Statistiques descriptives d'un ensemble de variables

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

INSA Rouen Normandie - ITI

February 25, 2024

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

Représentation des données

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

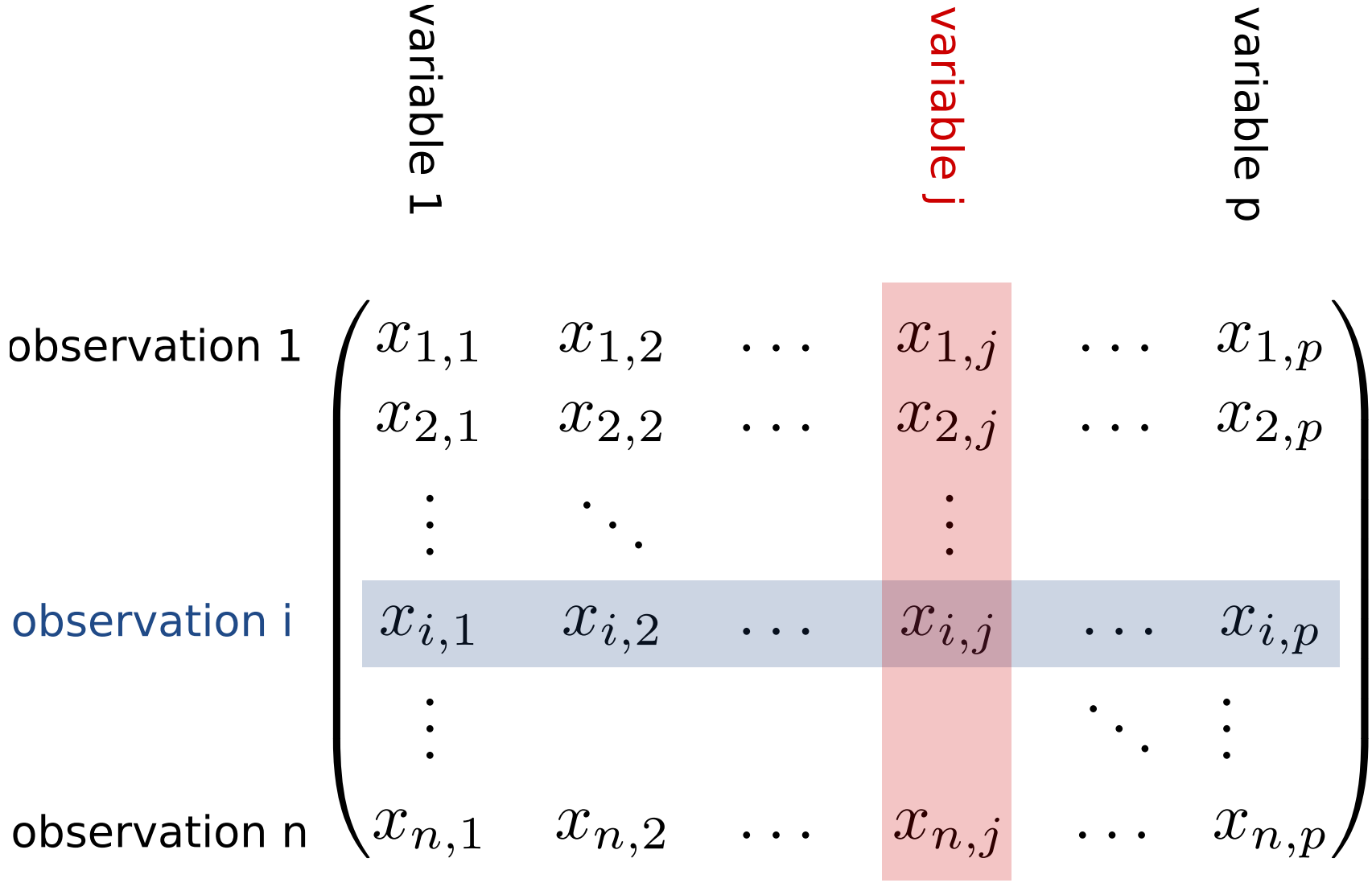
Observations

- 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
observation i
observation n
$x_{1,1}$ $x_{1,2}$ $\cdots$ $x_{1,j}$ $\cdots$ $x_{1,p}$
$x_{2,1}$ $x_{2,2}$ $\cdots$ $x_{2,j}$ $\cdots$ $x_{2,p}$
$\vdots$ $\ddots$ $\vdots$
$x_{i,1}$ $x_{i,2}$ $\cdots$ $x_{i,j}$ $\cdots$ $x_{i,p}$
$\vdots$ $\ddots$ $\vdots$
$x_{n,1}$ $x_{n,2}$ $\cdots$ $x_{n,j}$ $\cdots$ $x_{n,p}$

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

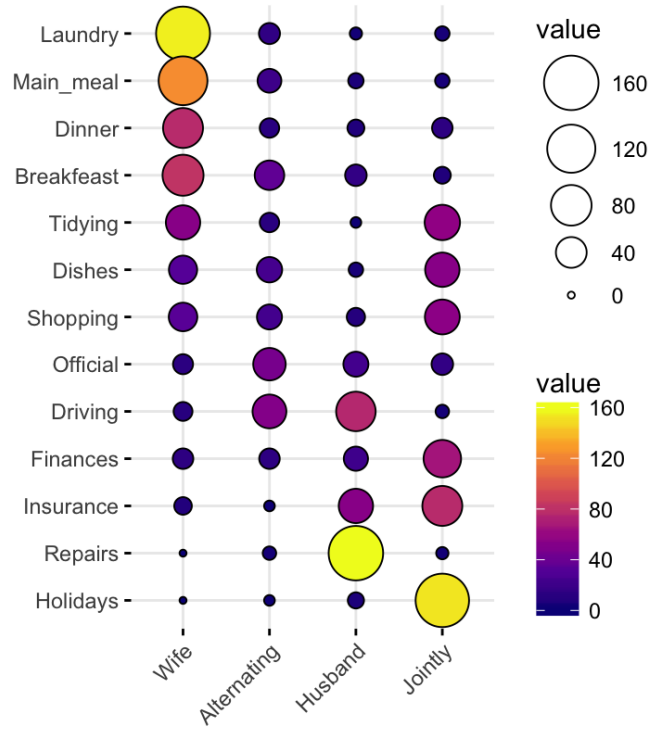
- Établir des liens de dépendance entre variables
- Visualiser les données en 2D ou 3D:
  - Points aberrants
  - Structure
- Repérer des groupes de variables liées
- Résumer les données pour les transmettre

Résumé graphique des données

Visualisation d'un ensemble de variables qualitatives

C'est difficile

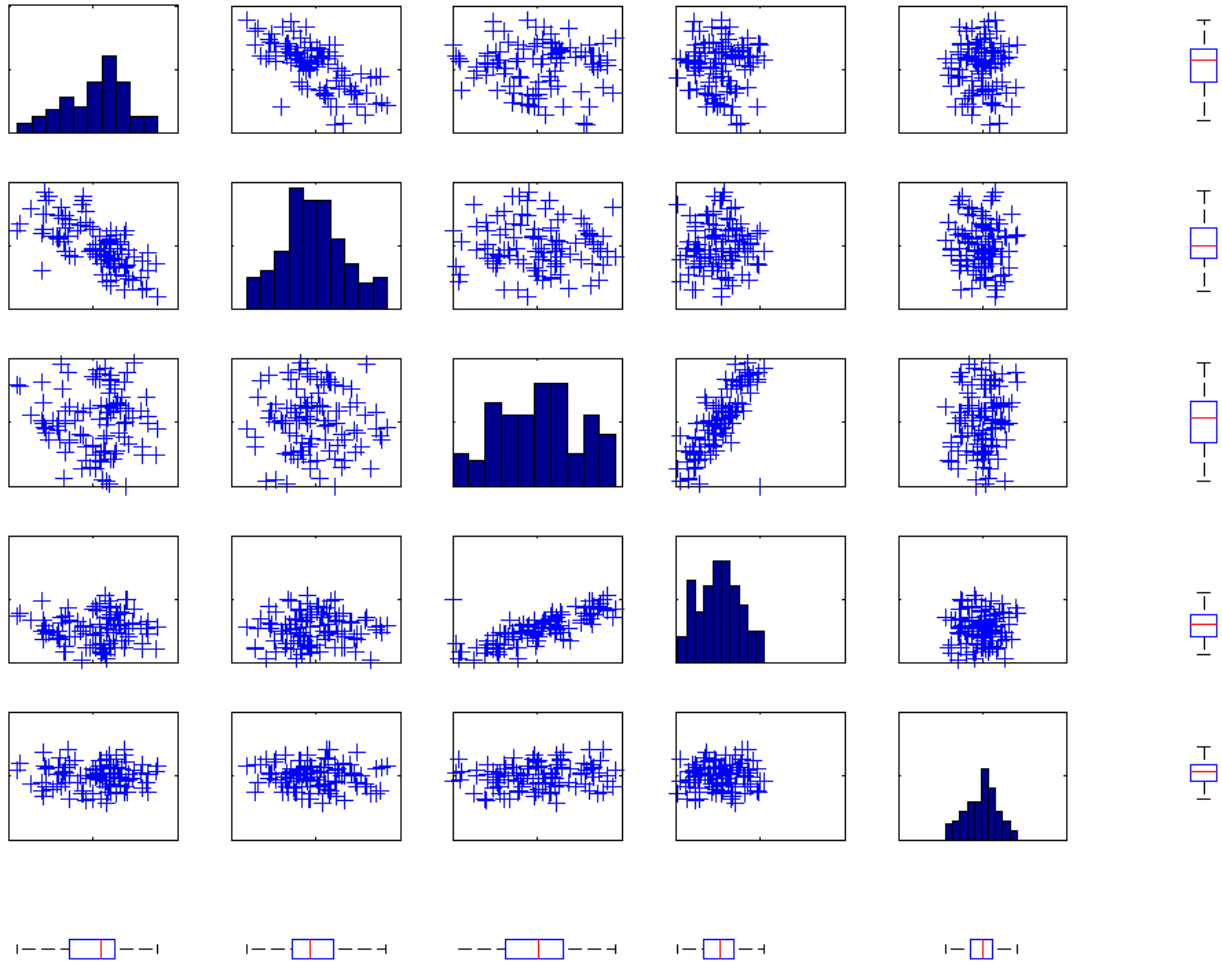
- 1 vs 1 tableau de contigence
- Balloon plots
- Ad hoc solution
- `www.ted.com/talks/view/id/783`



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

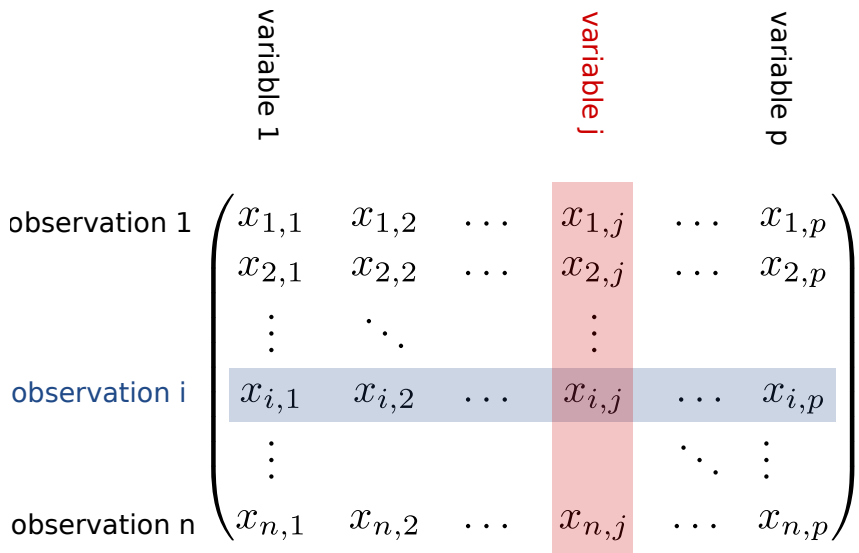


On ne représente que des projections

Variables quantitatives : qui se ressemble, s'assemble

$\mathbf{x}_i$ 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

Résumté 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 :

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

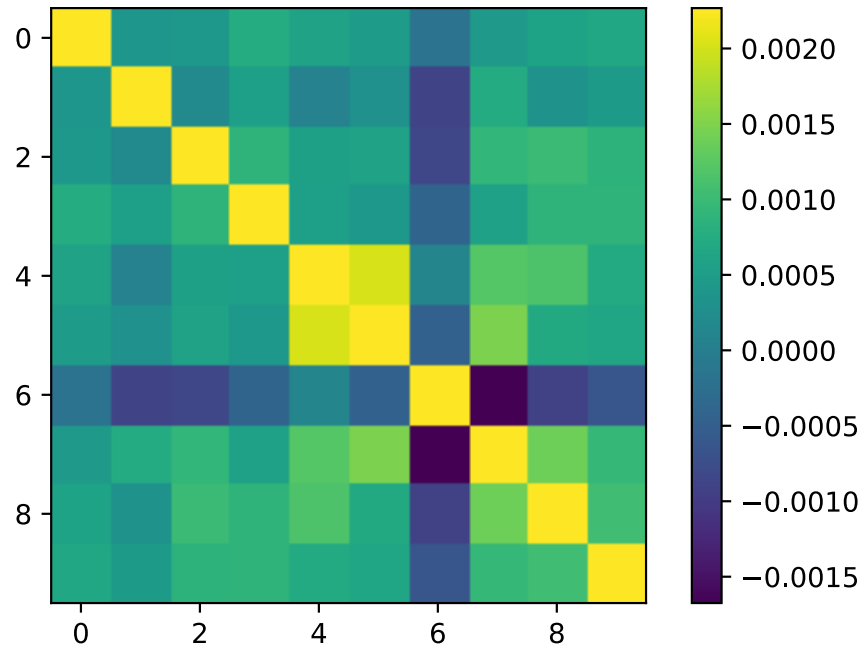
Matrice de variance covariance

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 $\mathbf{R}$

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

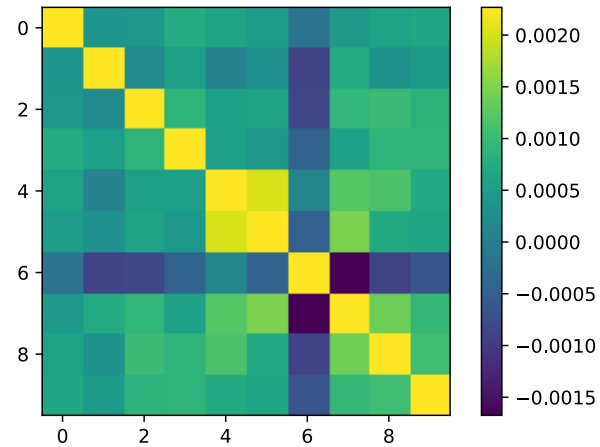
- $\mathbf{R} \in \mathbb{R}^{p \times p}$
- $\mathbf{R}(i,j) \in \{-1, 1\}$
- Données centrées **réduites**

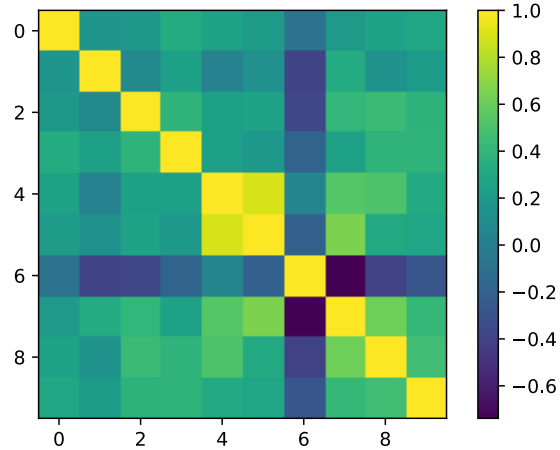
$$\mathbf{X}_r(i,j) = \frac{\mathbf{X}(i,j) - \overline{x}_j}{s_j}$$

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

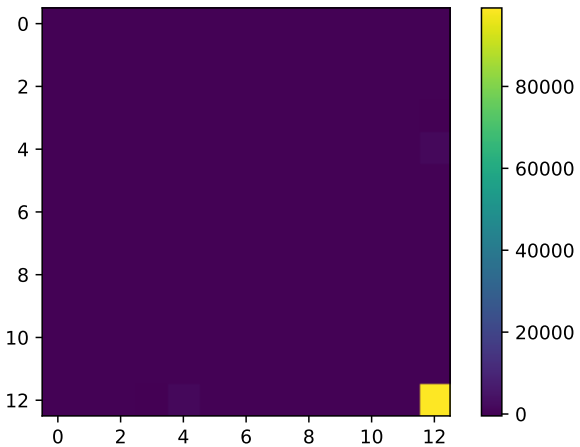
Matrice de covariance et corrélation

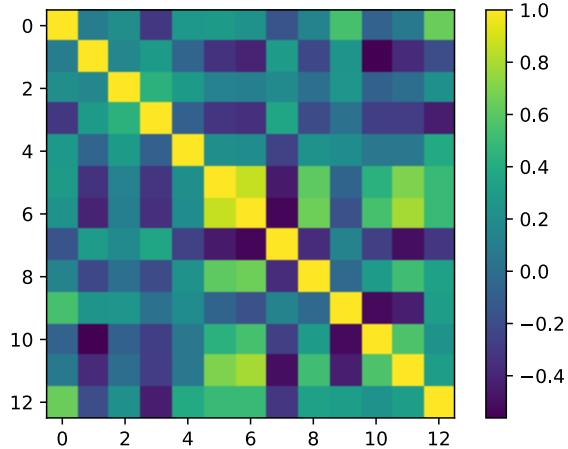
Sur diabetes





Sur wine





Normalisation des données

Pourquoi normaliser ?

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

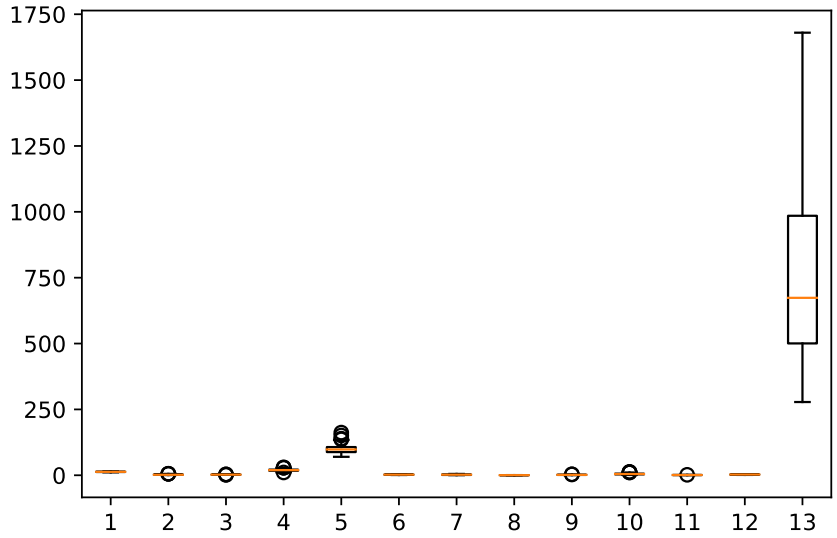
Comment normaliser ?

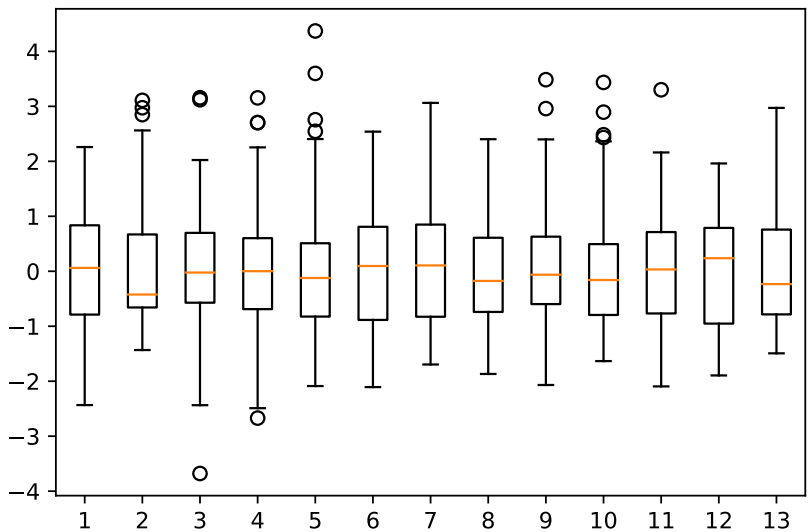
- Hypothèse gaussienne: centrer/réduire

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

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

(non) Normalisé



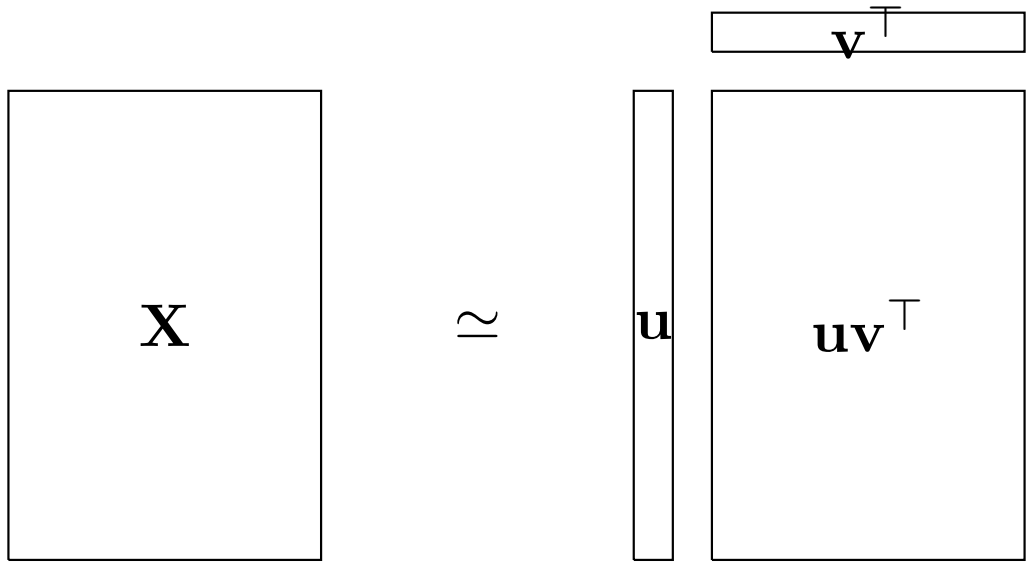


Résumer un tableau de données

Comment résumer l'information ?

- Résumer un tableau de données par deux vecteurs u et v
- $\mathbf{X} \in \mathbb{R}^{n \times p}$

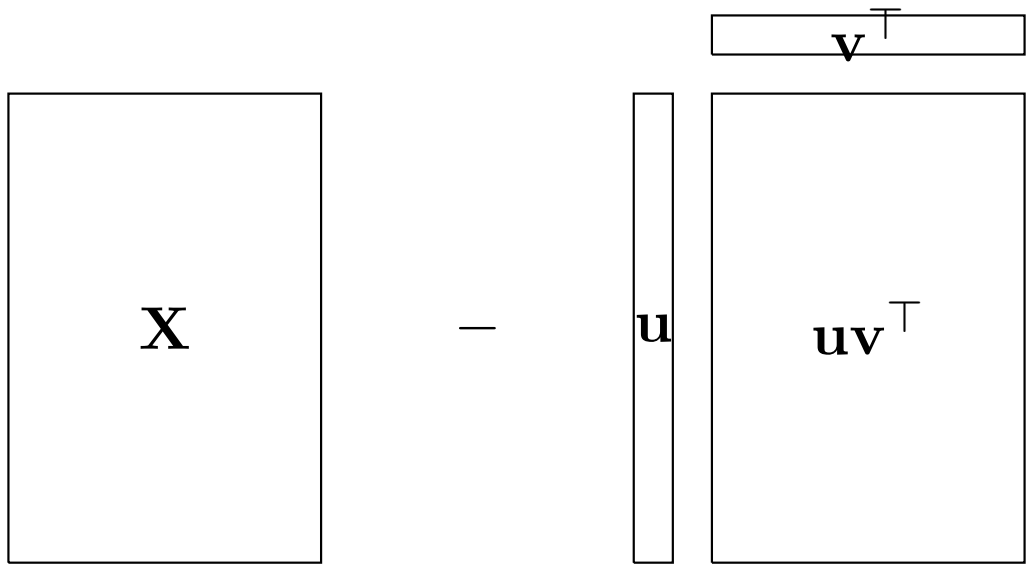
La meilleure représentation linéaire des observations est donnée par le couple de vecteurs $\mathbf{u} \in \mathbb{R}^n$ et $\mathbf{v} \in \mathbb{R}^p$ permettant au mieux de reconstruire la matrice $\mathbf{X}$.



Reconstruction de $\mathbf{X}$

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \quad \text{avec} \quad J(\mathbf{u},\mathbf{v}) = \sum_{i}^{n}\sum_{j}^{p}(x_{ij}-u_i v_j)^2$$

Aussi noté : $J(\mathbf{u},\mathbf{v}) = \|\mathbf{X}-\mathbf{u}\mathbf{v}^\top\|_F^2$



Fonction de coût

La fonction coût $\sum_i^n \sum_j^p (x_{ij} - u_i v_j)^2$ peut se réécrire :

$$\begin{aligned} J(\mathbf{u}, \mathbf{v}) &= \sum_i^n \sum_j^p x_{ij}^2 - 2 \sum_i^n \sum_j^p x_{ij} u_i v_j + \sum_i^n \sum_j^p (u_i v_j)^2 \\ &= \sum_i^n \sum_j^p x_{ij}^2 - 2 \sum_i^n \left(\sum_j^p x_{ij} v_j \right) u_i + \sum_i^n u_i^2 \sum_j^p v_j^2 \\ &= \sum_i^n \sum_j^p x_{ij}^2 - 2 (\mathbf{X}\mathbf{v})^\top \mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v}\|^2 \end{aligned}$$

et donc

$$\min_{\mathbf{u},\mathbf{v}} \underbrace{\|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2}_{J(\mathbf{u},\mathbf{v})} \Leftrightarrow \min_{\mathbf{u},\mathbf{v}} \underbrace{-2(\mathbf{X}\mathbf{v})^\top \mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v}\|^2}_{\mathcal{J}(\mathbf{u},\mathbf{v})}$$

Comment résoudre

$$\min_{\mathbf{u},\mathbf{v}} \mathcal{J}(\mathbf{u},\mathbf{v})?$$

Comment résoudre $\min_{\mathbf{u},\mathbf{v}} \mathcal{J}(\mathbf{u},\mathbf{v})$?

$\Rightarrow$ La méthode du gradient

Méthode du gradient

Minimisation d'une fonction de plusieurs variables

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \;=\; \|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2$$

Définition : Gradient

soit F une fonction de plusieurs (d) variables :

$$\begin{array}{rccc} F: & \mathbb{R}^d & \longmapsto & \mathbb{R} \\ & \mathbf{x} & \longrightarrow & F(\mathbf{x}) \end{array}$$

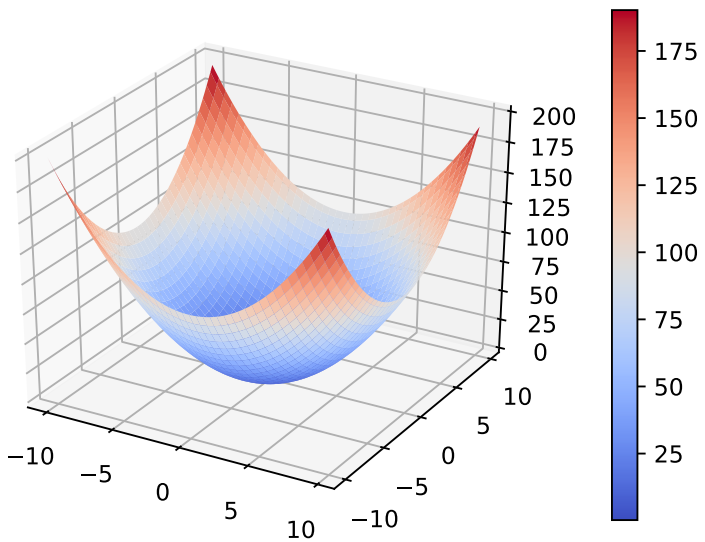
On appelle gradient de F au point $\mathbf{x}$ la fonction des dérivées partielles

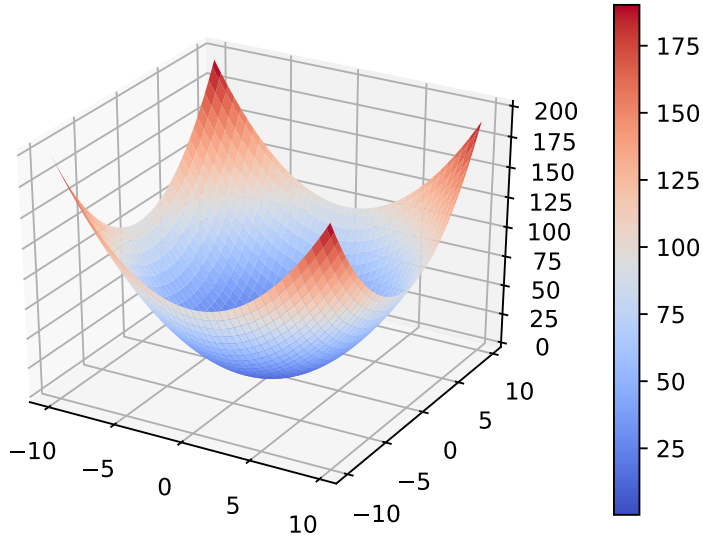
$$\begin{array}{rccc} \nabla_{\mathbf{x}} F: & \mathbb{R}^d & \longmapsto & \mathbb{R}^d \\ & \mathbf{x} & \longrightarrow & \nabla_{\mathbf{x}} F(\mathbf{x}) = \left(\frac{\partial F(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial F(\mathbf{x})}{\partial x_i}, \ldots, \frac{\partial F(\mathbf{x})}{\partial x_d} \right)^\top \end{array}$$

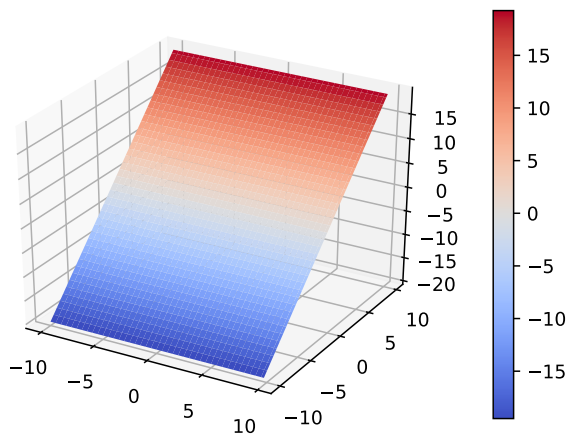
Minimisation d'une fonction de plusieurs variables

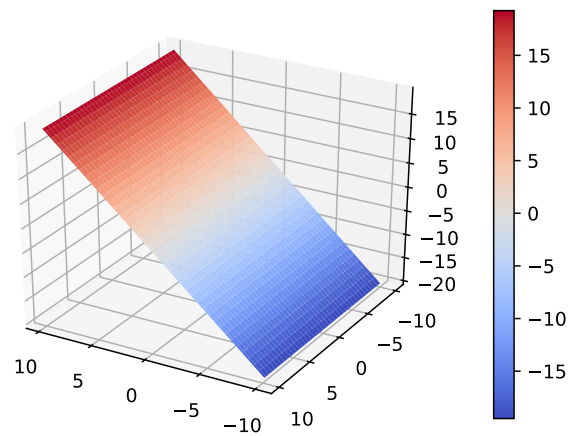
Condition d'optimalité

- $(\mathbf{u}^\star, \mathbf{v}^\star)$ est solution du problème de minimisation ssi le gradient de la fonction J s'annule en ce point
- $$\begin{cases} \nabla_{\mathbf{u}} J(\mathbf{u}^\star, \mathbf{v}^\star) = 0 \\ \nabla_{\mathbf{v}} J(\mathbf{u}^\star, \mathbf{v}^\star) = 0 \end{cases}$$
- F doit être convexe et différentiable
- Si F est strictement convexe : solution unique




200
175
150
125
100
75
50
25
0
−10
−5
0
5
10
175
150
125
100
75
50
25


15
10
5
0
−5
−10
−15
−20
−10
−5
0
5
10
15
10
5
0
−5
−10
−15


15
10
5
0
−5
−10
−15
−20
10
5
0
−5
−10
15
10
5
0
−5
−10
−15

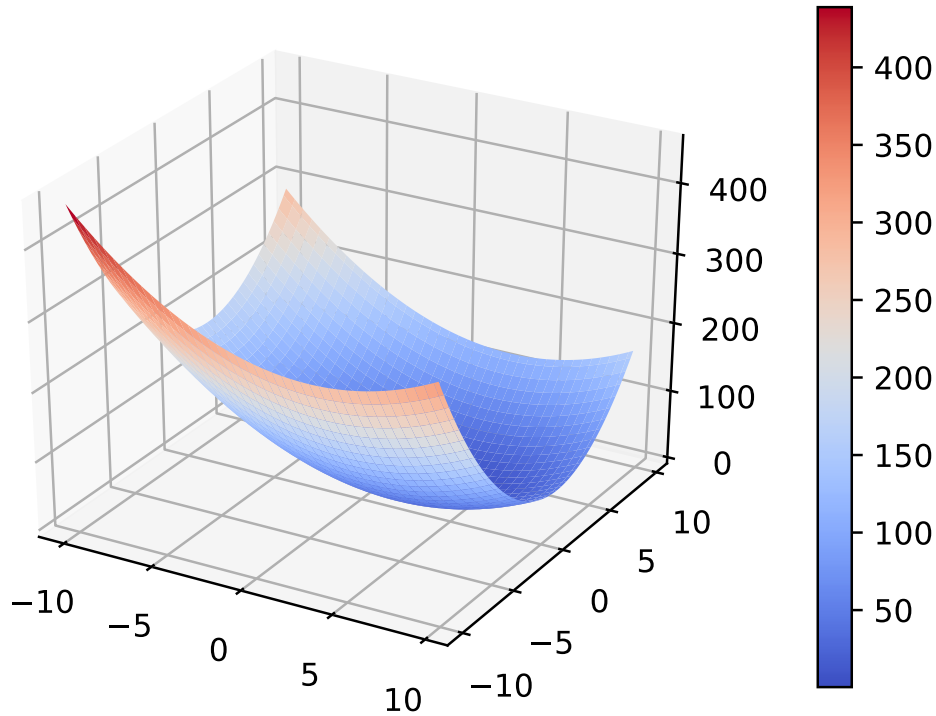
Minimisation d'une fonction de plusieurs variables

Exemple

$$\min_{x,y} J(x,y) \ = 2(x-a)^2 + (y-b)^2$$

Méthode de résolution du problème

1. Calcul du gradient $\nabla_{\mathbf{x}} F(x,y)$ et $\nabla_{\mathbf{y}} F(x,y)$
   - $\nabla_x = 4(x-a)$
   - $\nabla_y = 2(y-b)$
2. Résolution des équations $\nabla_{\mathbf{x}} F(x^\star, y^\star) = 0$ (2 équations à 2 inconnues)
   - $\nabla_x = 0 \Leftrightarrow x^\star = a$
   - $\nabla_y = 0 \Leftrightarrow y^\star = b$


400
300
200
100
0
−10
−5
0
5
10
−10
−5
0
5
10
400
350
300
250
200
150
100
50

Introduction à l'ACP

Reconstruction de $\mathbf{X}$

Rappel du problème

$$\min_{\mathbf{u},\mathbf{v}} \underbrace{\|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2}_{J(\mathbf{u},\mathbf{v})} \Leftrightarrow \min_{\mathbf{u},\mathbf{v}} \underbrace{-2(\mathbf{X}\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2 \|\mathbf{v}\|^2}_{\mathcal{J}(\mathbf{u},\mathbf{v})}$$

Résolution

Minimiser le cout c'est trouver $\mathbf{u}$ et $\mathbf{v}$ qui annulent le gradient :

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) = \|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2 \Leftrightarrow \begin{cases} \nabla_{\mathbf{u}}(\mathbf{u}) = 0 \\ \nabla_{\mathbf{v}}(\mathbf{v}) = 0 \end{cases}$$

Calcul du gradient

$$J(\mathbf{u},\mathbf{v}) = \sum_{i}^{n}\sum_{j}^{p} x_{ij}^2 - 2\sum_{i}^{n}\left(\sum_{j}^{p} x_{ij}v_j\right)u_i + \sum_{i}^{n} u_i^2 \sum_{j}^{p} v_j^2$$

Gradient

$$\begin{cases} \dfrac{\partial J(\mathbf{u},\mathbf{v})}{\partial u_i} = -2\displaystyle\sum_{j}^{p} x_{ij}v_j + 2u_i \sum_{j}^{p} v_j^2 \\ \dfrac{\partial J(\mathbf{u},\mathbf{v})}{\partial v_j} = -2\displaystyle\sum_{i}^{n} x_{ij}u_i + 2v_j \sum_{i}^{n} u_i^2 \end{cases}$$

Écriture matricielle du gradient[1]

$$\begin{cases} \nabla_{\mathbf{u}} J(\mathbf{u}) = -2\mathbf{X}\mathbf{v} + 2\|\mathbf{v}\|^2\mathbf{u} \\ \nabla_{\mathbf{v}} J(\mathbf{v}) = -2\mathbf{X}^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v} \end{cases}$$

[1] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

$\min_{\mathbf{u},\mathbf{v}} \|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2$

Conditions d'optimalité

$$\begin{cases} \nabla_{\mathbf{u}} \mathcal{J}(\mathbf{u}) = 0 & \Leftrightarrow \quad -\mathbf{X}\mathbf{v} + \|\mathbf{v}\|^2 \mathbf{u} = 0 \\ \nabla_{\mathbf{v}} \mathcal{J}(\mathbf{v}) = 0 & \Leftrightarrow \quad -\mathbf{X}^\top \mathbf{u} + \|\mathbf{u}\|^2 \mathbf{v} = 0 \end{cases}$$

Solutions

$$\begin{cases} \mathbf{X}\mathbf{v} = \|\mathbf{v}\|^2 \mathbf{u} \\ \mathbf{X}^\top \mathbf{u} = \|\mathbf{u}\|^2 \mathbf{v} \end{cases} \Rightarrow \mathbf{X}^\top \mathbf{X}\mathbf{v} = \|\mathbf{v}\|^2 \mathbf{X}^\top \mathbf{u} = \underbrace{\|\mathbf{v}\|^2 \|\mathbf{u}\|^2}_{\lambda} \mathbf{v}$$

- Solution $A\mathbf{v} = \lambda \mathbf{v}$
- p solutions $(\mathbf{v}_i, \lambda_i)$.

Quel vecteur propre choisir ?

À l'optimum

$$\mathcal{J}(\mathbf{u},\mathbf{v}) \;=\; -2(\mathbf{X}\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \qquad \text{et} \qquad \mathbf{X}\mathbf{v} = \|\mathbf{v}\|^2\mathbf{u}$$

$$\begin{aligned}\Rightarrow \mathcal{J}(\mathbf{u},\mathbf{v}) \;&=\; -2\|\mathbf{v}\|^2\mathbf{u}^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \\ &=\; -2\|\mathbf{v}\|^2\|\mathbf{u}\|^2 + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \\ &=\; -\|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \qquad\qquad = -\lambda\end{aligned}$$

$$\Rightarrow J(\mathbf{u},\mathbf{v}) \;= \|\mathbf{X}\|_F^2 - \lambda$$

Solution

La solution du problème est donc donnée par le vecteur propre associé à la plus grande valeur propre de la matrice $\mathbf{X}^\top\mathbf{X}$ car :

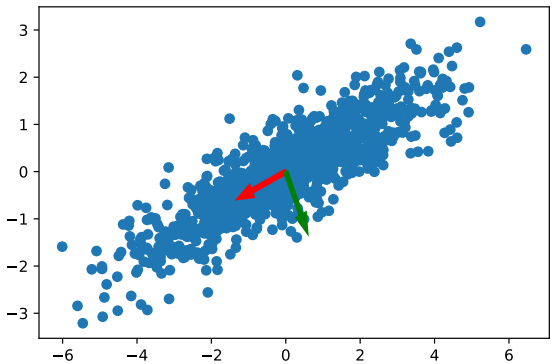
$$\nabla_J(\mathbf{v}) = 0 \quad \Leftrightarrow \quad \mathbf{X}^\top\mathbf{X}\mathbf{v} - \lambda\mathbf{v} = 0 \quad \text{avec} \quad J(\mathbf{u},\mathbf{v}) \;= \|\mathbf{X}\|^2 - \lambda$$

Toutes les v.p. de $\mathbf{X}^\top\mathbf{X}$ (sdp.) sont positives.

En résumé

$\min_{\mathbf{u},\mathbf{v}} \|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2$

- $\mathbf{u} = \mathbf{X}\mathbf{v}$: résumé de $\mathbf{X}$ en 1D
- $\mathbf{v}$: vecteur propre de λ_1 de Σ
- $\mathbf{v}$: projection de $\mathbb{R}^{n\times p} \to \mathbb{R}^{n\times 1}$
- $\mathbf{v}^\top$: projection de $\mathbb{R}^{n\times 1} \to \mathbb{R}^{n\times p}$



Vers l'ACP

- La matrice de projection peut être étendu à plusieurs dimensions
- Reconstructions de plus en plus précises

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

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

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

$$\begin{aligned}
\phi'(t) &= -2\mathbf{d}^\top\mathbf{X}^\top\mathbf{u} + \|\mathbf{u}\|^2(2t\|\mathbf{d}\|^2 + 2\mathbf{v}^\top\mathbf{d}) \\
\phi'(0) &= -2\mathbf{d}^\top\mathbf{X}^\top\mathbf{u} + 2\|\mathbf{u}\|^2\,\mathbf{v}^\top\mathbf{d}
\end{aligned}$$

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