

# MA431 : Analyse de données et réduction de dimension

## Analyse en Composantes Principales : une première approche

D. Barcelo

Grenoble INP ESISAR

2022/2023

- 1 Introduction
- 2 Analyse en composantes principales
- 3 Illustrations
- 4 Notions utiles pour l'ACP
- 5 Technique de l'ACP
- 6 Inertie dans  $\mathbb{R}^p$
- 7 Projection sur les axes principaux
- 8 Qualité de la projection
- 9 Exemple d'ACP
- 10 Nuage des variables
- 11 Pour aller plus loin

# Introduction

Problème :



Image : <https://www.scnsoft.de/blog/big-data-probleme-untersuchen>

# Introduction

Trop de données !

- Trop d'individus ( $n$ ),
- Trop de variables ( $p$ ),
- Impossibilité de visualiser,
- Temps de traitement trop important.

Comment faire pour traiter ces données sans perdre trop d'informations ?



On va réduire le nombre de dimensions (le nombre de variables) !

# Avantages

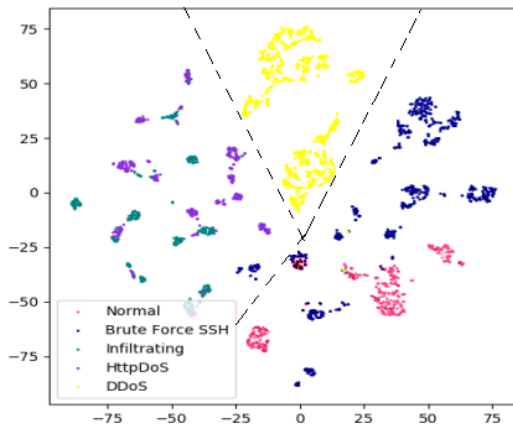



Image :   
[https://www.researchgate.net/figure/Visualization-of-unsupervised-intrusion-detection-algorithms-deep-learning-algorithm-b-Deep\\_fig4\\_336756530](https://www.researchgate.net/figure/Visualization-of-unsupervised-intrusion-detection-algorithms-deep-learning-algorithm-b-Deep_fig4_336756530)

# Avantages

Réduire le nombre de dimensions permet de :

- compresser les données pour diminuer l'espace de stockage,
- réduire le temps de calculs avec des données moins volumineuses,
- supprimer une partie du bruit présent dans les données initiales,
- échapper au fléau de la dimensionnalité !

# the curse of dimensionality

## La malédiction de la dimensionnalité ! (Bellman, 1961)

- Augmentation du nombre de dimensions : Croissance rapide du volume de l'espace occupé par les données.
- Croissance **exponentielle** de la quantité de données nécessaires pour un volume en grandes dimensions.
- Données **isolées** dans des espaces de très grandes dimensions.
- Données **moins statistiquement représentatives**.
- Solution : réduire la dimension pour passer de quelques millions de dimensions à seulement quelques centaines.



# Une solution

Une solution possible est :

l'Analyse en Composantes Principales (ACP)  
the Principal Component Analysis (PCA)

# Analyse en composantes principales ?

- Méthode de **réduction de dimension**
- Technique d'analyse de données basée sur **l'algèbre linéaire**
- Méthode **factorielle** de statistiques multidimensionnelles
- Plus précisément :
- Autrement dit :
- Méthode présentée : utiliser les **éléments propres** d'une matrice
- Méthode fréquemment utilisée en pratique : La décomposition en valeurs singulières d'une matrice (et non en éléments propres)

# Analyse en composantes principales ?

- Méthode de **réduction de dimension**
- Technique d'analyse de données basée sur **l'algèbre linéaire**
- Méthode **factorielle** de statistiques multidimensionnelles
- Plus précisément : **diminuer le nombre de dimensions en minimisant la perte d'information,**
- Autrement dit :
- Méthode présentée : utiliser les **éléments propres** d'une matrice
- Méthode fréquemment utilisée en pratique : La décomposition en valeurs singulières d'une matrice (et non en éléments propres)

# Analyse en composantes principales ?

- Méthode de **réduction de dimension**
- Technique d'analyse de données basée sur **l'algèbre linéaire**
- Méthode **factorielle** de statistiques multidimensionnelles
- Plus précisément : **diminuer le nombre de dimensions en minimisant la perte d'information**,
- Autrement dit : compresser les données en les déformant le moins possible.
- Méthode présentée : utiliser les **éléments propres** d'une matrice
- Méthode fréquemment utilisée en pratique : La décomposition en valeurs singulières d'une matrice (et non en éléments propres)

## ACP

- **ACP : Analyse en Composantes Principales**
- Méthode de **description et d'interprétation** de relations entre variables quantitatives,
- Objectif : remplacer les variables par des **combinaisons linéaires des variables** qui seraient plus représentatives des données,
- Avantage : obtenir une représentation graphique des relations entre  $p$  variables observées sur  $n$  individus en gardant les deux **variables les plus représentatives**.
- Historiquement : créée par Karl Pearson en 1901 mais véritablement utilisée que depuis les années 1970
- Depuis la fin des années 1990, regain d'intérêt pour la méthode, notamment dans la détection d'intrusion ou d'erreurs et surtout pour réduire la taille des données sans trop perdre de sens.

# Exemple de données en lien avec la sécurité

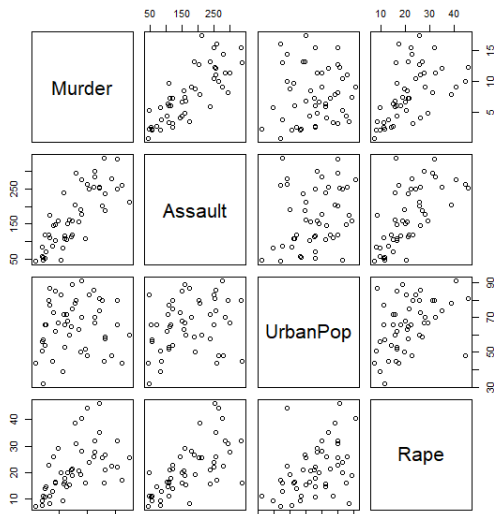
|               | Murder | Assault | UrbanPop | Rape |
|---------------|--------|---------|----------|------|
| Alabama       | 13.2   | 236     | 58       | 21.2 |
| Alaska        | 10.0   | 263     | 48       | 44.5 |
| Arizona       | 8.1    | 294     | 80       | 31.0 |
| Arkansas      | 8.8    | 190     | 50       | 19.5 |
| California    | 9.0    | 276     | 91       | 40.6 |
| Colorado      | 7.9    | 204     | 78       | 38.7 |
| Connecticut   | 3.3    | 110     | 77       | 11.1 |
| Delaware      | 5.9    | 238     | 72       | 15.8 |
| Florida       | 15.4   | 335     | 80       | 31.9 |
| Georgia       | 17.4   | 211     | 60       | 25.8 |
| Hawaii        | 5.3    | 46      | 83       | 20.2 |
| Idaho         | 2.6    | 120     | 54       | 14.2 |
| Illinois      | 10.4   | 249     | 83       | 24.0 |
| Indiana       | 7.2    | 113     | 65       | 21.0 |
| Iowa          | 2.2    | 56      | 57       | 11.3 |
| Kansas        | 6.0    | 115     | 66       | 18.0 |
| Kentucky      | 9.7    | 109     | 52       | 16.3 |
| Louisiana     | 15.4   | 249     | 66       | 22.2 |
| Maine         | 2.1    | 83      | 51       | 7.8  |
| Maryland      | 11.3   | 300     | 67       | 27.8 |
| Massachusetts | 4.4    | 149     | 85       | 16.3 |
| Michigan      | 12.1   | 255     | 74       | 35.1 |
| Minnesota     | 2.7    | 72      | 66       | 14.9 |

Showing 1 to 23 of 50 entries, 4 total columns

Comment représenter les données ? Etudier les liens entre les individus  
entre les variables ? Définir des profils d'individus ?

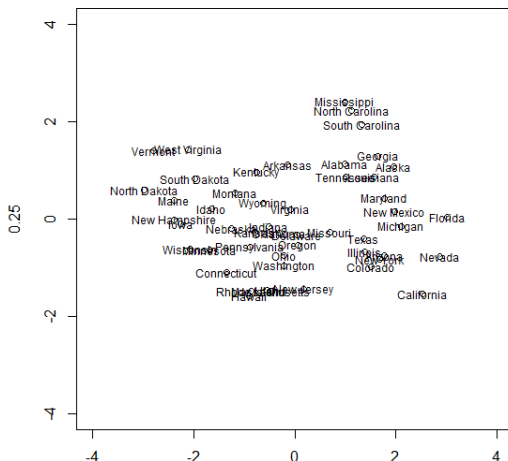
# Représentation graphique

On peut utiliser une représentation séparée :



# Représentation graphique

mais l'idéal serait d'avoir une représentation du type :





# Représentation graphique

Comment représenter un nuage de  $n$  individus caractérisés par  $p$  variables ?

# Représentation graphique

Comment représenter un nuage de  $n$  individus caractérisés par  $p$  variables ?

- 1  $n$  vecteurs de dimensions  $p$ .
- 2 Projection sur un espace de dimension 2.
- 3 Trouver **le plan de projection qui déforme le moins possible le nuage**.
- 4 Munir l'espace de dimension  $p$  d'une distance.
- 5 Déterminer un indicateur de la forme globale du nuage.

# Représentation graphique

Quel plan de projection choisir ?

# Représentation graphique

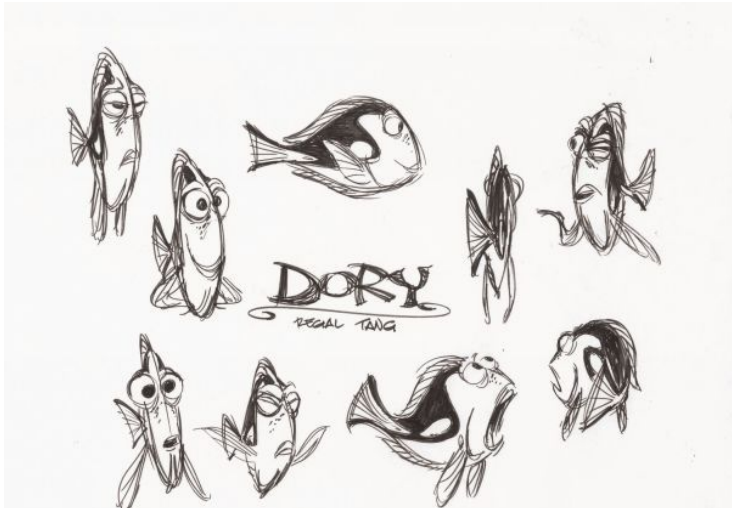
Quel plan de projection choisir ?

On pense à un poisson (exemple le plus fréquent)

# Représentation graphique

Quel plan de projection choisir ?

On pense à un poisson (exemple le plus fréquent)



# Les espaces

## Représentation graphique

Représenter graphiquement les individus (observations) :

# Les espaces

## Représentation graphique

Représenter graphiquement les individus (observations) : représenter  $n$  vecteurs dans un espace à  $p$  dimensions (les variables observées).

# Les espaces

## Représentation graphique

Représenter graphiquement les individus (observations) : représenter  $n$  vecteurs dans un espace à  $p$  dimensions (les variables observées).

On parle de **nuage des individus** dans l'**espace des variables** (noté  $E$ ).



# Les espaces

- On observe  $p$  variables **quantitatives**  $\{X_i\}_{i \in \llbracket 1;p \rrbracket}$  sur  $n$  individus.
- $E$  est munis du produit scalaire usuel et de la distance euclidienne associée.
- $\vec{x}$  et  $\vec{y}$  sont des vecteurs de  $E$  de représentation matricielle  $X$  et  $Y$ , on a :  $\vec{x} \cdot \vec{y} = {}^t X Y$ .
- Un individu correspond à un vecteur de  $E$ , donc un vecteur à  $p$  coordonnées.

# Présentation des données

On observe  $p$  variables sur  $n$  individus.

On étudie le nuage des individus dans l'espace des variables.

On note  $X$  la matrice des données et  $X_0$  la matrice des données centrées.

$$X = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ x_{2,1} & \dots & x_{2,p} \\ \dots & \dots & \dots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \text{ et } X_0 = \begin{pmatrix} x_{1,1} - \bar{x}_1 & \dots & x_{1,p} - \bar{x}_p \\ x_{2,1} - \bar{x}_1 & \dots & x_{2,p} - \bar{x}_p \\ \dots & \dots & \dots \\ x_{n,1} - \bar{x}_1 & \dots & x_{n,p} - \bar{x}_p \end{pmatrix}.$$

# Matrice de variance-covariance

## Matrice de Variance-Covariance

Soit  $X_0$  la matrice des données centrées. On pose  $V = \frac{1}{n} X_0^t \times X_0$ . On a alors :

$$V = \begin{pmatrix} \sigma_{x_1}^2 & \text{cov}(x_1; x_2) & \dots & \text{cov}(x_1; x_p) \\ \text{cov}(x_2; x_1) & \sigma_{x_2}^2 & \dots & \text{cov}(x_2; x_p) \\ \dots & \dots & \dots & \dots \\ \text{cov}(x_p; x_1) & \text{cov}(x_p; x_2) & \dots & \sigma_{x_p}^2 \end{pmatrix}.$$

# Inertie du nuage

## Inertie

On appelle **Inertie totale** d'un nuage de points la moyenne des carrés des distances euclidiennes des vecteurs au centre de gravité du nuage (noté  $\vec{\bar{X}}$ ).

$$\mathcal{I} = \frac{1}{n} \sum_{k=1}^n d^2(\vec{x}_k; \vec{\bar{X}})$$

# Inertie du nuage

## Inertie

On appelle **Inertie totale** d'un nuage de points la moyenne des carrés des distances euclidiennes des vecteurs au centre de gravité du nuage (noté  $\vec{\bar{x}}$ ).

$$\mathcal{I} = \frac{1}{n} \sum_{k=1}^n d^2(\vec{x}_k; \vec{\bar{x}})$$

**Interprétation :** L'inertie permet de mesurer la distance moyenne (au carré) des points du nuage par rapport au centre du nuage. C'est donc un indicateur de dispersion autour du point moyen et on peut l'interpréter comme une généralisation de la variance.

# Inertie

On a :

$$\mathcal{I} = \frac{1}{n} \sum_{k=1}^n d^2(\vec{x}_k; \vec{\bar{x}})$$

Donc :

# Inertie

On a :

$$\mathcal{I} = \frac{1}{n} \sum_{k=1}^n d^2(\vec{x}_k; \vec{\bar{x}})$$

Donc :

$$\mathcal{I} = \sum_{k=1}^n \mathbb{V}(\vec{x}_k)$$

## Inertie d'un nuage

On a :  $\mathcal{I} = tr(V)$ .

**Remarque :** Comme la matrice  $V$  est symétrique semi-définie positive, elle est donc diagonalisable. Toutes ses valeurs propres sont positives et ses vecteurs propres sont orthogonaux.

L'inertie est donc égale à la somme des valeurs propres de ces matrices.

# Recherche d'un espace de projection

- Objectif : obtenir par transformation de nouvelles données dont la quantité d'information est proche des données initiales.
- Recherche d'un espace de projection : critère, une inertie la plus proche possible de celle du nuage des individus initial.
- Choix du sous-espace **somme directe des sous-espaces propres** associés aux **plus grandes valeurs propres** du spectre de  $V$ .
- **Base de l'espace de projection** : vecteurs propres (orthogonaux) associés aux plus grandes valeurs propres. Il s'agit de combinaisons linéaires des vecteurs des axes initiaux qui représentaient les variables observées.
- **Vecteurs principaux** : vecteurs obtenus en projetant les individus initiaux sur l'espace de projection, on les appelle aussi composantes principales du nuage.



# Recherche d'un plan de projection

- Objectif : déformer le moins possible le nuage de données
- Recherche d'un plan de projection orthogonale : critère, une inertie la plus proche possible de celle du nuage.
- Choix du sous-espace **somme directe des sous-espaces propres** associés aux **plus grandes valeurs propres** du spectre de  $V$ .
- **Axes principaux** du nuage : Droites qui définissent le plan de projection
- **Plan principal** du nuage : plan de projection

# Méthode

On en déduit donc une technique pour **représenter graphiquement** le nuage des individus en minimisant les déformations possibles :

- i) Déterminer le **spectre** de la matrice de variance-covariance,
- ii) projeter les individus sur le plan engendré par les sous-espaces propres associés aux **deux plus grandes valeurs propres**.
- iii) Continuer à projeter si on souhaite avoir un pourcentage d'inertie restitué plus important

On appelle les coordonnées des projetés les **composantes principales**.

# Pourcentage d'information restitué

Projeter implique accepter une déformation du nuage.

On sait que

- l'inertie du nuage est :  $\mathcal{I} = tr(V)$ ,
- l'inertie du nuage projeté sur un sous-espace propre est :  $\mathcal{I}_{\lambda_i} = \lambda_i$

On peut donc calculer la part d'inertie, absolue ou relative, restitué par la projection orthogonale :

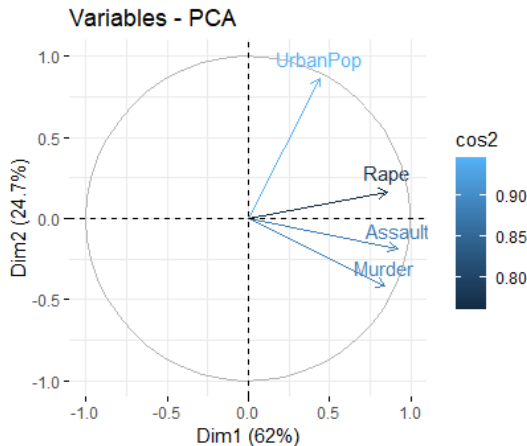
- part d'inertie absolue :  $\lambda_1 + \lambda_2$ ,
- part d'inertie relative :  $\frac{\lambda_1 + \lambda_2}{\mathcal{I}}$ .



# Représentation graphique des variables

- On observe  $p$  variables quantitatives sur  $n$  individus.
- On peut donc considérer que les  $p$  variables représentent  $p$  vecteurs d'un espace des individus de dimension  $n$ . On note  $F$  cet espace.
- On peut appliquer la même méthode et projeter les vecteurs variables sur les sous-espaces propres de  $X_0^t X_0$ .
- Les deux matrices  ${}^t X_0 X_0$  et  $X_0^t X_0$  ayant les mêmes valeurs propres, l'inertie expliquée par chaque axe principal sera la même que pour le nuage des individus.

# Représentation des variables



# Pour aller plus loin

- on pratique souvent l'ACP sur une matrice des corrélations pour éviter les problèmes d'échelles,
- la décomposition en valeurs singulières est souvent préférée aux valeurs propres,
- Interpréter les positions de deux individus sur le plan de projection est risqué si on ne connaît pas leur position vis à vis du plan.