Statistiques descriptives d'un ensemble de variables

Benoit Gaüzère, Stéphane Canu
benoit.gauzere@insa-rouen.fr

INSA Rouen Normandie - ITI

February 26, 2026

Contenu de ce Cours

- Limites de l'analyse bivariée (projections, effet Simpson)
- Représentation matricielle des données
- Matrices de variance/covariance et de corrélation
- Normalisation des données
- Dépendance fonctionnelle et dimension effective
- Transition vers l'ACP

Limites du bivarié : explosion dimensionnelle

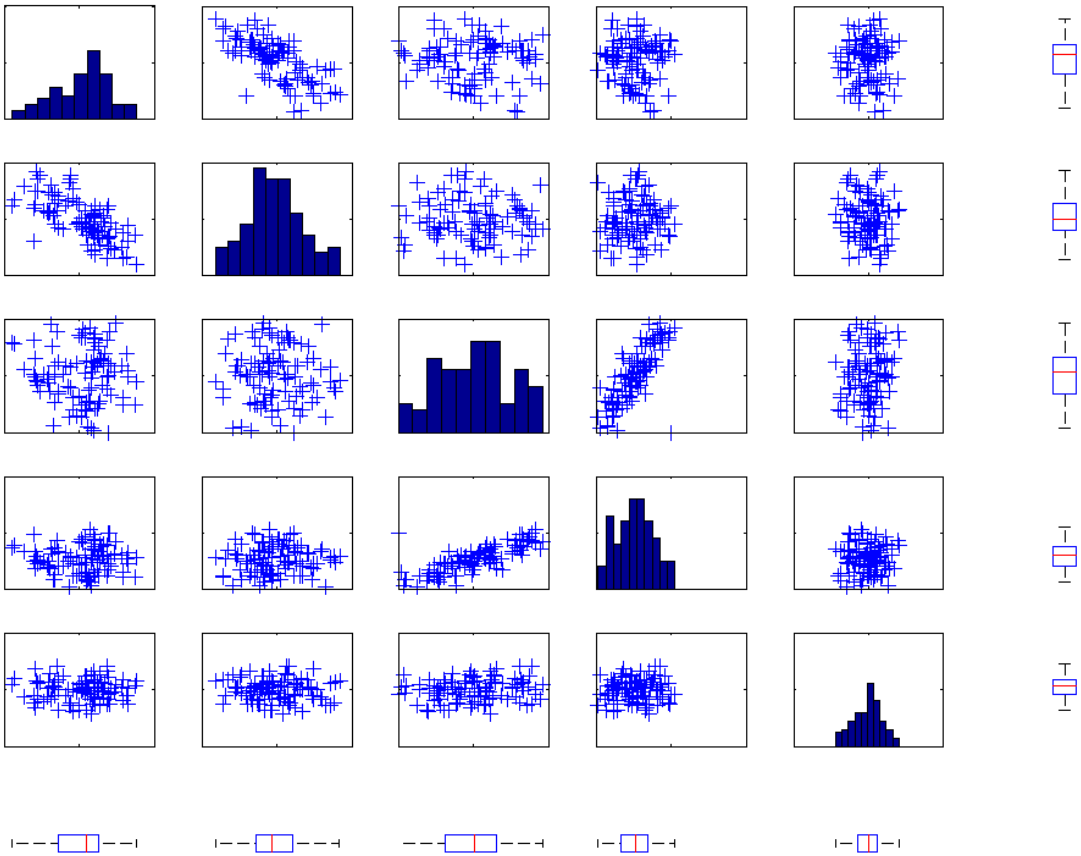
Nombre de relations bivariées

Si l'on a p variables, on a $\frac{p(p-1)}{2}$ relations deux à deux.

Conséquences

- Lecture difficile si p grand
- Multiplication des graphiques
- Perte d'une vision globale

Limites du bivarié : projections



On ne voit que des projections, jamais la structure complète.

Variables qualitatives

- Explosion des tableaux de contingence
- Visualisations ad hoc
- Difficile d'extraire une structure globale

Attention aux limites du bivarié !

Qui sont les meilleurs ?

Résultats d'admission

	Admis	Non Admis
Femmes	27 %	73 %
Hommes	73 %	27 %

Attention aux limites du bivarié !

Qui sont les meilleurs ?

Résultats d'admission

	Admis	Non Admis
Femmes	27 %	73 %
Hommes	73 %	27 %

Résultats par filière

Filière A

	Admis	Non Admis
Femmes	18	72
Hommes	1	9

20 % d'admission chez les femmes, 10 % chez les hommes

Filière B

	Admis	Non Admis
Femmes	9	1
Hommes	72	18

90 % d'admission chez les femmes, 80 % chez les hommes

L'effet Simpson

Observation

- Dans chaque filière, les taux d'admission vérifient Femmes $>$ Hommes, mais globalement Femmes $<$ Hommes.
- Une relation observée entre **Sexe** et **Admission** peut changer lorsqu'on introduit la variable **Filière**.

Interprétation

La filière est une variable de confusion. Les femmes postulent majoritairement à la filière la plus sélective, ce qui fait chuter leur moyenne globale.

L'effet Simpson

Décomposition probabiliste (Moyenne pondérée)

Soit Y l'Admission, X le Sexe, et Z la Filière :

$$P(Y \mid X) = \sum_z \underbrace{P(Y \mid X, Z = z)}_{\text{Taux de réussite}} \times \underbrace{P(Z = z \mid X)}_{\text{Poids de la filière}}$$

Le paradoxe s'explique ici : le "poids" $P(Z \mid X)$ est très déséquilibré selon le sexe (les femmes choisissant massivement la filière A où le taux de réussite est bas).

Ressources

- Une vidéo de 10 min sur ArteTV
- Paradoxe de Simpson sur Wikipedia
- Paradoxe de Simpson et Vaccination COVID

Pourquoi étudier toutes les variables conjointement ?

Objectifs

- Résumer
- Comprendre
- Prévoir
- Aider à la décision

Applications

- Établir des liens de dépendance entre variables
- Visualiser les données en 2D ou 3D:
  - Points aberrants
  - Structure
- Repérer des groupes de variables liées
- Résumer les données pour les transmettre

Dataset Diabetes

Chargement du dataset

- `https://scikit-learn.org/stable/datasets/toy_dataset.html#diabetes-dataset`

```
from sklearn.datasets import load_diabetes
X,y = load_diabetes(return_X_y=True)
```

- $\mathbf{X} \in \mathbb{R}^{446 \times 10}$
- $\mathbf{y} \in \mathbb{R}^{446}$
- Données $\Leftrightarrow$ Matrice

Représentation des données

$\mathbf{X} \in \mathbb{R}^{n \times p}$

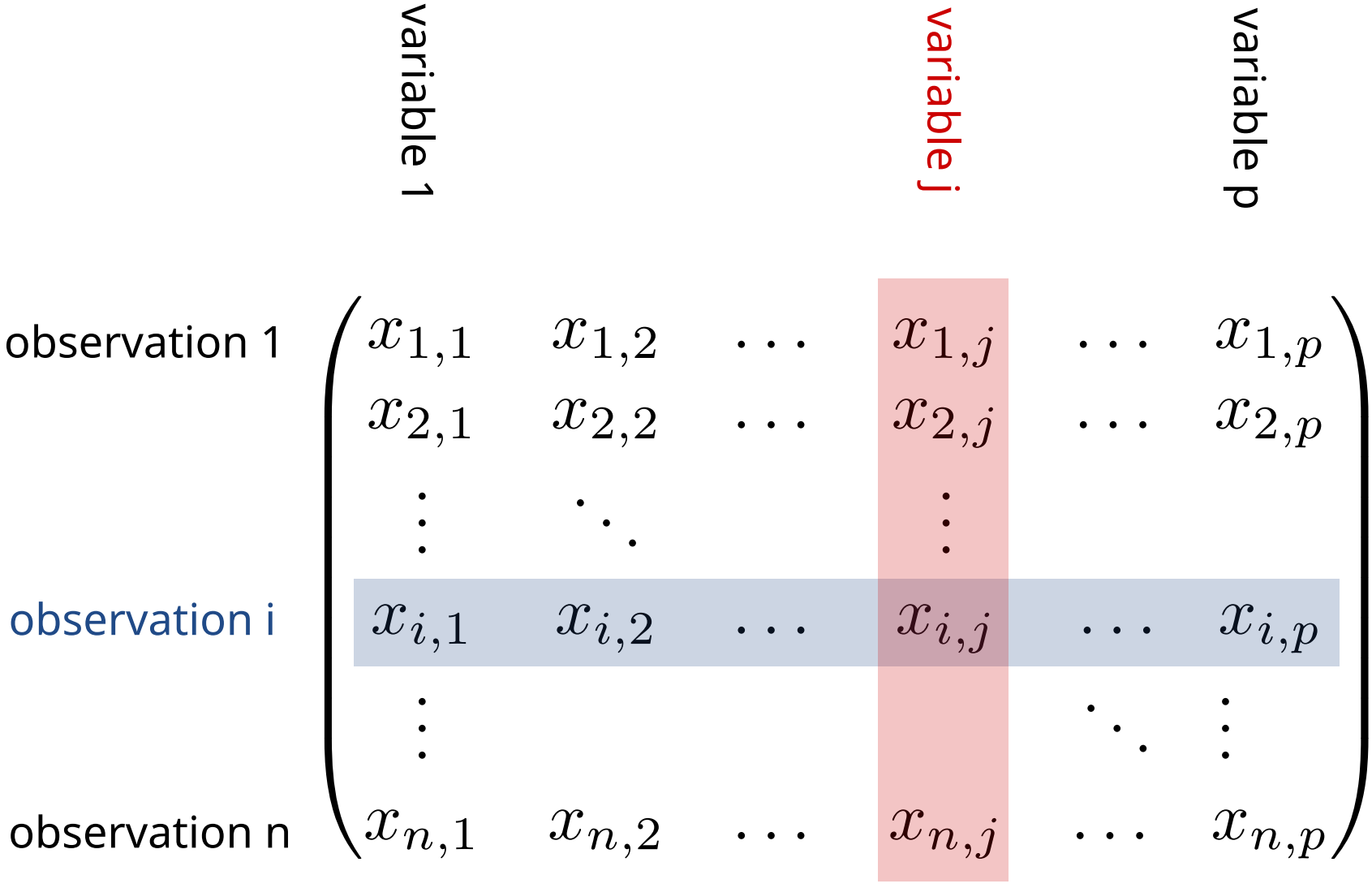
Observations

- n observations (le nombre de lignes)
- $\mathbf{X}(i,:) = \mathbf{x}_{i\bullet} = \mathbf{x}_i^\top$: la i-ème observation
- $\mathbf{x}_i \in \mathbb{R}^p$

Variables

- p variables (le nombre de colonnes)
- Chaque observation est décrite par p variables
- $\mathbf{X}(:,j) = \mathbf{x}_{\bullet j} = \mathbf{x}^{(j)}$: La j-ème variable pour toutes les observations

$\mathbf{X}(i,j)$: La j-ème variable de la i-ème observation.


variable 1
variable j
variable p
observation 1
observation i
observation n
$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,p} \\ \vdots & \ddots & & \vdots & & \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} & \cdots & x_{i,p} \\ \vdots & & & & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & \cdots & x_{n,p} \end{pmatrix}$

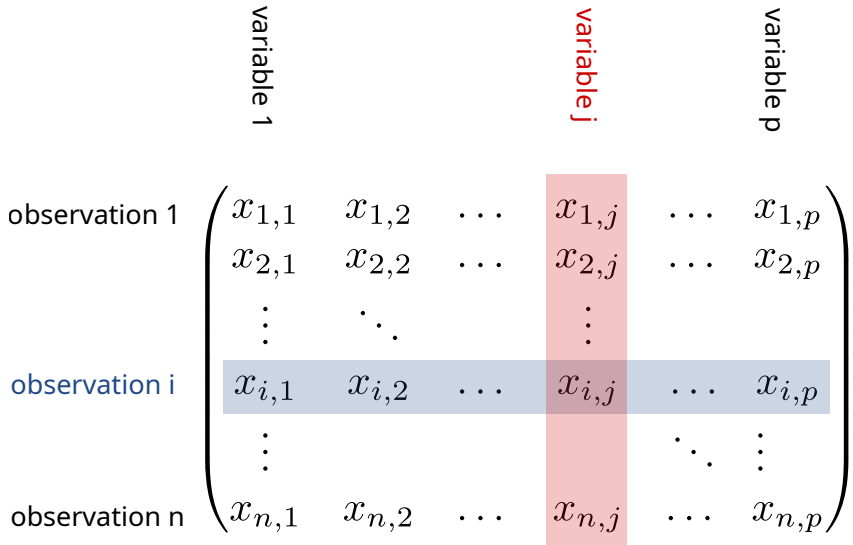
Exemples

https://scikit-learn.org/stable/datasets/toy_dataset.html

Variables quantitatives : qui se ressemble, s'assemble

$\mathbf{x}_i$ est proche de $\mathbf{x}_j$ si une $\mathrm{dist}(\mathbf{x}_i, \mathbf{x}_j)$ (par ex. $\|\mathbf{x}_i - \mathbf{x}_j\|$) est petite.

Variables Quantitatives : Dualité Observations / Variables



- Les individus sont des vecteurs lignes de $\mathbb{R}^p$
  - Chaque individu est un point de $\mathbb{R}^p$
  - $\mathbf{x}_i$ est proche de $\mathbf{x}_i$ si $||\mathbf{x}_i - \mathbf{x}_j||^2$ est petite
- Les variables sont des vecteurs colonnes de $\mathbb{R}^n$
  - Chaque variable est un point de $\mathbb{R}^n$
  - $\mathbf{X}(:, i)$ est proche de $\mathbf{X}(:, j)$ si $\cos(\mathbf{X}(:, i), \mathbf{X}(:, j)) \simeq 1$.

$\Rightarrow$ Similarité des obs **et** des variables

Résumé statistique d'un ensemble de variables

Matrice de variance covariance

Rappels

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \qquad \widehat{s_j^2} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2$$

covariance :

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)(x_{i,k} - \overline{x}_k)$$

corrélation :

$$r_{j,k} = \frac{s_{jk}}{s_j s_k}$$

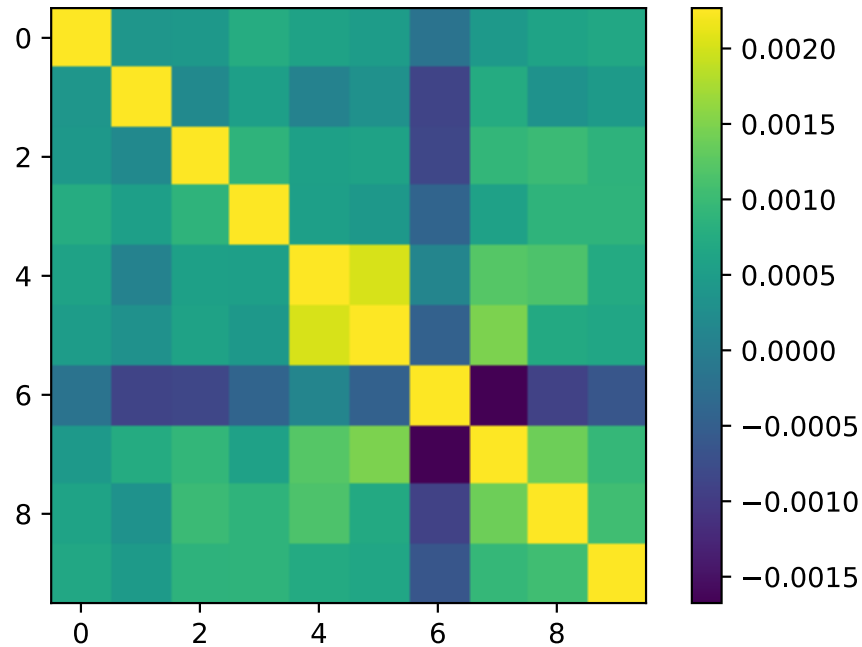
Matrice de variance covariance

Matrice Σ

$$\mathbf{\Sigma}(j,k) = s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)(x_{i,k} - \overline{x}_k)$$

- $\Sigma \in \mathbb{R}^{p \times p}$
- $\mathbf{\Sigma}(j,j) = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2 = \widehat{s_j^2}$
- Analyse la (co)variance des variables
- Données centrées : $\mathbf{X}_c(i,j) = \mathbf{X}(i,j) - \overline{\mathbf{X}(:,j)}$
- $\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}_c^\top \mathbf{X}_c$

Matrice de variance covariance



Matrice de corrélation $\mathbf{R}$

$$\mathbf{R}(j,k) = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)(x_{i,k} - \overline{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2 \sum_{i=1}^{n}(x_{i,k} - \overline{x}_k)^2}}$$

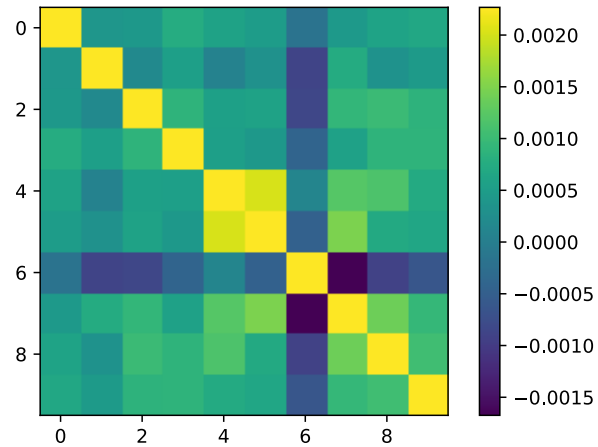
- $\mathbf{R} \in \mathbb{R}^{p \times p}$
- $\mathbf{R}(i,j) \in [-1,1]$
- Données centrées **réduites**

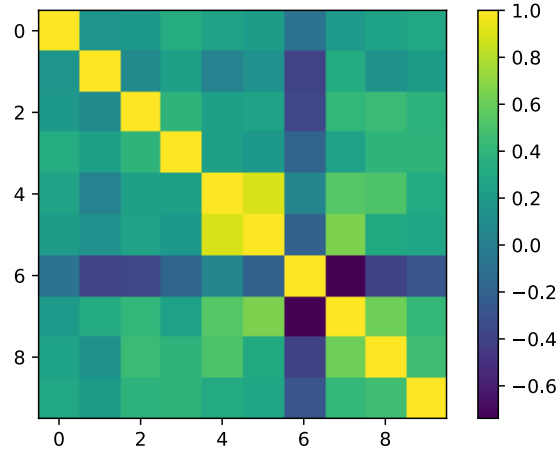
$$\mathbf{X}_r(i,j) = \frac{\mathbf{X}(i,j) - \overline{x}_j}{s_j}$$

- $\mathbf{R} = \frac{1}{n}\mathbf{X}_r^\top \mathbf{X}_r$
- $\mathbf{R}(i,i) = 1$

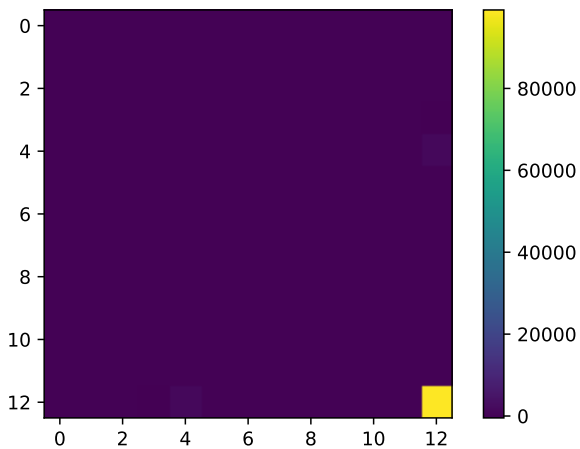
Matrice de covariance et corrélation

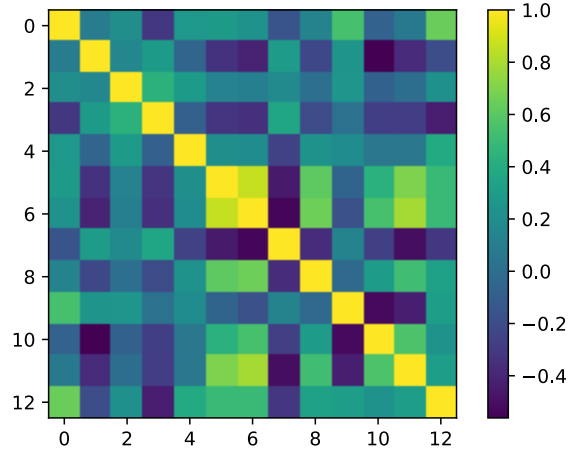
Sur diabetes





Sur wine





Normalisation des données

Pourquoi normaliser ?

- Éviter la surreprésentation d'une variable
- Stabilité numérique
- Comparaison plus facile

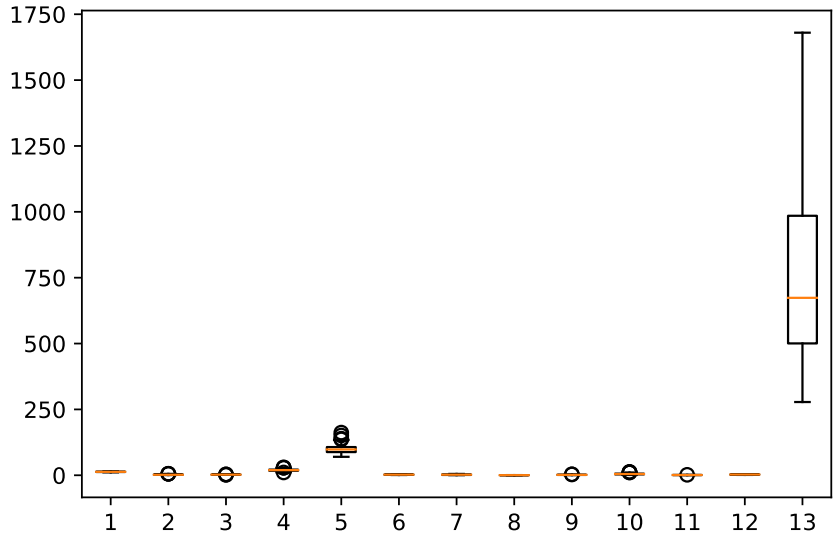
Comment normaliser ?

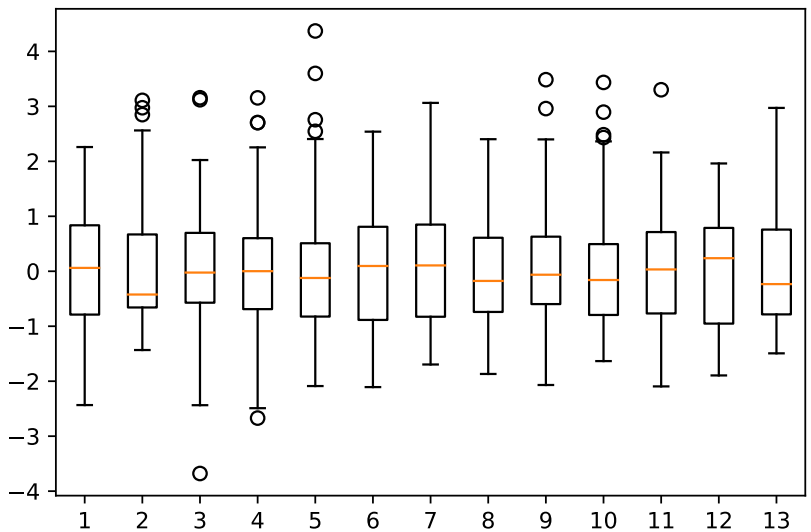
- Hypothèse gaussienne: centrer/réduire

$$\mathbf{X}_r(:, j) = \frac{\mathbf{X}(:, j) - \overline{x}_j}{s_j}$$

- Projection dans un intervalle $\{-1, 1\}, \{0, 1\}$

(non) Normalisé





Dépendance linéaire et redondance

Définition

Une variable peut être exprimée comme une combinaison linéaire parfaite d'autres variables.

$$\mathbf{x}^{(k)} = \alpha_1 \mathbf{x}^{(1)} + \cdots + \alpha_{k-1} \mathbf{x}^{(k-1)}$$

Par exemple, $\mathbf{x}^{(3)} = \mathbf{x}^{(1)} + \mathbf{x}^{(2)}$

Conséquences :

- Redondance d'information : $\mathbf{x}^{(k)}$ n'apporte aucune information nouvelle.
- La matrice de covariance devient **singulière** (non inversible).

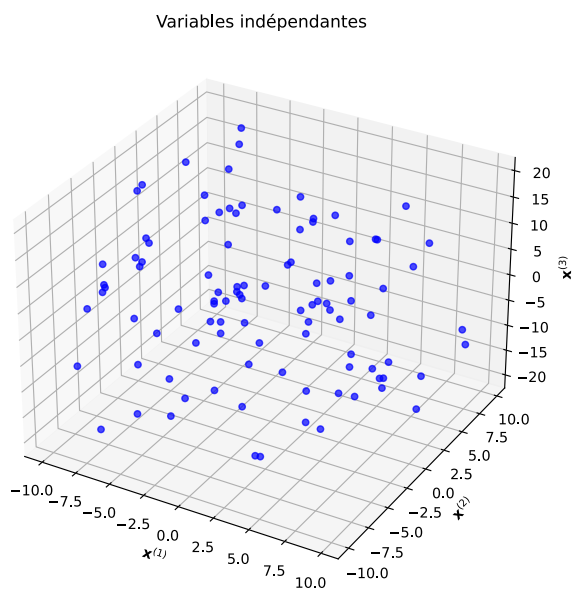
Dimension effective : La géométrie des données

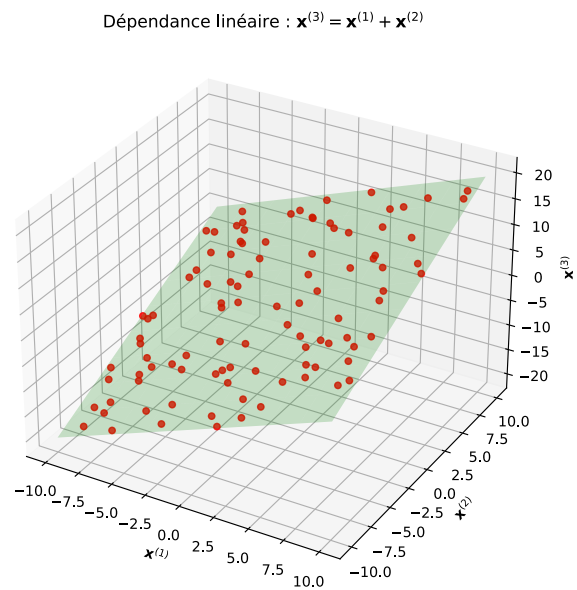
Le sous-espace des données

Considérons un jeu de données avec p variables, $\mathbf{X} \in \mathbb{R}^{n \times p}$. Les observations peuvent "vivre" dans un espace géométrique de dimension strictement inférieure à p, $\text{rank}(\mathbf{X}) < p$.

Intuition géométrique

Si nous avons 3 variables ($p = 3$) mais que $\mathbf{x}^{(3)} = \mathbf{x}^{(1)} + \mathbf{x}^{(2)}$, tous nos points de données sont "piégés" sur un **plan 2D** à l'intérieur de notre espace 3D. La dimension effective n'est que de 2 !





Conclusion

Représentation des données

- Tableau 2D
- Matrice $\mathbf{X} \in \mathbb{R}^{n \times p}$

Résumé qualitatif (graphiques)

- Var. qualitative : c'est compliqué
- Var. quantitative : Nuage de points 1 vs 1

Résumé quantitatif

- Ordre 1 : Tendance centrale
- Ordre 2 : Matrice de corrélation
- ACP : prochain cours