

Statistiques descriptives d'un ensemble de variables

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

INSA Rouen Normandie - ITI

March 6, 2025

Contenu de ce Cours

- ▶ Représentations des données sous forme de matrices
- ▶ Matrice de variance/covariance et de corrélation
- ▶ Normalisation des données
- ▶ Résumé un tableau de données par deux vecteurs
 - ▶ Méthode du Gradient
 - ▶ Introduction à l'ACP

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
- ▶ $\mathbf{X}(:, j) = \mathbf{x}_{\bullet j}$: 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	$x_{1,1}$	$x_{1,2}$	\dots	$x_{1,j}$	\dots	$x_{1,p}$
	$x_{2,1}$	$x_{2,2}$	\dots	$x_{2,j}$	\dots	$x_{2,p}$
	\vdots	\ddots		\vdots		
observation i	$x_{i,1}$	$x_{i,2}$	\dots	$x_{i,j}$	\dots	$x_{i,p}$
	\vdots				\ddots	\vdots
observation n	$x_{n,1}$	$x_{n,2}$	\dots	$x_{n,j}$	\dots	$x_{n,p}$

Exemples

https:

`//scikit-learn.org/stable/datasets/toy_dataset.html`

Pourquoi étudier toutes les variables conjointement ?

Objectifs

- ▶ Résumer
- ▶ Comprendre
- ▶ Prévoir
- ▶ Aider à la décision

Applications

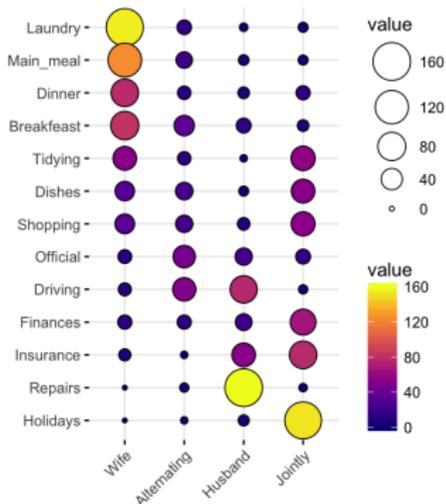
- ▶ Établir des liens de dépendance entre variables
- ▶ Visualiser les données en 2D ou 3D:
 - ▶ Points aberrants
 - ▶ Structure
- ▶ Repérer des groupes de variables liées
- ▶ Résumer les données pour les transmettre

Résumé graphique des données

Visualisation d'un ensemble de variables qualitatives

C'est difficile

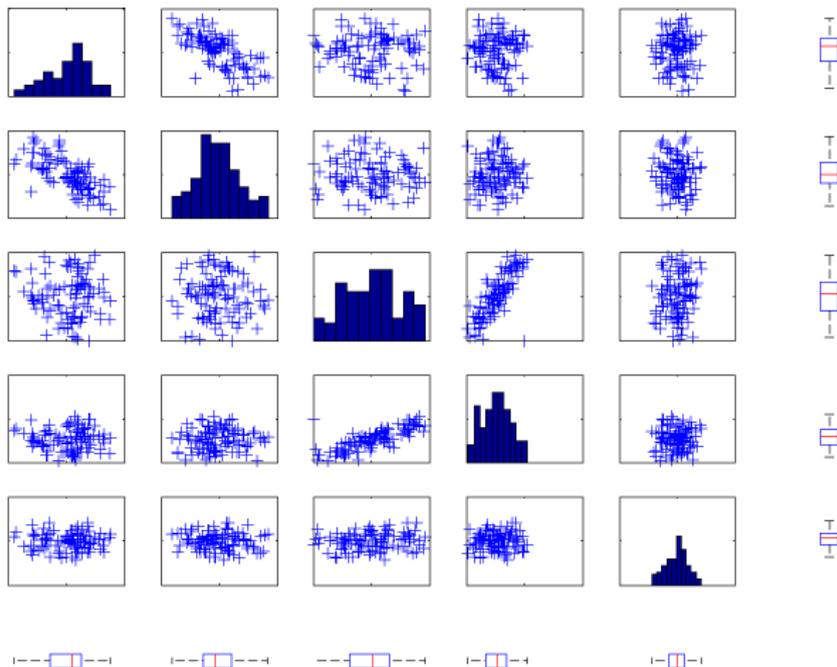
- ▶ 1 vs 1 tableau de contingence
- ▶ Balloon plots
- ▶ Ad hoc solution
- ▶ www.ted.com/talks/view/id/783



Critique

- ▶ Les réponses viennent avant les questions
- ▶ Nous recherchons une aiguille d'information dans la botte de foin des données - *S. Canu*

Visualisation d'un nuage de points de variables quantitatives



On ne représente que des projections

Variables quantitatives : qui se ressemble, s'assemble



x_i est proche de x_j si une $\text{dist}(x_i, x_j)$
(par ex. $\|x_i - x_j\|$) est petite.

Variables Quantitatives : Dualité Observations / Variables

	variable 1		variable j		variable p	
observation 1	$x_{1,1}$	$x_{1,2}$	\dots	$x_{1,j}$	\dots	$x_{1,p}$
	$x_{2,1}$	$x_{2,2}$	\dots	$x_{2,j}$	\dots	$x_{2,p}$
	\vdots	\ddots		\vdots		
observation i	$x_{i,1}$	$x_{i,2}$	\dots	$x_{i,j}$	\dots	$x_{i,p}$
	\vdots				\ddots	\vdots
observation n	$x_{n,1}$	$x_{n,2}$	\dots	$x_{n,j}$	\dots	$x_{n,p}$

- ▶ 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}_j 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 $\cos(\mathbf{X}(:, i), \mathbf{X}(:, j)) \simeq 1$.
- ⇒ Similarité des obs **et** des variables

Résumé statistique d'un ensemble de variables

Matrice de variance covariance

Rappels

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{i,j} \qquad \hat{s}_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2$$

covariance :

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

corrélation :

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

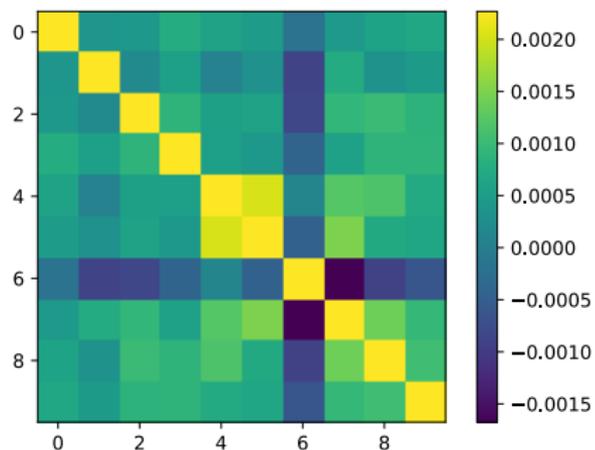
Matrice de variance covariance

Matrice Σ

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

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

Matrice de variance covariance



Matrice de corrélation \mathbf{R}

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

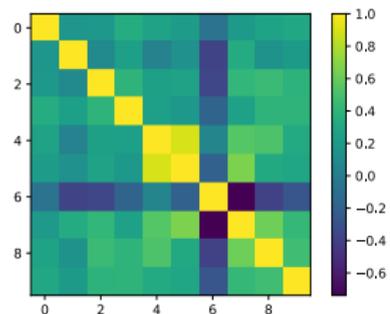
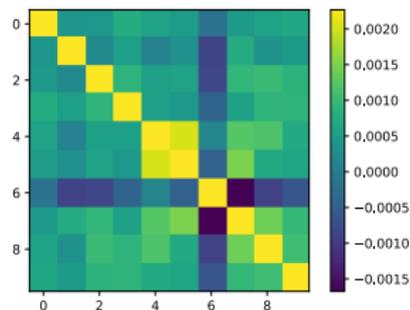
- ▶ $\mathbf{R} \in \mathbb{R}^{p \times p}$
- ▶ $\mathbf{R}(i, j) \in \{-1, 1\}$
- ▶ Données centrées **réduites**

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

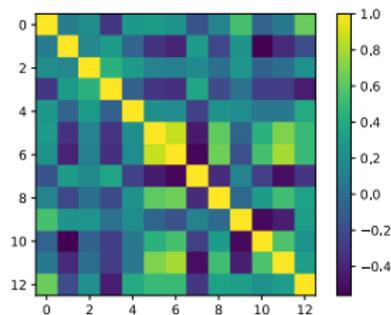
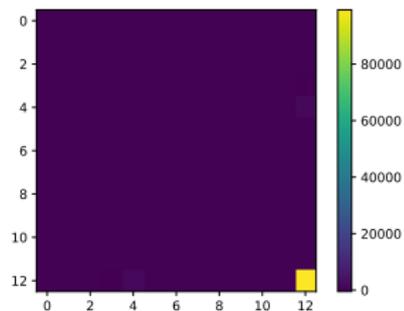
- ▶ $\mathbf{R} = \frac{1}{n} \mathbf{X}_r^\top \mathbf{X}_r$
- ▶ $\mathbf{R}(i, i) = 1$

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

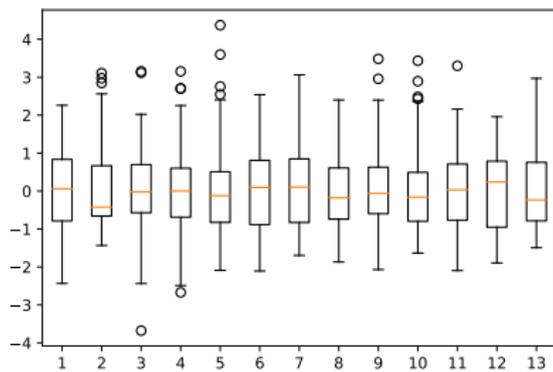
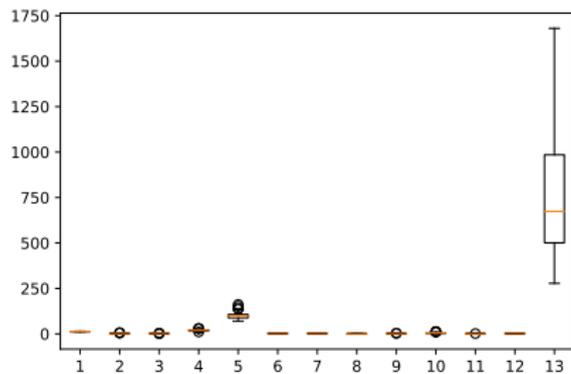
Comment normaliser ?

- ▶ Hypothèse gaussienne: centrer/réduire

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

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

(non) Normalisé



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 : Tendances centrale
- ▶ Ordre 2 : Matrice de corrélation
- ▶ ACP : prochain cours

Bonus

Un autre mode de calcul du gradient

$\mathcal{J}(\mathbf{u}, \mathbf{v}) = -2(\mathbf{X}\mathbf{v})^\top \mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v}\|^2$, le minimum est atteint lorsque le gradient s'annule.

gradient d'une fonction : $J : \mathbb{R}^n \longrightarrow \mathbb{R}$

par définition le gradient $\nabla_J(\mathbf{v})$ s'obtient en posant $\phi(t) = J(\mathbf{v} + t\mathbf{d})$ où \mathbf{d} est un vecteur de \mathbb{R}^p et en calculant la valeur de la dérivée de ϕ au point zéro qui vérifie $\phi'(0) = \mathbf{d}^\top \nabla_J(\mathbf{v})$.

Faisons le calcul

$$\begin{aligned}\phi(t) &= \mathcal{J}(\mathbf{u}, \mathbf{v} + t\mathbf{d}) \\ &= -2(\mathbf{X}(\mathbf{v} + t\mathbf{d}))^\top \mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v} + t\mathbf{d}\|^2 \\ &= -2(\mathbf{v}^\top \mathbf{X}^\top \mathbf{u} + t\mathbf{d}^\top \mathbf{X}^\top \mathbf{u}) + \|\mathbf{u}\|^2 (\|\mathbf{v}\|^2 + t^2 \|\mathbf{d}\|^2 + 2t \mathbf{v}^\top \mathbf{d})\end{aligned}$$

$$\phi'(t) = -2\mathbf{d}^\top \mathbf{X}^\top \mathbf{u} + \|\mathbf{u}\|^2 (2t\|\mathbf{d}\|^2 + 2\mathbf{v}^\top \mathbf{d})$$

$$\phi'(0) = -2\mathbf{d}^\top \mathbf{X}^\top \mathbf{u} + 2\|\mathbf{u}\|^2 \mathbf{v}^\top \mathbf{d}$$

$$= \mathbf{d}^\top (-2\mathbf{X}^\top \mathbf{u} + 2\|\mathbf{u}\|^2 \mathbf{v}) \Rightarrow \nabla_{\mathbf{v}} \mathcal{J}(\mathbf{v}) = -2\mathbf{X}^\top \mathbf{u} + 2\|\mathbf{u}\|^2 \mathbf{v}$$