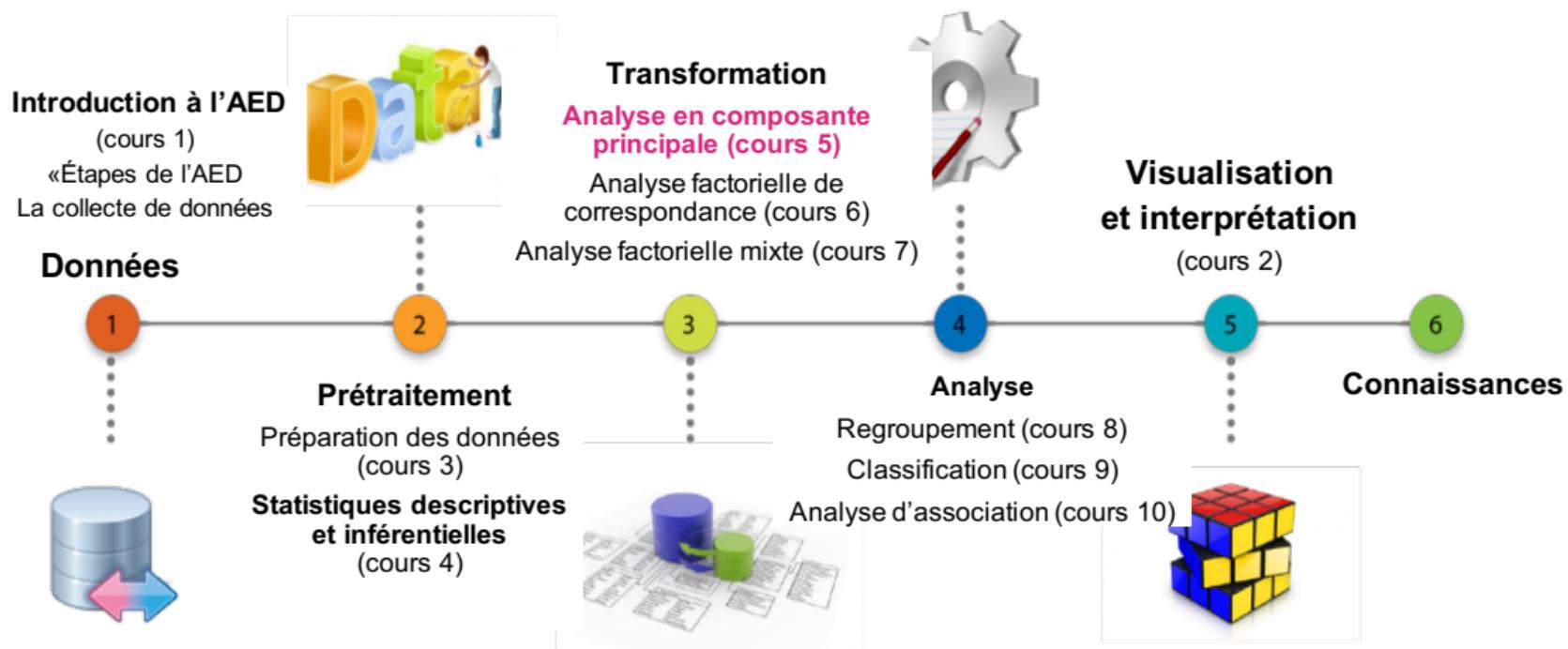


# Cours 5 - Analyse en composante principale (ACP)

Neila Mezghani

9 & 16 mars 2021



# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

# Plan du cours

- 1** Introduction
- 2** Les données et leurs caractéristiques
- 3** Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4** Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Définition

Selon le *Dictionary of Statistics and Methodology*, l'analyse en composantes principales est :

*« Un ensemble de méthodes permettant de procéder à des transformations linéaires d'un grand nombre de variables inter corrélées de manière à obtenir un nombre relativement limité de composantes non corrélées. Cette approche facilite l'analyse en regroupant les données en des ensembles plus petits et en permettant d'éliminer les problèmes de multi colinéarité entre les variables ».*

## Objectif

- L'analyse en composantes principales (ACP) est une méthode d'analyse de données très connue en statistique et dans les sciences expérimentales.
- Elle consiste à rechercher les directions de l'espace qui représentent le mieux les corrélations dans un ensemble de données.
- Ceci a pour objectifs de réduire la dimension des caractéristiques, de les visualiser et d'interpréter et analyser les corrélations entre ces données.

## Illustration (1)

La figure représente une photographie réelle d'un chameau. Il s'agit d'une image représentée dans un espace à 3 dimensions.

- Nous désirons passer à un espace de dimension moindre, par exemple, un espace de 2 dimensions.
- La question qui se pose alors : dans quel plan faudrait-il **projeter** les données pour réduire la dimension de 3 à 2 ?

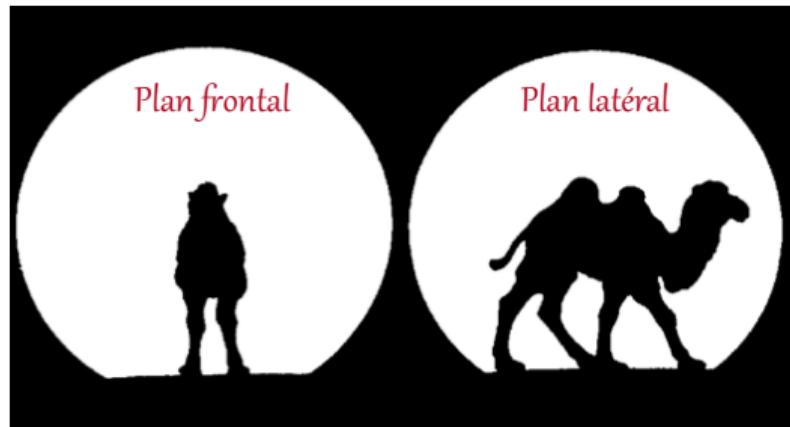


## Illustration (2)

Photographie réelle (3D)



Projection



La projection dans le plan latéral montre une meilleure préservation de l'information que la projection dans le plan frontal

# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Les données (1)

- Les données, aussi appelée « observations » ou bien « points », sont généralement représentées sous la forme d'un tableau rectangulaire (ou matrice) à  $N$  lignes représentant les individus et  $K$  colonnes correspondant aux variables.
- On note  $\mathbf{M}$  la matrice de dimension  $(N, K)$  contenant les observations.

$$\mathbf{M} = \begin{pmatrix} x_1^1 & \cdot & \cdot & x_1^K \\ x_2^1 & \cdot & \cdot & x_2^K \\ \cdot & \cdot & \cdot & \cdot \\ x_N^1 & \cdot & \cdot & x_N^K \end{pmatrix}$$

où  $x_i^j$  est la valeur de l'individu  $i$  pour la variable  $j$ .

## Les données (2)

- On notera :  
 $\mathbf{x}_i = (x_i^1, \dots, x_i^K)'$  le vecteur des variables de l'individu  $i$   
et  
 $\mathbf{x}^j = (x_1^j, \dots, x_N^j)'$  le vecteur des individus de la variable  $j$ .
- La matrice  $M$  est généralement centrée et réduite pour des fins de normalisations afin de remédier à l'écart entre les valeurs des différentes variables et afin de mieux interpréter les résultats.

## Les données (3)

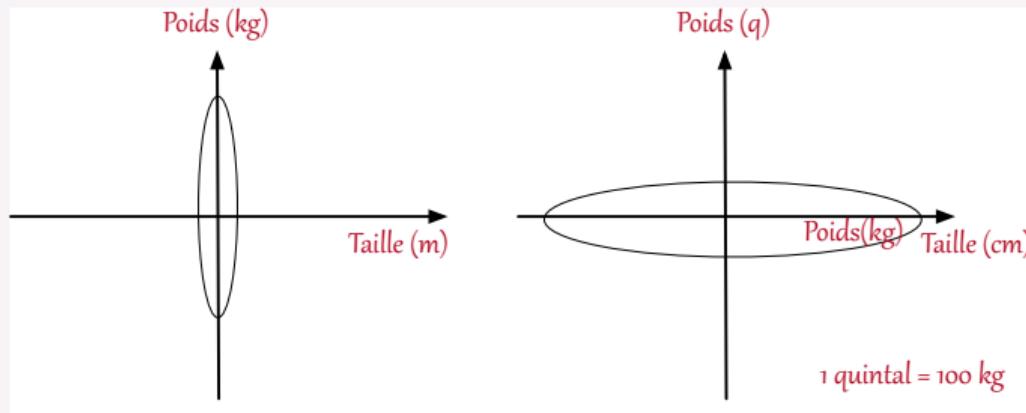
- Le centrage consiste à enlever la moyenne à chaque variable alors que la réduction consiste à diviser les valeurs de chaque variable par l'écart-type.  
Soit  $\mathbf{X}$  la matrice centrée réduite obtenue à partir de  $\mathbf{M}$ .

$$\mathbf{X} = \begin{pmatrix} \frac{x_1^1 - \bar{x}_1}{\sigma_1} & \dots & \frac{x_1^p - \bar{x}_K}{\sigma_K} \\ \frac{x_2^1 - \bar{x}_1}{\sigma_1} & \dots & \frac{x_2^K - \bar{x}_K}{\sigma_K} \\ \vdots & \ddots & \vdots \\ \frac{x_N^1 - \bar{x}_1}{\sigma_1} & \dots & \frac{x_N^K - \bar{x}_K}{\sigma_K} \end{pmatrix}$$

$\bar{x}_{j \in \{1, \dots, K\}}$  est la moyenne de la variable  $j$  et  $\sigma_{i \in \{1, \dots, K\}}$  est la variance.

## Pourquoi centrer et réduire les données ?

- Centrer les données = retirer la moyenne  $\Rightarrow$  Centrer les données en termes d'individus = Mettre le centre de gravité à l'origine du nuage des points des individus.
- Réduire les données = Diviser par l'écart type  $\Rightarrow$  Les dispersions deviennent homogènes (parce qu'on divise par l'écart type)



## Exemple : Les données (1)

Soit la base données « autos » décrivant les caractéristiques d'autos (Saporta, 2006 ; page 428)

- ① CYL : Cylindrée en cm<sup>3</sup>
- ② PUISS : La puissance en chevaux
- ③ LONG : Longueur en cm
- ④ LARG : Largeur en cm
- ⑤ POIDS : Poids en kg
- ⑥ VMAX : Vitesse maximale en km h

## Exemple : Les données (2)

Les 18 observations

Les 6 variables

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

$x_1^2$

Le vecteur  $\mathbf{x}_1 = (1350, 79, 393, 161, 870, 165)$  est caractéristique de l'observation 1

Le vecteur  $\mathbf{x}^1 = (1350, 1588, 1294, \dots, 1294)$  est caractéristique de la variable 1

## Nuage de points (1)

- La matrice  $\mathbf{X}$  permet de définir deux nuages de points.
- Le premier correspond au nuage des variables, qui est constitué par les coordonnées des vecteurs variables tracées dans le repère dont les axes représentent les individus. La dimension de cet espace correspond au nombre d'individus, c'est-à-dire  $N$  dimension.
- Le deuxième, le nuage des individus, est constitué des coordonnées des vecteurs individus tracées dans le repère dont les axes représentent les variables. Cet espace est de dimension  $K$ .

## Nuage de points (2)

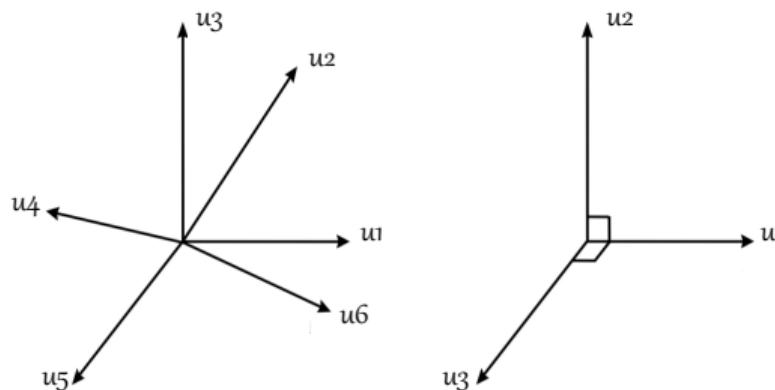
- Analyser la typologie des individus  $\Rightarrow$  former des groupes d'individus semblables
  - Terme clé : Ressemblance
- Analyser la typologie des variables  $\Rightarrow$  former des groupes de variables liées
  - Termes clé : Liaison - Corrélation
- Analyser le/les groupes de variables qui expliquent le plus la variabilité inter-individus ?

## Nuage de points (3)

- Comme les espaces de représentation des individus et des variables sont souvent de grandes dimensions, il est impossible de les visualiser selon des nuages de points dès que la dimension dépasse 3.

Espace ne pouvant pas être visualisé

Espace pouvant être visualisé



## Exemple : Nuage de points

Les 18 observations

Les 6 variables

Modèle	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

$x_1^2$

Nuage des individus

→ Un ensemble de 18 points dont les coordonnées sur les 6 axes correspondent aux valeurs des variables

Nuage de points des variables

→ Un ensemble de 6 points dont les coordonnées sur les 18 axes correspondent aux valeurs des individus



Nuage de points des variables et des individus  
Impossible à visualiser

## Exercice : Étape 1 - Lecture et préparation des données

---

- ➊ À partir du fichier Cars-Saporta.xls, téléchargez le contenu de la base de données et affichez son contenu.
- ➋ Réalisez une standardisation des données en vue d'une analyse ACP.
- ➌ Vérifiez que les données sont bien centrées et réduites.

# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

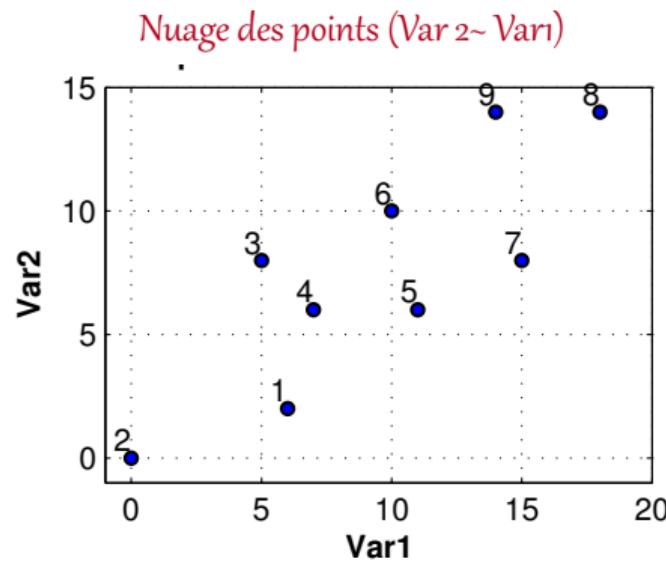
# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Projection des données (1)

- Comme les espaces de représentation des individus et des variables sont souvent de grandes dimensions, il est impossible de les visualiser selon des nuages de points dès que la dimension dépasse 3.
- On va déterminer un sous-espace de dimension  $k < K$  ( $k$  nouveaux axes dans le cas des individus) sur lequel on pourrait projeter les nuages de points tout en respectant, le plus possible, la configuration initiale des données  $\Rightarrow$  **L'importance du choix de la projection**

## Exemple : Illustration de l'importance du choix de la projection (1)

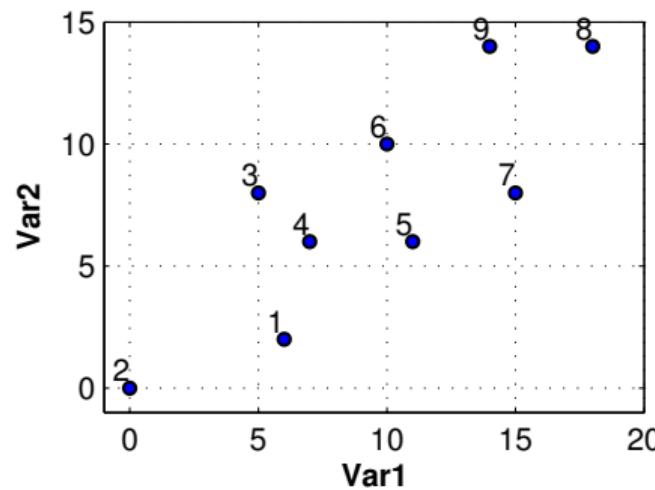


Ensemble des données  
(9 observations, 2 variables)

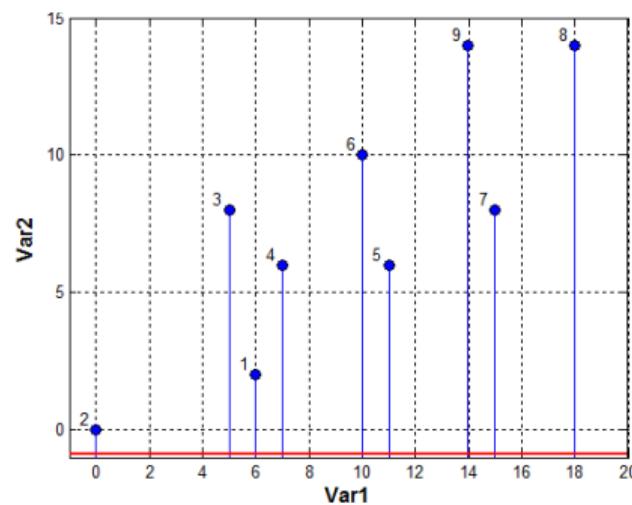
Station	Var1	Var2
1	6	2
2	0	0
3	5	8
4	7	6
5	11	6
6	10	10
7	15	8
8	18	14
9	14	14

## Exemple : Illustration de l'importance de la projection (2)

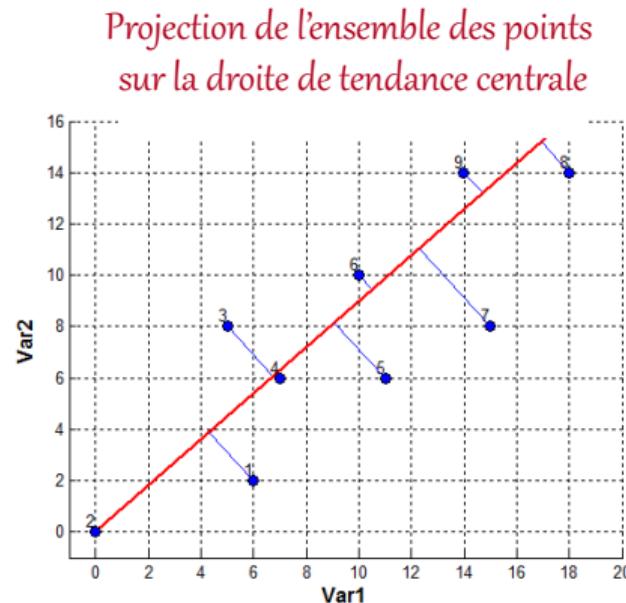
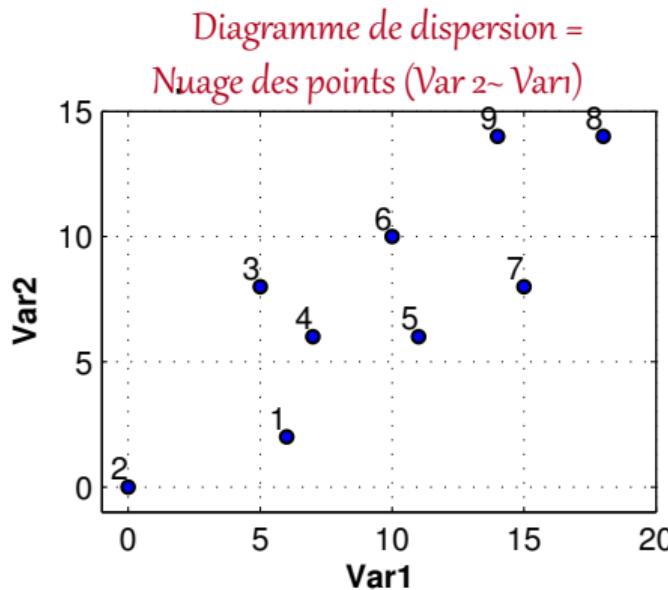
Nuage des points (Var 2~Var1)



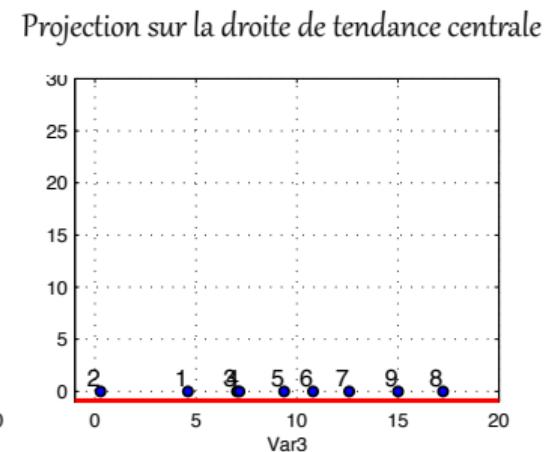
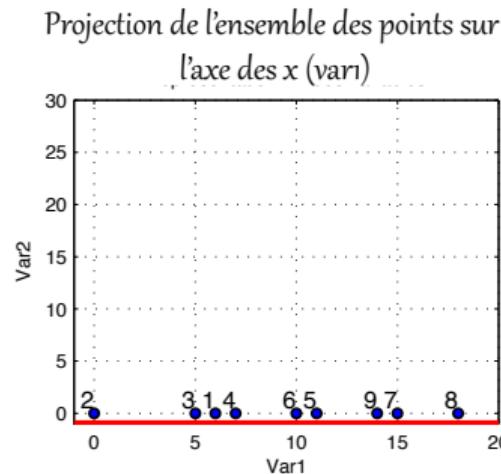
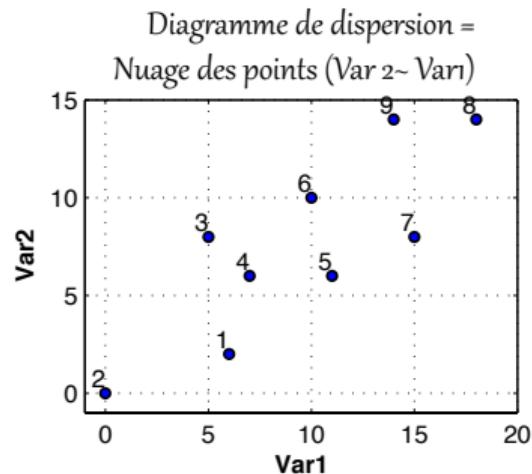
Projection de l'ensemble des points sur l'axe des x



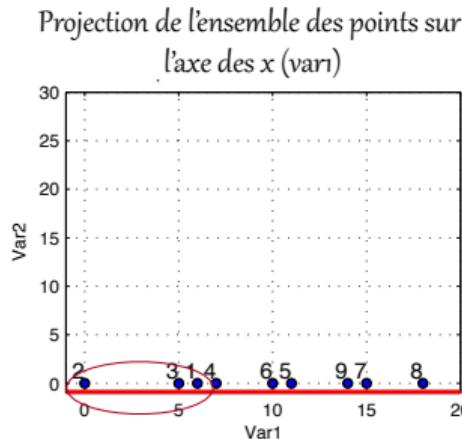
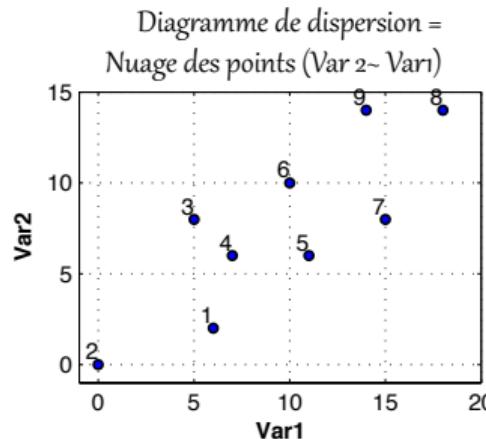
## Exemple : Illustration de l'importance du choix de la projection (3)



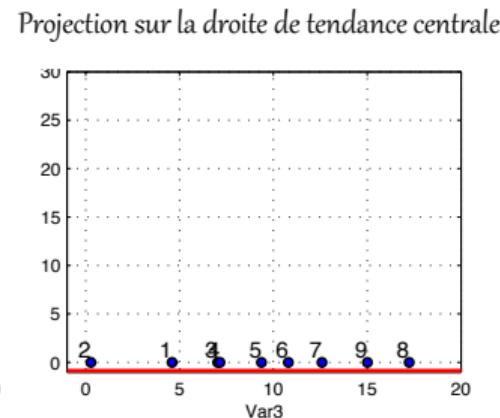
## Exemple : Illustration de l'importance du choix de la projection (4)



## Exemple : Illustration de l'importance du choix de la projection (5)



Projection n'est pas très représentative →  
Induit beaucoup de perte d'informations :  
perte du voisinage entre les individus

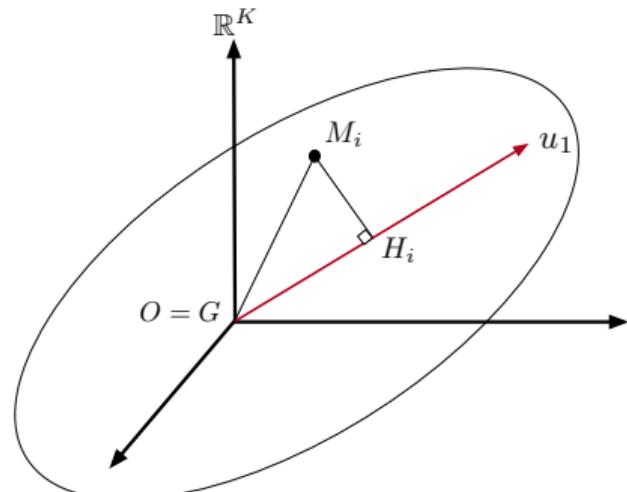


Perte d'informations est moindre  
Respect de la notion de voisinage

**La question est alors :** Quelle est le meilleur espace de projection des données ? Et comment le déterminer ?

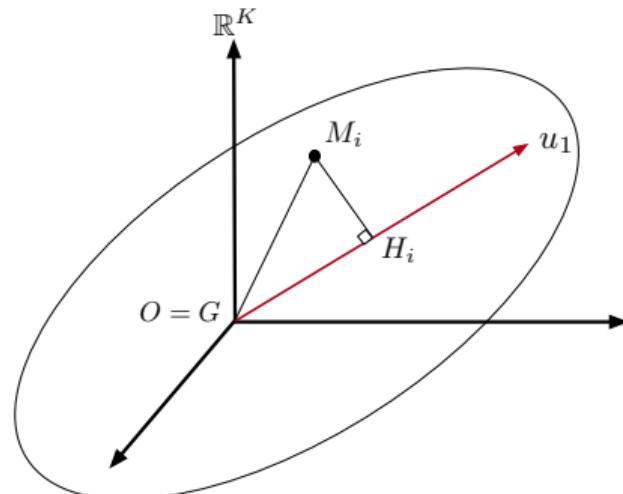
## Projection des données (2)

- On commence par un seul axe  $\Rightarrow$  on va chercher une représentation axiale des données (1D).
- Soit le point  $M_i$  dans l'espace  $\mathbb{R}^K$
- On cherche donc une direction (un axe) représentée par le vecteur unitaire  $\mathbf{u}_1$
- On va projeter le point  $M_i$  sur cet axe.



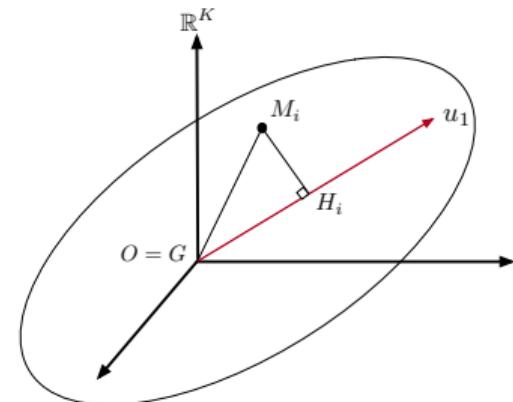
## Projection des données (3)

- On va projeter le point  $M_i$  sur cet axe. Soit  $H_i$  la coordonnée de cette projection sur l'axe  $\mathbf{u}_1$
- L'objectif est donc de trouver le vecteur  $\mathbf{u}_1$  qui permet de déformer le moins possible le nuage  $\Rightarrow OH_i$  le plus grand possible (pour s'approcher de  $OM_i$ )



## Projection des données (4)

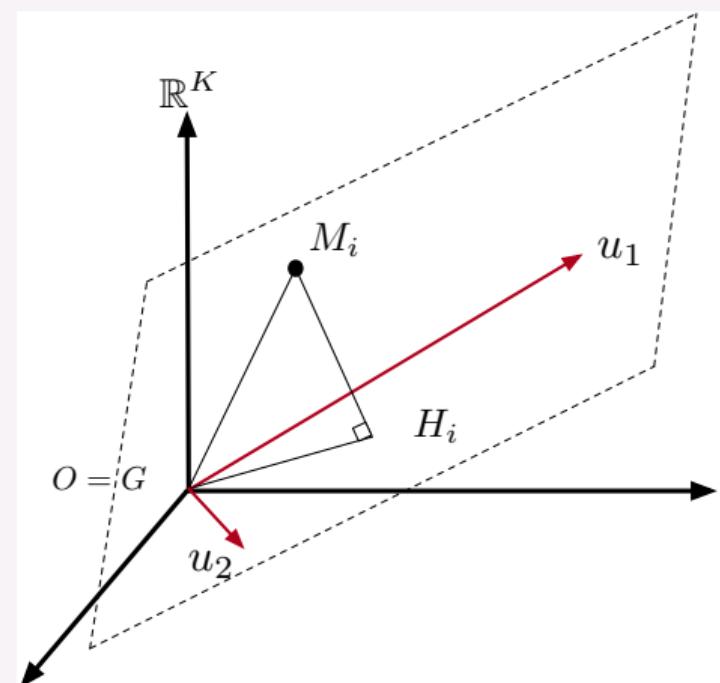
- Trouver  $u_1$  tel que :  $\sum_i OH_i^2$  est maximale
- Or  $OH_i^2$  est la distance par rapport à  $O$  qui est aussi le centre de gravité puisque les données sont centrées
  - ⇒ L'écart au point moyen
  - ⇒  $\sum_i OH_i^2$  qui est la somme des écarts au point moyen est la variance.



- **En conclusion :** on cherche une direction de variance maximum  $\Rightarrow$  de variabilité maximum = on dit aussi d'inertie maximale

## Projection des données (5)

- Maintenant, on va étendre l'analyse en 2D  $\Rightarrow$  On veut chercher la meilleure représentation plane de  $M_i$  de manière à déformer le moins possible le nuage.
- Trouver le plan tel que :  $\sum_i OH_i^2$  est maximale  $\Rightarrow$  on cherche le plan d'inertie maximum.

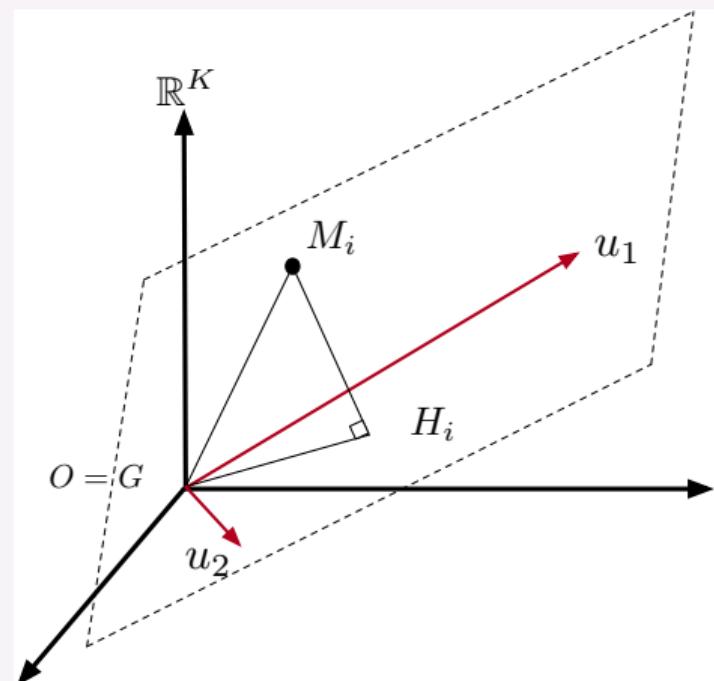


## Projection des données (6)

- Les deux solutions (1D et 2D) sont emboitées  $\implies$  Le meilleur plan contient le meilleur axe avec :

$$u_1 \perp u_2$$

- L'extension de cette analyse peut être faite sur 3, 4, ... axes
- **En conclusion :** On cherche une suite d'axes orthogonaux d'inertie maximum.



## Valeurs propres et vecteurs propres (1)

- La matrice de corrélation  $\mathbf{R}$  est une matrice carré qui a la structure suivante :

$$\mathbf{R} = \begin{pmatrix} 1 & \cdot & \cdot & \rho_{1,K} \\ \rho_{2,1} & 1 & \cdot & \rho_{2,K} \\ \cdot & \cdot & \cdot & \cdot \\ \rho_{K,1} & \cdot & \cdot & 1 \end{pmatrix}$$

- Cette matrice comprend des valeurs propres (eigenvalue) et des vecteurs propres (eigenvectors)

## Valeurs propres et vecteurs propres (2)

La solution pour déterminer les axes d'inertie maximum est la suivante :

- On montre donc que la solution de ce problème de maximisation est  $\mathbf{u}_1$ , le vecteur propre de  $\mathbf{R}$  associé à la plus grande valeur propre  $\lambda_1$ .
- Ensuite, le vecteur  $\mathbf{u}_2$  orthogonal à  $\mathbf{u}_1$  pour que l'inertie des points projetés sur cette direction soit maximale (qui correspond à la valeur propre  $\lambda_2$ ).
- Et ainsi de suite...

## Facteurs principaux et composantes principales (1)

- Plus généralement, le sous-espace à  $q$  dimensions recherché est engendré par les  $q$  premiers vecteurs propres de la matrice  $\mathbf{R}$  associés aux plus grandes valeurs propres.
- Les vecteurs  $\mathbf{u}_j$  sont appelés les **facteurs principaux**, alors que les **composantes principales** sont les variables définies par les facteurs principaux :

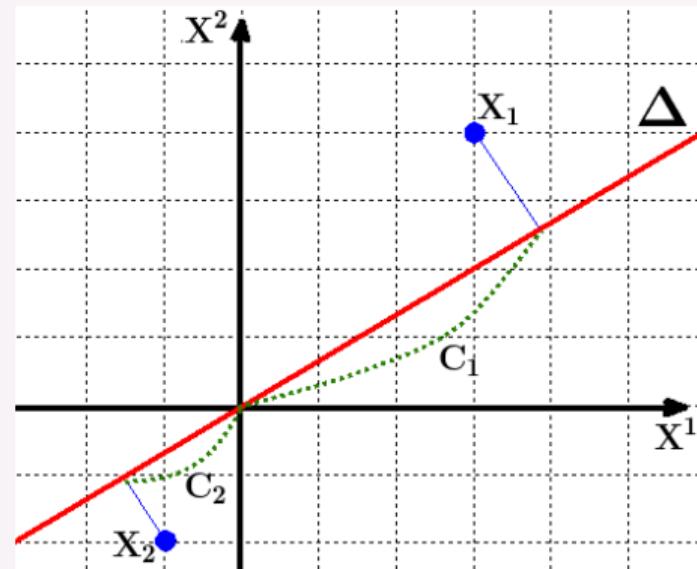
$$\mathbf{c}^j = X\mathbf{u}_j \quad (1)$$

- Les **composantes principales** contiennent les coordonnées des projections orthogonales des individus sur les axes définis par les  $\mathbf{u}_j$ .

## Facteurs principaux et composantes principales (2)

- La liste des coordonnées des individus sur  $\Delta_u$  forme une nouvelle variable artificielle  $\mathbf{c}$ .
- $\mathbf{c}$  est une combinaison linéaire des variables initiales.

$$\mathbf{c} = (c_1, \dots, c_n)' = \sum_{j=1}^K x^j u_j = X\mathbf{u}. \quad (2)$$



# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Étapes de l'ACP (1)

### ① Structuration des données :

- ◊ Soit un ensemble de  $N$  données décrites chacune par  $K$  variables. Ces données peuvent être structurées selon la matrice  $\mathbf{M}$ .
- ◊ On centre et réduit cette matrice pour obtenir la matrice de données  $\mathbf{X}$  et la matrice de covariance  $\mathbf{V}$ , laquelle est aussi égale à la matrice de corrélation  $\mathbf{R}$ .

## Étapes de l'ACP (2)

### ❷ Détermination des valeurs propres :

- ◊ Soient  $\lambda_{j \in \{1, \dots, K\}}$  les  $K$  valeurs propres de la matrice de corrélation  $\mathbf{R}$  ordonnées de telle manière que :

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$$

et leurs vecteurs propres :

$$V_{j \in \{1, \dots, K\}}$$

- ◊ Les vecteurs propres constituent les axes principaux.
- ◊ Sachant que :

$$\sum_{j \in \{1, \dots, K\}} \lambda_j = K \quad (3)$$

## Étapes de l'ACP (3)

### ③ Projection des données :

- ◊ Les vecteurs propres constituent les axes principaux  $\implies$  On peut donc projeter les données sur ces axes en faisant le produit scalaire des coordonnées d'une donnée par chacun des vecteurs propres  
(via `sklearn.decomposition.PCA`)
- ◊ L'importance d'une valeur propre par rapport aux autres peut être mesurée par l'inertie :

$$\mathcal{I} = \frac{\lambda_j}{K}$$

(via `acp.explained_variance_ratio_`)

## Étapes de l'ACP (4)

### ④ Détermination du nombre d'axes à retenir

Le nombre d'axes factoriels à retenir est souvent déterminé selon :

- Le critère de Kaiser
- La règle de Coude

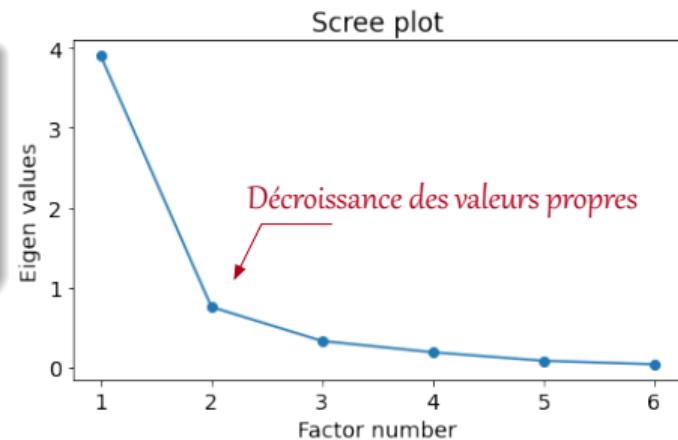
## Étapes de l'ACP : Critère de Kaiser

- La règle de Kaiser repose sur une idée simple= Le critère le plus utilisé dans la pratique.
- Elle consiste à retenir les composantes principales correspondant à des valeurs propres supérieures à 1.
- Dans une ACP normée, la somme des valeurs propres étant égale au nombre de variables, leur moyenne vaut 1. Nous considérons par conséquent qu'un axe est intéressant si sa valeur propre est supérieure 1.

## Étapes de l'ACP : Règle de coude

L'idée de la règle de coude consiste à détecter les coudes = cassures signalant un changement de structure.

(1) Propose d'étudier la courbe de décroissance des valeurs propres  $\lambda_{j \in \{1, \dots, K\}}$  et de déterminer la valeur de  $\lambda_j$  pour laquelle la décroissance diminue beaucoup plus lentement .

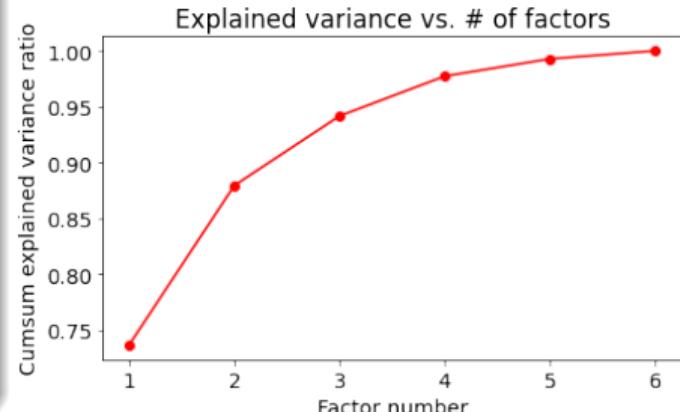


## Étapes de l'ACP : Règle de coude

L'idée de la règle de coude consiste à détecter les coudes = cassures signalant un changement de structure.

(2) Propose de s'appuyer sur les pourcentages d'inertie expliquée par les différents sous-espaces  $E_\alpha$  et à repérer l'endroit à partir duquel le pourcentage d'inertie diminue beaucoup plus lentement

$$I_\alpha = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_\alpha}{\sum_{\alpha=1}^K \lambda_\alpha}$$



## Exemple : Les données

- CYL : Cylindrée en cm<sup>3</sup>
- PUISS : La puissance en chevaux
- LONG : Longueur en cm
- LARG : Largeur en cm
- POIDS : Poids en kg
- VMAX : Vitesse maximale en km h

Modèle	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

## Exemple : Les données centrées réduites

Matrice X: Les données centrées et réduites

```
print(X.round(3))
None

[[-0.775 -0.283 -1.885 -1.097 -1.569  0.57 ]
 [-0.12    0.02   1.606  2.001  0.234  0.146]
 [-0.929 -0.839 -0.442  0.258 -0.217 -0.532]
 [-1.127 -1.293 -1.001 -1.097 -1.118 -0.617]
 [-0.128  0.676  0.256 -0.516  0.197  0.57 ]
 [-0.921 -0.132 -0.209  0.452  0.009  0.146]
 [ 0.452 -0.283  0.721  0.452  0.61   -0.363]
 [-0.183 -1.495 -0.442 -0.71   -0.517 -1.549]
 [ 2.841  2.191  0.861  1.226  1.812  1.841]
 [-1.281 -1.495 -1.606 -1.872 -1.982 -1.549]
 [-0.17   1.232 -0.256 -0.904 -0.141  1.417]
 [ 0.458 -0.132  0.535  1.033  0.61   -0.024]
 [ 1.008  1.535  1.652  0.452  2.188  0.146]
 [ 0.994  0.676  0.209  0.645  0.009  0.739]
 [-0.522 -0.233 -0.116 -0.129  0.377 -1.21 ]
 [ 0.378 -0.081  0.303 -0.323  0.121  0.57 ]
 [ 0.956  0.777  1.187  1.226  0.309  1.248]
 [-0.929 -0.839 -1.373 -1.097 -0.93   -1.549]]
```

Matrice X: vérification de la moyenne et des écarts type

```
import math
Moy =np.mean(X.round(3),axis=0)
print('Les moyennes des variables de X \n', Moy.round(3))
print('Les écarts types des variables de X\n', np.std(X,axis=0,ddof=0))

Les moyennes des variables de X
[ 0.  0.  0. -0.  0.  0.]
Les écarts types des variables de X
[1.  1.  1.  1.  1.  1.]
```

## Exemple : Analyse ACP

```
C = acp.fit_transform(X)
n_component= acp.n_components_
print(C.round(3))

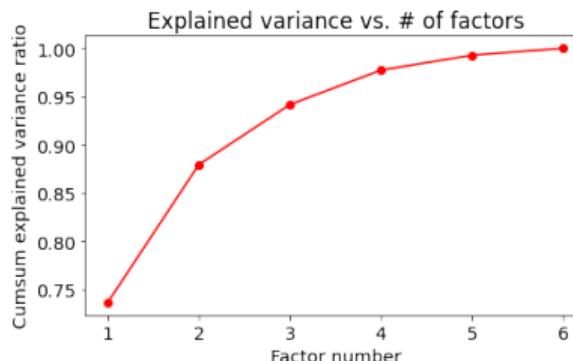
[[ -2.139  1.786  0.572  0.202 -0.301  0.054]
 [ 1.561 -1.527  1.315 -0.211  0.149 -0.327]
 [-1.119 -0.675  0.457 -0.168 -0.375  0.272]
 [-2.574  0.113  0.149 -0.017  0.227  0.263]
 [ 0.428  0.696 -0.193 -0.628  0.264 -0.037]
 [-0.304 -0.196  0.676 -0.556 -0.445  0.2  ]
 [ 0.684 -0.933 -0.257  0.203  0.209  0.154]
 [-1.948 -0.98   -0.62   0.63   0.293  0.109]
 [ 4.41   1.064 -0.594  0.847 -0.375  0.044]
 [-3.986  0.236 -0.303  0.265  0.278 -0.329]
 [ 0.438  1.912  0.025 -0.759  0.168 -0.054]
 [ 1.018 -0.842  0.217  0.303 -0.185  0.185]
 [ 2.941 -0.559 -1.244 -0.772  0.054 -0.057]
 [ 1.315  0.487  0.283  0.582 -0.067 -0.253]
 [-0.691 -0.898 -0.628 -0.358 -0.377 -0.122]
 [ 0.386  0.356 -0.076  0.103  0.527  0.339]
 [ 2.29   0.104  0.796  0.236  0.338 -0.157]
 [-2.709 -0.144 -0.574  0.096 -0.382 -0.283]]
```

*La matrice des composantes principales  
= les coordonnées des projections orthogonales  
des individus sur les axes définis par les facteurs principaux*

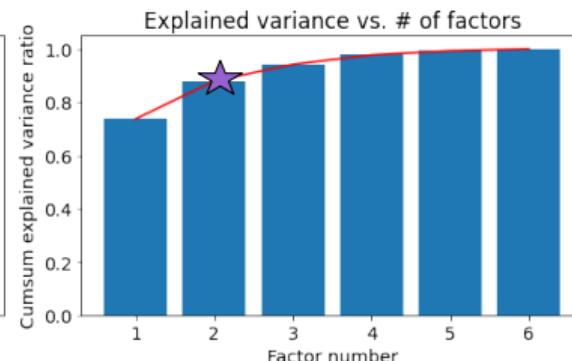
## Exemple : Règle de coude

La variance expliquée en fonction de l'axe principal

```
print(acp.explained_variance_ratio_.round(2))  
[0.74 0.14 0.06 0.04 0.02 0.01]
```



Le premier axe principal permet d'expliquer 74% de la variance



Les deux premiers axes principaux sont suffisants pour expliquer 88% de la variance totale

Apartir de deux composantes (facteurs) le pourcentage de la variance diminue plus lentement

On retient les deux premières composantes principales

# Exemple : Les composantes principales (1)

*Coordonnées des individus dans l'espace factoriel*

```
print(C.round(3))
```

```
[[ -2.139  1.786  0.572  0.202 -0.301  0.054]
 [ 1.561 -1.527  1.315 -0.211  0.149 -0.327]
 [-1.119 -0.675  0.457 -0.168 -0.375  0.272]
 [-2.574  0.113  0.149 -0.017  0.227  0.263]
 [ 0.428  0.696 -0.193 -0.628  0.264 -0.037]
 [-0.304 -0.196  0.676 -0.556 -0.445  0.2 ]
 [ 0.684 -0.933 -0.257  0.203  0.209  0.154]
 [-1.948 -0.98 -0.62  0.63  0.293  0.109]
 [ 4.41  1.064 -0.594  0.847 -0.375  0.044]
 [-3.986  0.236 -0.303  0.265  0.278 -0.329]
 [ 0.438  1.912  0.025 -0.759  0.168 -0.054]
 [ 1.018 -0.842  0.217  0.303 -0.185  0.185]
 [ 2.941 -0.559 -1.244 -0.772  0.054 -0.057]
 [ 1.315  0.487  0.283  0.582 -0.067 -0.253]
 [-0.691 -0.898 -0.628 -0.358 -0.377 -0.122]
 [ 0.386  0.356 -0.076  0.103  0.527  0.339]
 [ 2.29  0.104  0.796  0.236  0.338 -0.157]
 [-2.709 -0.144 -0.574  0.096 -0.382 -0.283]]
```

Modèle	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Alfasud TI	-2,139	1,786	0,572	0,202	-0,301	0,054
Audi 100	1,561	-1,527	1,315	-0,211	0,149	-0,327
Simca 1300	-1,119	-0,675	0,457	-0,168	-0,375	0,272
Citroen GS Clu	-2,574	0,113	0,149	-0,017	0,227	0,263
Fiat 132	0,428	0,696	-0,193	-0,628	0,264	-0,037
Lancia Beta	-0,304	-0,196	0,676	-0,556	-0,445	0,200
Peugeot 504	0,684	-0,933	-0,257	0,203	0,209	0,154
Renault 16 TL	-1,948	-0,980	-0,620	0,630	0,293	0,109
Renault 30	4,410	1,064	-0,594	0,847	-0,375	0,044
Toyota Coroll	-3,986	0,236	-0,303	0,265	0,278	-0,329
Alfetta 1.66	0,438	1,912	0,025	-0,759	0,168	-0,054
Princess 1800	1,018	-0,842	0,217	0,303	-0,185	0,185
Datsun 200L	2,941	-0,559	-1,244	-0,772	0,054	-0,057
Taunus 2000	1,315	0,487	0,283	0,582	-0,067	-0,253
Rancho	-0,691	-0,898	-0,628	-0,358	-0,377	-0,122
Mazda 9295	0,386	0,356	-0,076	0,103	0,527	0,339
Opel Rekord	2,290	0,104	0,796	0,236	0,338	-0,157
Lada 1300	-2,709	-0,144	-0,574	0,096	-0,382	-0,283

## Exemple : Les composantes principales (2)

*Coordonnées des individus sur les deux premiers axes principaux*

```
print(C.round(3))
```

```
[[ -2.139  1.786  0.572  0.202 -0.301  0.054]
 [ 1.561 -1.527  1.315 -0.211  0.149 -0.327]
 [-1.119 -0.675  0.457 -0.168 -0.375  0.272]
 [-2.574  0.113  0.149 -0.017  0.227  0.263]
 [ 0.428  0.696 -0.193 -0.628  0.264 -0.037]
 [-0.304 -0.196  0.676 -0.556 -0.445  0.2 ]
 [ 0.684 -0.933 -0.257  0.203  0.209  0.154]
 [-1.948 -0.98 -0.62  0.63  0.293  0.109]
 [ 4.41  1.064 -0.594  0.847 -0.375  0.044]
 [-3.986  0.236 -0.303  0.265  0.278 -0.329]
 [ 0.438  1.912  0.025 -0.759  0.168 -0.054]
 [ 1.018 -0.842  0.217  0.303 -0.185  0.185]
 [ 2.941 -0.559 -1.244 -0.772  0.054 -0.057]
 [ 1.315  0.487  0.283  0.582 -0.067 -0.253]
 [-0.691 -0.898 -0.628 -0.358 -0.377 -0.122]
 [ 0.386  0.356 -0.076  0.103  0.527  0.339]
 [ 2.29  0.104  0.796  0.236  0.338 -0.157]
 [-2.709 -0.144 -0.574  0.096 -0.382 -0.283]]
```

Modèle	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Alfasud TI	-2,139	1,786	0,572	0,202	-0,301	0,054
Audi 100	1,561	-1,527	1,315	-0,211	0,149	-0,327
Simca 1300	-1,119	-0,675	0,457	-0,168	-0,375	0,272
Citroen GS Clu	-2,574	0,113	0,149	-0,017	0,227	0,263
Fiat 132	0,428	0,696	-0,193	-0,628	0,264	-0,037
Lancia Beta	-0,304	-0,196	0,676	-0,556	-0,445	0,200
Peugeot 504	0,684	-0,933	-0,257	0,203	0,209	0,154
Renault 16 TL	-1,948	-0,980	-0,620	0,630	0,293	0,109
Renault 30	4,410	1,064	-0,594	0,847	-0,375	0,044
Toyota Coroll	-3,986	0,236	-0,303	0,265	0,278	-0,329
Alfetta 1.66	0,438	1,912	0,025	-0,759	0,168	-0,054
Princess 1800	1,018	-0,842	0,217	0,303	-0,185	0,185
Datsun 200L	2,941	-0,559	-1,244	-0,772	0,054	-0,057
Taunus 2000	1,315	0,487	0,283	0,582	-0,067	-0,253
Rancho	-0,691	-0,898	-0,628	-0,358	-0,377	-0,122
Mazda 9295	0,386	0,356	-0,076	0,103	0,527	0,339
Opel Rekord	2,290	0,104	0,796	0,236	0,338	-0,157
Lada 1300	-2,709	-0,144	-0,574	0,096	-0,382	-0,283

## Exercice : Étape 2 - Analyse ACP

- ❶ Calculez la matrice de corrélation des variables.
- ❷ Calculez les valeurs propres de la matrice de corrélation.
- ❸ Calculez la somme des valeurs propres de la matrice de corrélation.
- ❹ Calculez la proportion de variance expliquée.
- ❺ Déterminez le nombre d'axe à retenir en utilisant le critère de coude.
- ❻ Réalisez une ACP.

# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Représentation des individus

- Coordonnées factorielles
- Qualité globale de la représentation
- Contribution d'un individu sur un axe ( $Ctr$ )
- Qualité de représentation d'un individu sur un axe ( $Qlt$ )

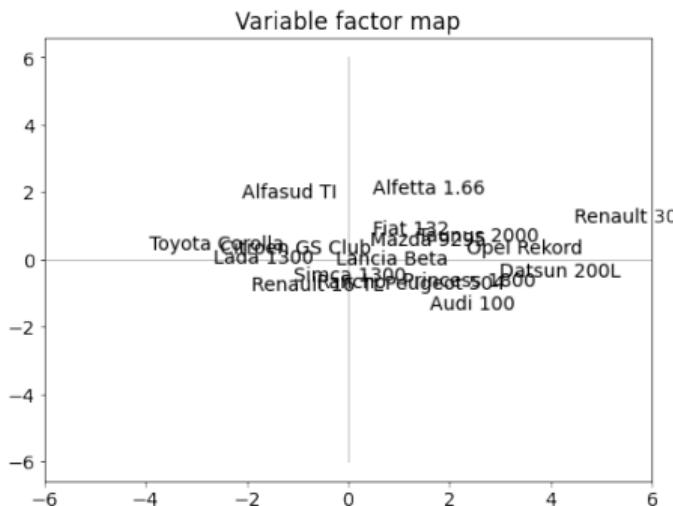
## Coordonnées factoriels

- Les  **coordonnées factoriels = composantes principales** contiennent les coordonnées des projections orthogonales des individus sur les axes définis par les  $\mathbf{u}_j$ .
- Les vecteurs  $\mathbf{u}_j = \text{facteurs principaux}$  (`acp.components_`)

• Par exemple :  $\mathbf{u}_1 = \begin{pmatrix} 0.424 \\ 0.421 \\ 0.421 \\ 0.386 \\ 0.430 \\ 0.358 \end{pmatrix}$  et  $\mathbf{u}_2 = \begin{pmatrix} 0.124 \\ 0.415 \\ -0.411 \\ -0.446 \\ -0.242 \\ 0.619 \end{pmatrix}$

## Exemple : Projection des individus sur les axes principaux (1)

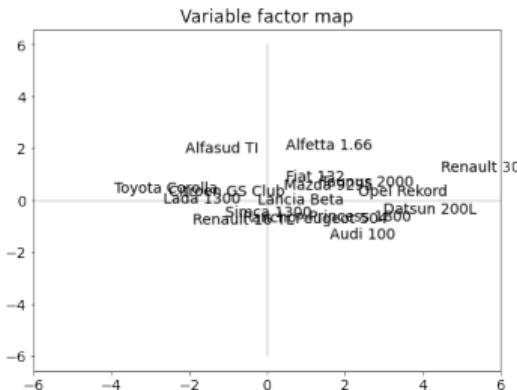
*Coordonnées des individus sur les deux premiers axes principaux*



Modèle	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Alfasud TI	-2,139	1,786	0,572	0,202	-0,301	0,054
Audi 100	1,561	-1,527	1,315	-0,211	0,149	-0,327
Simca 1300	-1,119	-0,675	0,457	-0,168	-0,375	0,272
Citroen GS Club	-2,574	0,113	0,149	-0,017	0,227	0,263
Fiat 132	0,428	0,696	-0,193	-0,628	0,264	-0,037
Lancia Beta	-0,304	-0,196	0,676	-0,556	-0,445	0,200
Peugeot 504	0,684	-0,933	-0,257	0,203	0,209	0,154
Renault 16 TL	-1,948	-0,980	-0,620	0,630	0,293	0,109
Renault 30	4,410	1,064	-0,594	0,847	-0,375	0,044
Toyota Corolla	-3,986	0,236	-0,303	0,265	0,278	-0,329
Alfetta 1.66	0,438	1,912	0,025	-0,759	0,168	-0,054
Princess 1800	1,018	-0,842	0,217	0,303	-0,185	0,185
Datsun 200L	2,941	-0,559	-1,244	-0,772	0,054	-0,057
Taunus 2000	1,315	0,487	0,283	0,582	-0,067	-0,253
Rancho	-0,691	-0,898	-0,628	-0,358	-0,377	-0,122
Mazda 929	0,386	0,356	-0,076	0,103	0,527	0,339
Opel Rekord	2,290	0,104	0,796	0,236	0,338	-0,157
Lada 1300	-2,709	-0,144	-0,574	0,096	-0,382	-0,283

## Exemple : Projection des individus sur les axes principaux (2)

*Coordonnées des individus sur les deux premiers axes principaux*



Modèle	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6
Alfasud TI	-2,139	1,786	0,572	0,202	-0,301	0,054
Audi 100	1,561	-1,527	1,315	-0,211	0,149	-0,327
Simca 1300	-1,119	-0,675	0,457	-0,168	-0,375	0,272
Citroën GS Club	-2,574	0,113	0,149	-0,017	0,227	0,263
Fiat 132	0,428	0,696	-0,193	-0,628	0,264	-0,037
Lancia Beta	-0,304	-0,196	0,676	-0,556	-0,445	0,200
Peugeot 504	0,684	-0,933	-0,257	0,203	0,209	0,154
Renault 16 TL	-1,948	-0,980	-0,620	0,630	0,293	0,109
Renault 30	4,410	1,064	-0,594	0,847	-0,375	0,044
Toyota Corolla	-3,986	0,236	-0,303	0,265	0,278	-0,329
Alfetta 1.66	0,438	1,912	0,025	-0,759	0,168	-0,054
Princess 1800	1,018	-0,842	0,217	0,303	-0,185	0,185
Datsun 200L	2,941	-0,559	-1,244	-0,772	0,054	-0,057
Taunus 2000	1,315	0,487	0,283	0,582	-0,067	-0,253
Ranchero	-0,691	-0,898	-0,628	-0,358	-0,377	-0,122
Mazda 9295	0,386	0,356	-0,076	0,103	0,527	0,339
Opel Rekord	2,290	0,104	0,796	0,236	0,338	-0,157
Lada 1300	-2,709	-0,144	-0,574	0,096	-0,382	-0,283

- Projection des différents modèles de voitures sur les deux premiers axes factoriels = plan factoriel
- Analyse des proximités entre les différents modèles

## Qualité globale de la représentation

- La qualité globale de la représentation que permet le sous-espace  $E_k$  est mesurée par le pourcentage d'inertie  $I_k$  :

$$I_k = \frac{\lambda_1 + \lambda_2 + \dots + \lambda_k}{\sum_{\alpha=1}^p \lambda_\alpha} \times 100 \quad (4)$$

- Dans notre cas :  $\lambda_1 = 4.421$ ,  $\lambda_2 = 0.856$ ,  $\lambda_3 = 0.373$ ,  $\lambda_4 = 0.214$ ,  $\lambda_5 = 0.093$ ,  $\lambda_6 = 0.043$
- Donc la qualité globale du plan factoriel  $I_2 = 87.9\% \approx 88\% \implies$  La qualité globale est préservée à 88%

## Contribution d'un individu sur un axe

- La contribution relative d'un individu  $i$  à la formation de l'axe factoriel  $\alpha$  est égale à l'inertie relative de cet individu sur l'axe factoriel  $i$ .
- Elle est défini par :

$$Ctr(i, \alpha) = \frac{(c_i^\alpha)^2}{\lambda_\alpha} = \frac{(\text{Score de } i \text{ sur l'axe } \alpha)^2}{\lambda_\alpha} \quad (5)$$

## Exemple : Contribution des individus (1)

```
C = acp.fit_transform  
print(C.round(3))
```

```
[[ -2.139  1.786  
[  1.561 -1.527  
[-1.119 -0.675  
[-2.574  0.113  
[  0.428  0.696  
[-0.304 -0.196  
[  0.684 -0.933  
[-1.948 -0.98  
[  4.41    1.064  
[-3.986  0.236  
[  0.438  1.912  
[  1.018 -0.842  
[  2.941 -0.559  
[  1.315  0.487  
[-0.691 -0.898  
[  0.386  0.356  
[  2.28    0.104
```

```
eigval, vectors = np.linalg.eig(corr_matrix )  
print('Les valeurs propres sont:', '\n', eigval)
```

Les valeurs propres sont:  
[4.421 0.856 0.373 0.214 0.093 0.043]

$$Ctr(i, \alpha) = \frac{(c_{\alpha}^i)^2}{\lambda_{\alpha}} = \frac{(\text{Score de } i \text{ sur l'axe } \alpha)^2}{N\lambda_{\alpha}}$$

$$Ctr(1, 1) = \frac{(c_1^1)^2}{\lambda_1} = \frac{(-2, 139)^2}{18 \times 4.42}$$

## Exemple : Contribution des individus (2)

		<b>id</b>	<b>CTR_1</b>	<b>CTR_2</b>
0		Alfasud TI	0.057493	0.206933
1		Audi 100	0.030640	0.151329
2		Simca 1300	0.015746	0.029525
3	Citroen GS Club		0.083244	0.000827
4		Fiat 132	0.002300	0.031398
5		Lancia Beta	0.001163	0.002497
6		Peugeot 504	0.005878	0.056499
7		Renault 16 TL	0.047711	0.062384
8		Renault 30	0.2444369	0.073419
9	Toyota Corolla		0.199640	0.003622
10		Alfetta 1.66	0.002407	0.237357
11		Princess 1800	0.013028	0.045978
12		Datsun 200L	0.108701	0.020292
13		Taunus 2000	0.021727	0.015361
14		Rancho	0.006002	0.052300
15		Mazda 9295	0.001870	0.008233
16		Opel Rekord	0.065888	0.000707
17		Lada 1300	0.092194	0.001340

$$Ctr(1, 1) = \frac{(c_1^1)^2}{\lambda_1} = \frac{(-0,42)^2}{18 \times 4.42} = 0.057$$

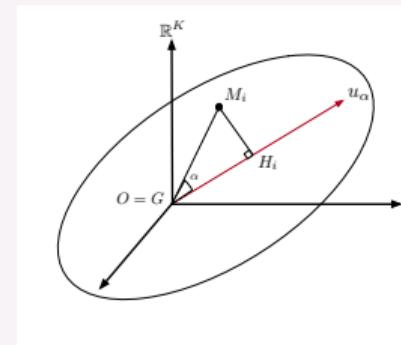
La Renault 30 et la Toyota sont déterminants pour le 1er axe

L'Alfetta-1.66 et l'Alfasud et l'Audi 100 sont déterminants pour le 2eme axe

## Qualité de représentation d'un individu sur un axe (1)

- L'inertie totale  $I$  du nuage de points est donnée par :

$$I = \frac{1}{N} \sum_{i=1}^N d^2(i, G) \quad (6)$$



- La contribution d'un individu  $i$  dans l'inertie totale est donnée par :

$$d_i^2 = \sum_{j=1}^K (x_i^j)^2$$

## Qualité de représentation d'un individu sur un axe (2)

- Le rapport entre ces deux quantités est appelé **Qualité de représentation d'un individu  $i$  sur un axe  $\alpha$** .
- Ce rapport est noté  $Qlt(i, \alpha)$  :

$$Qlt(i, \alpha) = \frac{(c_i^\alpha)^2}{d_i^2}$$

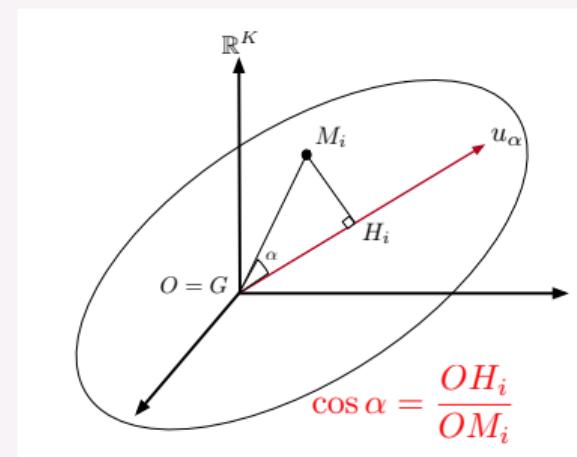
- Si  $Qlt(i, \alpha)$  est proche de 1, l'individu est bien représenté par cet axe
- Si  $Qlt(i, \alpha)$  est proche de 0, l'individu est très mal représenté par cet axe.

## Qualité de représentation d'un individu sur un axe (3)

- La **Qualité de représentation d'un individu  $i$  sur un axe  $\alpha$** , qui est noté  $Qlt(i, \alpha)$  :

$$Qlt(i, \alpha) = \frac{(c_i^\alpha)^2}{d_i^2} = \cos^2(i, \alpha)$$

- Si  $\cos^2(i, \alpha)$  est proche de 1,  
 l'individu est bien représenté par l'axe
- Si  $\cos^2(i, \alpha)$  est proche de 0,  
 l'individu est mal représenté par l'axe.



# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Représentation des variables

- Coefficients de corrélation linéaire  $r(c, x^j)$
- Cercle de corrélation
- Qualité de représentation des variables
- Contribution des variables à un axe

## Coefficients de corrélation linéaire $r(c, x^j)$

- Le calcul des coefficients de corrélation linéaire  $r(c, x^j)$  entre les composantes principales et les variables initiales permet d'interpréter l'importance de leur relation.
- On s'intéresse aux coefficients les plus forts en valeur absolue et près de 1.
- Dans le cas de données centrées réduites, le calcul de  $r(c, x^j)$  est particulièrement simple. On montre que :

$$r(c, x^j) = \sqrt{\lambda} u_i \quad (7)$$

## Cercle de corrélation

- Pour un couple de composantes principales  $(c^1, c^2)$ , ces corrélations sont souvent représentées sur un graphique appelé « cercle de corrélation » sur lequel chaque variable  $x^j$  est repérée par une abscisse  $r(c^1, x^j)$  et une ordonnée  $r(c^2, x^j)$ .
- Le cercle de corrélation permet de détecter les éventuels groupes de variables qui se ressemblent ou, au contraire, qui s'opposent.
- Les variables les plus intéressantes sont généralement celles qui sont assez proches de l'un des axes et qui sont assez loin de l'origine.

## Qualité de représentation des variables

- On peut calculer la qualité de représentation de la variable  $j$  par l'axe  $k$  en montant la corrélation au carré :

$$Qlt(j, k) = \cos^2 = r^2(c, x^j)$$

## Contribution des variables à un axe

- La contribution de la variable  $j$  à l'inertie de l'axe  $k$  est également basée sur le carré de la corrélation, mais relativisée par l'importance de l'axe :

$$Ctr(j, k) = \frac{r^2(c, x^j)}{\lambda_k}$$

## Exemple : Coefficients de corrélation linéaire $r(c_i, x^j)$

Les valeurs propres sont:

[4.421 0.856 0.373 0.214 0.093 0.043]

```
# Vecteurs propres à partir de l'acp
acp.components_.round(3)
```

```
array([[ 0.425,  0.422,  0.421,  0.387,  0.431,  0.359],
       [ 0.124,  0.416, -0.412, -0.446, -0.243,  0.62 ],
      [-0.354, -0.185,  0.068,  0.605, -0.484,  0.485],
      [ 0.808, -0.358, -0.28 ,  0.212, -0.302, -0.074],
      [ 0.152, -0.294,  0.731, -0.478, -0.305,  0.189],
      [-0.059, -0.633, -0.19 , -0.11 ,  0.581,  0.459]])
```

$$r(c_1^1, x^1) = \sqrt{\lambda_1} u_1^1$$

$$r(c_1^1, x^1) = \sqrt{4,42} \times 0,425$$

$$r(c_1^1, x^1) = 0,893$$

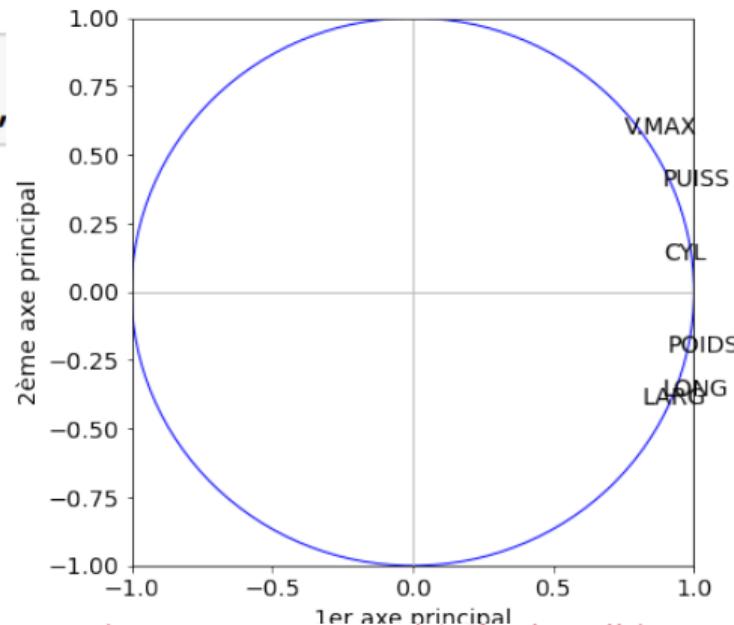
```
#Affichage des corrélations sur tous les axes
print(pd.DataFrame({'id':M.columns,'COR_1':corvar[:,0],'COR_2':corvar[:,1],
```

	id	COR_1	COR_2	COR_3	COR_4	COR_5	COR_6
0	CYL	0.893464	0.114906	-0.215983	0.373615	0.046176	-0.012254
1	PUISS	0.886858	0.384689	-0.112948	-0.165485	-0.089481	-0.131711
2	LONG	0.886155	-0.381029	0.041310	-0.129390	0.222555	-0.039593
3	LARG	0.813536	-0.412736	0.369448	0.097854	-0.145672	-0.022797
4	POIDS	0.905187	-0.224532	-0.295865	-0.139547	-0.092779	0.120846

## Exemple : Cercle de corrélation

```
#Affichage des deux premiers axes
print(pd.DataFrame({'id':M.columns,
```

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



- Toutes les variables sont fortement corrélées (positivement) au 1er axe principal (effet de taille)
- Sur le 2ème axe: les variables de tailles (Poids, Long, Larg) s'opposent aux variables V.Max, Puis et Cyl

## Exercice : Étape 3 - Interprétation des résultats de l'ACP (1)

---

- ① Représentez les observations dans le premier plan factoriel.
- ② Déterminez les deux premiers axes factoriels.
- ③ Évaluez la qualité globale de la représentation.
- ④ Calculez la contribution de chaque individu à l'inertie totale.
- ⑤ Évaluez la qualité de représentation de chaque individu sur les deux premiers axes.

## Exercice : Étape 3 - Interprétation des résultats de l'ACP (2)

---

- ⑥ Représentez le cercle de corrélation des variables.
- ⑦ Évaluez la qualité de représentation des variables.
- ⑧ Calculez la contribution de chaque variable aux deux premiers axes factoriels

# Plan du cours

- 1 Introduction
- 2 Les données et leurs caractéristiques
- 3 Principe de l'ACP
  - Projection des données
  - Étapes de l'ACP
- 4 Interprétation des résultats de l'ACP
  - Représentation des individus
  - Représentation des variables
  - Projection d'individus supplémentaires

## Projection d'individus supplémentaires

- La projection des individus supplémentaires est une analyse importante de l'analyse en composantes principales.
- L'objectif est de positionner de nouveau individus par rapport à ceux (les individus actifs, l'échantillon d'apprentissage, l'échantillon de référence) qui ont contribué à la construction de l'espace factoriel.

## Exemple : Projection d'individus supplémentaires

Peugeot 604	2664	136	472	177	1410	180
Peugeot 304	1288	74	414	157	915	160