

# SVD

**Nicolas Verzelen, Joseph Salmon (Pierre Pudlo)**

INRA / Université de Montpellier



# Plan

Algèbre linéaire

SVD

Pseudo-inverse

L'approche SVD pour les moindres carrés

# La décomposition spectrale

## Théorème spectral

Une matrice symétrique  $S \in \mathbb{R}^{n \times n}$  est diagonalisable en base orthonormée, i.e., il existe  $\lambda_1 \geq \dots \geq \lambda_n$  et une matrice orthogonale  $U \in \mathbb{R}^{n \times n}$  telle que :

$$S = U \operatorname{diag}(\lambda_1, \dots, \lambda_n) U^\top \text{ ou } SU = U \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

Rem: Si l'on écrit  $U = [\mathbf{u}_1, \dots, \mathbf{u}_n]$  cela signifie que :

$$S = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^\top, \quad \text{avec } \forall i \in \llbracket 1, n \rrbracket, \quad S\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Rappel : une matrice  $U \in \mathbb{R}^{n \times n}$  est dite orthogonale si elle vérifie  $U^\top U = UU^\top = \operatorname{Id}_n$  ou  $\forall (i, j) \in \llbracket 1, n \rrbracket^2 \mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$

Vocabulaire : les  $\lambda_i$  sont les **valeurs propres** de  $S$  et les  $\mathbf{u}_i \in \mathbb{R}^n$  sont les **vecteurs propres** associés

# La décomposition en valeurs singulières ( : *Singular Value Decomposition, SVD*)

## Théorème

Pour toute matrice  $X \in \mathbb{R}^{n \times p}$ , il existe une matrice orthogonale  $U \in \mathbb{R}^{n \times n}$  et une matrice orthogonale  $V \in \mathbb{R}^{p \times p}$ , telles que

$$U^T X V = \text{diag}(s_1, \dots, s_{\min(n,p)}) = \Sigma \in \mathbb{R}^{n \times p}$$

avec  $s_1 \geq s_2 \geq \dots \geq s_{\min(n,p)} \geq 0$ , ou encore :

$$X = U \Sigma V^T$$

avec  $U = [\mathbf{u}_1, \dots, \mathbf{u}_n]$  et  $V = [\mathbf{v}_1, \dots, \mathbf{v}_p]$

Rappel : 
$$\begin{cases} \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}, & \forall (i, j) \in \llbracket 1, n \rrbracket^2 \\ \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{i,j}, & \forall (i, j) \in \llbracket 1, p \rrbracket^2 \end{cases}$$

Démonstration : diagonaliser  $X^T X$  Golub et Van Loan (1996)

# SVD : visualisation

The diagram illustrates the SVD decomposition of a matrix  $X$  into three components:  $U$ ,  $\Sigma$ , and  $V^T$ .

Matrix  $X$  is shown as a rectangle with height  $n$  and width  $p$ .

Matrix  $U$  is shown as a square with height  $n$  and width  $n$ .

Matrix  $\Sigma$  is shown as a rectangle with height  $n$  and width  $p$ . It contains the singular values  $s_1, \dots, s_n$  along the diagonal, with a large  $0$  indicating the rest of the matrix is zero.

Matrix  $V^T$  is shown as a square with height  $p$  and width  $p$ .

The decomposition is represented by the equation:

$$X = U \cdot \Sigma \cdot V^T$$

# SVD : visualisation

The diagram illustrates the SVD decomposition of a matrix  $X$  into three matrices:  $U$ ,  $\Sigma$ , and  $V^T$ .

Matrix  $X$  is a rectangle with height  $n$  and width  $p$ .

Matrix  $U$  is a square with height  $n$  and width  $n$ .

Matrix  $\Sigma$  is a rectangle with height  $n$  and width  $p$ . It contains the singular values  $s_1, \dots, s_p$  along its diagonal and a large  $0$  in the bottom right corner, indicating that the remaining elements are zero.

Matrix  $V^T$  is a square with height  $p$  and width  $p$ .

The decomposition is shown as:

$$X = U \cdot \Sigma \cdot V^T$$

## SVD la suite

Vocabulaire : les  $s_j$  sont les **valeurs singulières** de  $X$  ; les  $\mathbf{u}_j$  (resp.  $\mathbf{v}_j$ ) sont les **vecteurs singuliers** à gauche (resp. droite)

Propriété variationnelle de la plus grande valeur singulière

$$s_1 = \begin{cases} \max_{\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^p} \mathbf{u}^\top X \mathbf{v} \\ \text{s.c. } \|\mathbf{u}\|^2 = 1 \text{ et } \|\mathbf{v}\|^2 = 1 \end{cases}$$

Lagrangien :  $\mathcal{L}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top X \mathbf{v} - \lambda_1(\|\mathbf{u}\|^2 - 1) - \lambda_2(\|\mathbf{v}\|^2 - 1)$

$$\text{CNO : } \begin{cases} \nabla_{\mathbf{u}} \mathcal{L} = X \mathbf{v} - 2\lambda_1 \mathbf{u} = 0 \\ \nabla_{\mathbf{v}} \mathcal{L} = X^\top \mathbf{u} - 2\lambda_2 \mathbf{v} = 0 \end{cases} \iff \begin{cases} X \mathbf{v} = 2\lambda_1 \mathbf{u} \\ X^\top \mathbf{u} = 2\lambda_2 \mathbf{v} \end{cases} \Rightarrow \begin{cases} X^\top X \mathbf{v} = \alpha \mathbf{v} \\ X X^\top \mathbf{u} = \alpha \mathbf{u} \end{cases}$$

avec  $\alpha = 4\lambda_1\lambda_2$ , et donc  $\mathbf{v}$  et  $\mathbf{u}$  sont des vecteurs propres de  $X^\top X$  et de  $XX^\top$

## SVD réduite

On part de la SVD  $X = U\Sigma V^\top$

### SVD réduite

On ne garde que les éléments utiles avec  $r = \min(n, p)$  :

$$X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top = U_r \text{diag}(s_1, \dots, s_r) V_r^\top$$

avec  $s_i > 0, \forall i \in \llbracket 1, r \rrbracket$  et  $U_r = [\mathbf{u}_1, \dots, \mathbf{u}_r], V_r = [\mathbf{v}_1, \dots, \mathbf{v}_r]$

Rem: Quand on en garde que les  $r = \text{rang}(X)$  valeurs singulières non-nulles, on parle alors de **SVD compacte**.

Rem: les matrices  $\mathbf{u}_i \mathbf{v}_i^\top$  sont toutes de rang 1

Rem: les  $\mathbf{u}_i$  (resp. les  $\mathbf{v}_i^\top$ ) sont orthonormés et engendrent le même espace que celui engendré par les colonnes (resp. les lignes) de  $X$

$$\text{vect}(\mathbf{x}_1, \dots, \mathbf{x}_p) = \text{vect}(\mathbf{u}_1, \dots, \mathbf{u}_r)$$



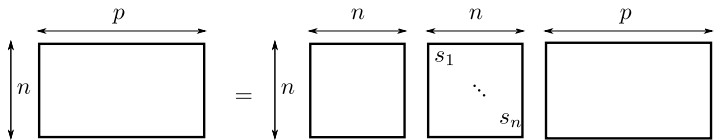
# SVD réduite

The diagram illustrates the reduced SVD decomposition of a matrix  $X$ . On the left, matrix  $X$  is shown as a rectangle with height  $n$  and width  $p$ . This is followed by an equals sign. To the right of the equals sign are three matrices:  $U$ ,  $\Sigma$ , and  $V^T$ . Matrix  $U$  is a square with height  $n$  and width  $n$ . Matrix  $\Sigma$  is a rectangle with height  $n$  and width  $p$ ; its diagonal elements are labeled  $s_1, \dots, s_n$ , and it contains a large  $0$  in the bottom right corner. Matrix  $V^T$  is a square with height  $p$  and width  $p$ . Below the matrices, the equation is written as  $X = U \cdot \Sigma \cdot V^T$ . The word "SVD" is centered below the equation.

$$X = U \cdot \Sigma \cdot V^T$$

SVD

# SVD réduite

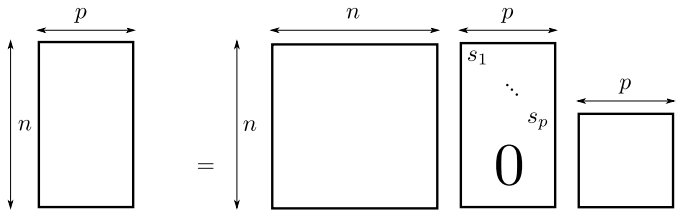


The diagram illustrates the reduced SVD decomposition of a matrix  $X$ . On the left, matrix  $X$  is represented by a rectangle with height  $n$  and width  $p$ . This is followed by an equals sign. To the right of the equals sign are three matrices:  $U$ ,  $\Sigma$ , and  $V^T$ . Matrix  $U$  is a square with height  $n$  and width  $n$ . Matrix  $\Sigma$  is a square with height  $n$  and width  $n$ , containing the singular values  $s_1, \dots, s_n$  along its diagonal. Matrix  $V^T$  is a rectangle with height  $n$  and width  $p$ . The matrices are separated by dots, indicating their product.

$$X = U \cdot \Sigma \cdot V^T$$

SVD réduite

# SVD réduite



The diagram illustrates the reduced SVD decomposition of a matrix  $X$ . On the left, matrix  $X$  is represented by a rectangle with height  $n$  and width  $p$ . This is followed by an equals sign. To the right of the equals sign are three matrices:  $U$ ,  $\Sigma$ , and  $V^T$ , separated by dot operators. Matrix  $U$  is a square with height  $n$  and width  $n$ . Matrix  $\Sigma$  is a rectangle with height  $n$  and width  $p$ ; it contains the singular values  $s_1, \dots, s_p$  along its diagonal and a large  $0$  in the bottom-right corner, indicating it is a truncated SVD. Matrix  $V^T$  is a square with height  $p$  and width  $p$ .

$$X = U \cdot \Sigma \cdot V^T$$

SVD

# SVD réduite

The diagram illustrates the reduced SVD decomposition of a matrix  $X$ . On the left, matrix  $X$  is represented as a rectangle with height  $n$  and width  $p$ . This is followed by an equals sign. To the right of the equals sign, the decomposition is shown as the product of three matrices:  $U$ ,  $\Sigma$ , and  $V^T$ . Matrix  $U$  is a rectangle with height  $n$  and width  $p$ . Matrix  $\Sigma$  is a square with width  $p$ , containing the singular values  $s_1, \dots, s_p$ . Matrix  $V^T$  is a square with width  $p$ .

$$X = U \cdot \Sigma \cdot V^T$$

SVD réduite

# SVD et meilleure approximation

## Théorème (meilleure approximation de rang $k$ )

Soit  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$  la SVD compacte de  $X \in \mathbb{R}^{n \times p}$ .

On note  $X_k = \sum_{i=1}^k s_i \mathbf{u}_i \mathbf{v}_i^\top$ , pour tout  $k \in \llbracket 1, r \rrbracket$ . Ainsi,

$$\min_{Z \in \mathbb{R}^{n \times p} : \text{rang}(Z)=k} \|X - Z\|_2 = \|X - X_k\|_2 = s_{k+1}$$

Rem: la norme spectrale de  $X$  est définie par

$$\|X\|_2 = \sup_{u \in \mathbb{R}^p, \|u\|=1} \|Xu\| = s_1(X)$$

Rem: crucial pour l'analyse en composante principale (ACP)

# Pseudo-inverse

## Définition

Si  $X \in \mathbb{R}^{n \times p}$  admet pour SVD  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$  avec  $r = \text{rang}(X)$ , alors sa **pseudo-inverse**  $X^+ \in \mathbb{R}^{p \times n}$  est définie par :

$$X^+ = \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$$

Rem: si  $X = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{n \times n}$  est inversible alors  $X^+ = X^{-1}$

Démonstration : 
$$X X^+ = \sum_{j=1}^n s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^n \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$$

# Pseudo-inverse

## Définition

Si  $X \in \mathbb{R}^{n \times p}$  admet pour SVD  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$  avec  $r = \text{rang}(X)$ , alors sa **pseudo-inverse**  $X^+ \in \mathbb{R}^{p \times n}$  est définie par :

$$X^+ = \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$$

Rem: si  $X = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{n \times n}$  est inversible alors  $X^+ = X^{-1}$

Démonstration :

$$\begin{aligned} X X^+ &= \sum_{j=1}^n s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^n \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i \mathbf{u}_i^\top \end{aligned}$$

# Pseudo-inverse

## Définition

Si  $X \in \mathbb{R}^{n \times p}$  admet pour SVD  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$  avec  $r = \text{rang}(X)$ , alors sa **pseudo-inverse**  $X^+ \in \mathbb{R}^{p \times n}$  est définie par :

$$X^+ = \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$$

Rem: si  $X = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{n \times n}$  est inversible alors  $X^+ = X^{-1}$

Démonstration :

$$\begin{aligned} X X^+ &= \sum_{j=1}^n s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^n \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \delta_{i,j} \mathbf{u}_j \mathbf{u}_i^\top = \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top = \text{Id}_n \end{aligned}$$



# Pseudo-inverse

## Définition

Si  $X \in \mathbb{R}^{n \times p}$  admet pour SVD  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$  avec  $r = \text{rang}(X)$ , alors sa **pseudo-inverse**  $X^+ \in \mathbb{R}^{p \times n}$  est définie par :

$$X^+ = \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$$

Rem: si  $X = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{v}_i^\top \in \mathbb{R}^{n \times n}$  est inversible alors  $X^+ = X^{-1}$

Démonstration :

$$\begin{aligned} X X^+ &= \sum_{j=1}^n s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^n \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \mathbf{u}_j \mathbf{v}_j^\top \mathbf{v}_i \mathbf{u}_i^\top \\ &= \sum_{j=1}^n \sum_{i=1}^n s_j \frac{1}{s_i} \delta_{i,j} \mathbf{u}_j \mathbf{u}_i^\top = \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top = \text{Id}_n \end{aligned}$$

## SVD et numérique

Les fonctions SVD et pseudo-inverse sont disponibles dans les bibliothèques numériques classiques, par exemple Numpy

► SVD :  $U, s, V = \text{np.linalg.svd}(X)$

Attention dans ce cas :  $X = U \cdot \text{np.diag}(s) \cdot V$   
On accède aux variantes compactes ou non par l'option *cf.*  
`full_matrices=True/False`

► Pseudo-inverse :  $X_{\text{inv}} = \text{np.linalg.pinv}(X)$

---

**Exercice:** Évaluer numériquement le théorème de meilleure approximation de rang fixé. Pour cela calculer l'erreur d'approximation obtenue pour une matrice tirée aléatoirement selon une loi gaussienne (e.g., de taille  $9 \times 6$ , pour  $k = 3$ )

---

# Plan

Algèbre linéaire

L'approche SVD pour les moindres carrés

- SVD et moindres carrés

- Analyse du biais par la SVD

- Analyse de la variance par la SVD

- Stabilité numérique

## Retour sur les moindres carrés

Partons de la SVD de  $X$ ,  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top \beta - \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) - \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

## Retour sur les moindres carrés

Partons de la SVD de  $X$ ,  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top \beta - \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) - \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) \right\|^2 + \left\| \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

## Retour sur les moindres carrés

Partons de la SVD de  $X$ ,  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top \beta - \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) - \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) \right\|^2 + \left\| \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \sum_{i=1}^r (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y})^2 + \sum_{i=r+1}^n (\mathbf{u}_i^\top \mathbf{y})^2$$

## Retour sur les moindres carrés

Partons de la SVD de  $X$ , 
$$X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top \beta - \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) - \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) \right\|^2 + \left\| \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \sum_{i=1}^r (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y})^2 + \sum_{i=r+1}^n (\mathbf{u}_i^\top \mathbf{y})^2$$

Rem: choisir  $\beta = \sum_{i=1}^r \frac{\mathbf{y}^\top \mathbf{u}_i}{s_i} \mathbf{v}_i$  annule le 1<sup>er</sup> terme du 2<sup>d</sup> membre

## Retour sur les moindres carrés

Partons de la SVD de  $X$ ,  $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top \beta - \sum_{i=1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) - \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \left\| \sum_{i=1}^r \mathbf{u}_i (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y}) \right\|^2 + \left\| \sum_{i=r+1}^n \mathbf{u}_i \mathbf{u}_i^\top \mathbf{y} \right\|^2$$

$$\|X\beta - \mathbf{y}\|^2 = \sum_{i=1}^r (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y})^2 + \sum_{i=r+1}^n (\mathbf{u}_i^\top \mathbf{y})^2$$

Rem: choisir  $\beta = \sum_{i=1}^r \frac{\mathbf{y}^\top \mathbf{u}_i}{s_i} \mathbf{v}_i$  annule le 1<sup>er</sup> terme du 2<sup>d</sup> membre



## Retour sur les moindres carrés (suite)

$$\|X\beta - \mathbf{y}\|^2 = \sum_{i=1}^r (s_i \mathbf{v}_i^\top \beta - \mathbf{u}_i^\top \mathbf{y})^2 + \sum_{i=r+1}^n (\mathbf{u}_i^\top \mathbf{y})^2 \geq \sum_{i=r+1}^n (\mathbf{u}_i^\top \mathbf{y})^2$$

avec égalité si  $\beta = \sum_{i=1}^r \frac{\mathbf{y}^\top \mathbf{u}_i}{s_i} \mathbf{v}_i$ , or  $X^+ = \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$  !

Ainsi **UNE** solution des moindres carrés peut s'écrire :

$$\boxed{\hat{\beta} = X^+ \mathbf{y}} \in \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|X\beta - \mathbf{y}\|^2$$

Rem: l'ensemble de toutes les solutions est :

$$\left\{ X^+ \mathbf{y} + \sum_{i=r+1}^p \alpha_i \mathbf{v}_i, (\alpha_{r+1}, \dots, \alpha_p) \in \mathbb{R}^{p-r} \right\}$$

Rem:  $X^+ \mathbf{y}$  est la solution de norme  $\|\cdot\|$  minimale

# Le biais dans le cas général

Sous l'hypothèse de bruit "blanc" (i.e.,  $\mathbb{E}(\varepsilon) = 0$ ) :

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}(X^+ \mathbf{y}) = \mathbb{E}(X^+ X \beta^* + X^+ \varepsilon) = X^+ X \beta^* \\ &= \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \sum_{j=1}^r s_j \mathbf{u}_j \mathbf{v}_j^\top \beta^* \\ &= \sum_{j=1}^r \mathbf{v}_j \mathbf{v}_j^\top \beta^* = \Pi_l \beta^*\end{aligned}$$

- $\Pi_l$  : projecteur sur l'espace des lignes de  $X$

$$\Pi_l = \sum_{i=1}^r \mathbf{v}_i \mathbf{v}_i^\top = X^+ X$$

- $\Pi_c$  : projecteur sur l'espace des colonnes de  $X$

$$\Pi_c = \sum_{i=1}^r \mathbf{u}_i \mathbf{u}_i^\top = X X^+$$

Rem: si  $r = \text{rang}(X) = n$  on retrouve que les MCO sont sans biais

# Variance dans le cas général

## Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que  $X$  est de plein rang :

$$\text{Cov}(\hat{\beta}) = \sigma^2 X^+ (X^+)^{\top}$$

Démonstration : notons  $V = \text{Cov}(\hat{\beta})$

$$\begin{aligned} V &= \mathbb{E} \left[ (\hat{\beta} - \mathbb{E}\hat{\beta})(\hat{\beta} - \mathbb{E}\hat{\beta})^{\top} \right] = \mathbb{E} \left[ (\hat{\beta} - X^+ X \beta^*)(\hat{\beta} - X^+ X \beta^*)^{\top} \right] \\ &= \mathbb{E} \left[ (X^+ \varepsilon)(X^+ \varepsilon)^{\top} \right] \end{aligned}$$

# Variance dans le cas général

## Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que  $X$  est de plein rang :

$$\text{Cov}(\hat{\beta}) = \sigma^2 X^+ (X^+)^{\top}$$

Démonstration : notons  $V = \text{Cov}(\hat{\beta})$

$$\begin{aligned} V &= \mathbb{E} \left[ (\hat{\beta} - \mathbb{E}\hat{\beta})(\hat{\beta} - \mathbb{E}\hat{\beta})^{\top} \right] = \mathbb{E} \left[ (\hat{\beta} - X^+ X \beta^*)(\hat{\beta} - X^+ X \beta^*)^{\top} \right] \\ &= \mathbb{E} \left[ (X^+ \epsilon)(X^+ \epsilon)^{\top} \right] \\ &= \mathbb{E} \left[ X^+ \epsilon \epsilon^{\top} (X^+)^{\top} \right] \end{aligned}$$

# Variance dans le cas général

## Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que  $X$  est de plein rang :

$$\text{Cov}(\hat{\beta}) = \sigma^2 X^+ (X^+)^{\top}$$

Démonstration : notons  $V = \text{Cov}(\hat{\beta})$

$$\begin{aligned} V &= \mathbb{E} \left[ (\hat{\beta} - \mathbb{E}\hat{\beta})(\hat{\beta} - \mathbb{E}\hat{\beta})^{\top} \right] = \mathbb{E} \left[ (\hat{\beta} - X^+ X \beta^*)(\hat{\beta} - X^+ X \beta^*)^{\top} \right] \\ &= \mathbb{E} \left[ (X^+ \epsilon)(X^+ \epsilon)^{\top} \right] \\ &= \mathbb{E} \left[ X^+ \epsilon \epsilon^{\top} (X^+)^{\top} \right] \\ &= \sigma^2 X^+ (X^+)^{\top} = \sum_{i=1}^r \frac{\sigma^2}{s_i^2} \mathbf{v}_i \mathbf{v}_i^{\top} \end{aligned}$$

# Variance dans le cas général

## Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que  $X$  est de plein rang :

$$\text{Cov}(\hat{\beta}) = \sigma^2 X^+ (X^+)^{\top}$$

Démonstration : notons  $V = \text{Cov}(\hat{\beta})$

$$\begin{aligned} V &= \mathbb{E} \left[ (\hat{\beta} - \mathbb{E}\hat{\beta})(\hat{\beta} - \mathbb{E}\hat{\beta})^{\top} \right] = \mathbb{E} \left[ (\hat{\beta} - X^+ X \beta^*)(\hat{\beta} - X^+ X \beta^*)^{\top} \right] \\ &= \mathbb{E} \left[ (X^+ \epsilon)(X^+ \epsilon)^{\top} \right] \\ &= \mathbb{E} \left[ X^+ \epsilon \epsilon^{\top} (X^+)^{\top} \right] \\ &= \sigma^2 X^+ (X^+)^{\top} = \sum_{i=1}^r \frac{\sigma^2}{s_i^2} \mathbf{v}_i \mathbf{v}_i^{\top} \end{aligned}$$

Rem: si  $\text{rang}(X) = n$  on retrouve  $\text{Cov}(\hat{\beta}) = \sigma^2 (X^{\top} X)^{-1}$

# Variance dans le cas général

## Matrice de variance/covariance des moindres carrés

Sous l'hypothèse de modèle homoscédastique et que  $X$  est de plein rang :

$$\text{Cov}(\hat{\beta}) = \sigma^2 X^+ (X^+)^{\top}$$

Démonstration : notons  $V = \text{Cov}(\hat{\beta})$

$$\begin{aligned} V &= \mathbb{E} \left[ (\hat{\beta} - \mathbb{E}\hat{\beta})(\hat{\beta} - \mathbb{E}\hat{\beta})^{\top} \right] = \mathbb{E} \left[ (\hat{\beta} - X^+ X \beta^*)(\hat{\beta} - X^+ X \beta^*)^{\top} \right] \\ &= \mathbb{E} \left[ (X^+ \epsilon)(X^+ \epsilon)^{\top} \right] \\ &= \mathbb{E} \left[ X^+ \epsilon \epsilon^{\top} (X^+)^{\top} \right] \\ &= \sigma^2 X^+ (X^+)^{\top} = \sum_{i=1}^r \frac{\sigma^2}{s_i^2} \mathbf{v}_i \mathbf{v}_i^{\top} \end{aligned}$$

Rem: si  $\text{rang}(X) = n$  on retrouve  $\text{Cov}(\hat{\beta}) = \sigma^2 (X^{\top} X)^{-1}$

# Risque de prédiction

Hypothèse de modèle homoscedastique :  $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \text{Id}_n$

Risque (quadratique) de prédiction  $\mathbb{E}\|X\beta^* - X\hat{\beta}\|^2$

Sous l'hypothèse de modèle homoscedastique :

$$R_{\text{pred}}(\beta^*, \hat{\beta}) = \mathbb{E} \left[ (\hat{\beta} - \beta^*)^\top (X^\top X) (\hat{\beta} - \beta^*) \right] = \sigma^2 \text{rang}(X)$$

Preuve (début identique) :

$$\begin{aligned} R_{\text{pred}}(\beta^*, \hat{\beta}) &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] \\ &\quad + \beta^{*\top} (\Pi_l - \text{Id}_p)^\top (X^\top X) (\Pi_l - \text{Id}_p) \beta^* \\ &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] = \text{tr}[\mathbb{E}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] \end{aligned}$$



# Risque de prédiction

Hypothèse de modèle homoscedastique :  $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \text{Id}_n$

Risque (quadratique) de prédiction  $\mathbb{E}\|X\beta^* - X\hat{\beta}\|^2$

Sous l'hypothèse de modèle homoscedastique :

$$R_{\text{pred}}(\beta^*, \hat{\beta}) = \mathbb{E} \left[ (\hat{\beta} - \beta^*)^\top (X^\top X) (\hat{\beta} - \beta^*) \right] = \sigma^2 \text{rang}(X)$$

Preuve (début identique) :

$$\begin{aligned} R_{\text{pred}}(\beta^*, \hat{\beta}) &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] \\ &\quad + \beta^{*\top} (\Pi_l - \text{Id}_p)^\top (X^\top X) (\Pi_l - \text{Id}_p) \beta^* \\ &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] = \text{tr}[\mathbb{E}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] \\ &= \mathbb{E}[\text{tr}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] = \mathbb{E}[\text{tr}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] \end{aligned}$$

# Risque de prédiction

Hypothèse de modèle homoscédastique :  $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \text{Id}_n$

Risque (quadratique) de prédiction  $\mathbb{E}\|X\beta^* - X\hat{\beta}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\beta^*, \hat{\beta}) = \mathbb{E} \left[ (\hat{\beta} - \beta^*)^\top (X^\top X) (\hat{\beta} - \beta^*) \right] = \sigma^2 \text{rang}(X)$$

Preuve (début identique) :

$$\begin{aligned} R_{\text{pred}}(\beta^*, \hat{\beta}) &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] \\ &\quad + \beta^{*\top} (\Pi_l - \text{Id}_p)^\top (X^\top X) (\Pi_l - \text{Id}_p) \beta^* \\ &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] = \text{tr}[\mathbb{E}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] \\ &= \mathbb{E}[\text{tr}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] = \mathbb{E}[\text{tr}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] \\ &= \text{tr}[\mathbb{E}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] = \text{tr} \Pi_c \mathbb{E}(\varepsilon \varepsilon^\top) \Pi_c^\top \end{aligned}$$

# Risque de prédiction

Hypothèse de modèle homoscédastique :  $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \text{Id}_n$

Risque (quadratique) de prédiction  $\mathbb{E}\|X\beta^* - X\hat{\beta}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\beta^*, \hat{\beta}) = \mathbb{E} \left[ (\hat{\beta} - \beta^*)^\top (X^\top X) (\hat{\beta} - \beta^*) \right] = \sigma^2 \text{rang}(X)$$

Preuve (début identique) :

$$\begin{aligned} R_{\text{pred}}(\beta^*, \hat{\beta}) &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] \\ &\quad + \beta^{*\top} (\Pi_l - \text{Id}_p)^\top (X^\top X) (\Pi_l - \text{Id}_p) \beta^* \\ &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] = \text{tr}[\mathbb{E}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] \\ &= \mathbb{E}[\text{tr}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] = \mathbb{E}[\text{tr}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] \\ &= \text{tr}[\mathbb{E}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] = \text{tr} \Pi_c \mathbb{E}(\varepsilon \varepsilon^\top) \Pi_c^\top \\ &= \sigma^2 \text{tr}(\Pi_c) = \sigma^2 \text{rang}(\Pi_c) = \sigma^2 r = \sigma^2 \text{rang}(X) \end{aligned}$$

# Risque de prédiction

Hypothèse de modèle homoscédastique :  $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \text{Id}_n$

Risque (quadratique) de prédiction  $\mathbb{E}\|X\beta^* - X\hat{\beta}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\text{pred}}(\beta^*, \hat{\beta}) = \mathbb{E} \left[ (\hat{\beta} - \beta^*)^\top (X^\top X) (\hat{\beta} - \beta^*) \right] = \sigma^2 \text{rang}(X)$$

Preuve (début identique) :

$$\begin{aligned} R_{\text{pred}}(\beta^*, \hat{\beta}) &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] \\ &\quad + \beta^{*\top} (\Pi_l - \text{Id}_p)^\top (X^\top X) (\Pi_l - \text{Id}_p) \beta^* \\ &= \mathbb{E} \left[ (X^+ \varepsilon)^\top (X^\top X) (X^+ \varepsilon) \right] = \text{tr}[\mathbb{E}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] \\ &= \mathbb{E}[\text{tr}(\varepsilon^\top \Pi_c^\top \Pi_c \varepsilon)] = \mathbb{E}[\text{tr}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] \\ &= \text{tr}[\mathbb{E}(\Pi_c \varepsilon \varepsilon^\top \Pi_c^\top)] = \text{tr} \Pi_c \mathbb{E}(\varepsilon \varepsilon^\top) \Pi_c^\top \\ &= \sigma^2 \text{tr}(\Pi_c) = \sigma^2 \text{rang}(\Pi_c) = \sigma^2 r = \sigma^2 \text{rang}(X) \end{aligned}$$

## Quelques mots de stabilité numérique

Prenons  $\hat{\beta} = X^+ \mathbf{y}$  comme solution des moindres carrés.

Supposons qu'on observe maintenant non plus  $\mathbf{y}$  mais  $\mathbf{y} + \Delta$  où  $\Delta$  est une erreur très petite :  $\|\Delta\| \ll \|\mathbf{y}\|$ .

Alors l'estimateur des moindres carrés pour  $\mathbf{y} + \Delta$  par  $X$  donne

$$\hat{\beta}^\Delta = X^+(\mathbf{y} + \Delta)$$

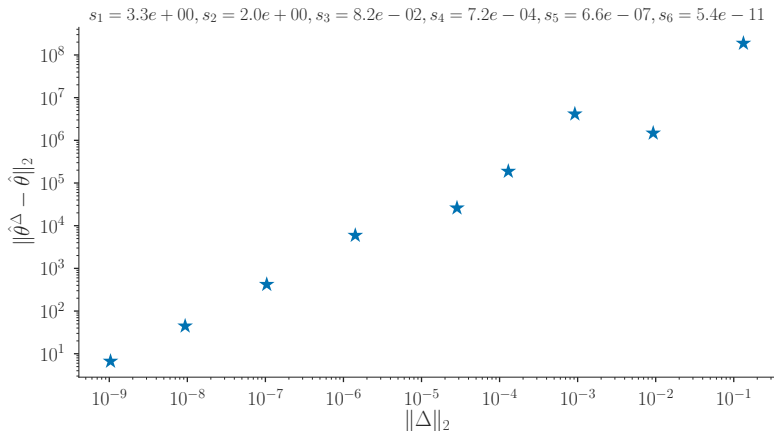
$$\hat{\beta}^\Delta = \hat{\beta} + X^+ \Delta$$

$$\hat{\beta}^\Delta = \hat{\beta} + \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \Delta$$

Rem: Noter l'influence des “petites” valeurs singulières.

# Exemple de problème de conditionnement

$X \in \mathbb{R}^{10 \times 6}$  dont les valeurs singulières sont ci-dessous :



Amplification des erreurs

## Prochains cours :

Remèdes possibles contre les mauvais “conditionnements”

- ▶ Régulariser le spectre / les valeurs singulières
- ▶ Contraindre les coefficients de  $\hat{\beta}$  à n'être pas trop grands

Une solution rendant ces deux points de vue équivalents : *Ridge Regression* / Régularisation de Tychonoff

# References