affichage
plt.show()
```


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
V.MAX
PUISS
CYL
POIDS
LONG
LARG

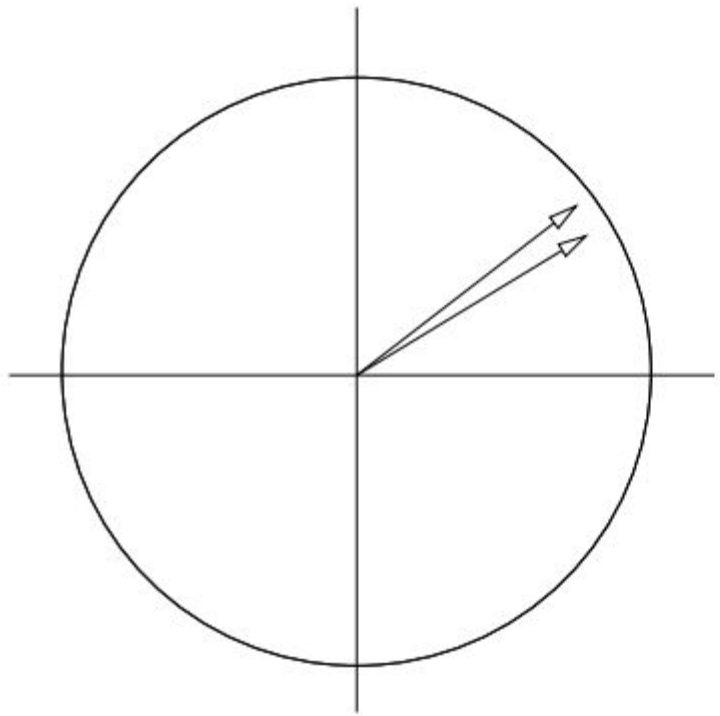
Aucune interprétation



Non corrélation



Corrélation positive



Corrélation négative