# SVD

**Nicolas Verzelen, Joseph Salmon**

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 $\lambda_1 \geq \ldots \geq \lambda_n$ et une matrice orthogonale $U \in \mathbb{R}^{n \times n}$ telle que :

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

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

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

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

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

$$U^\top X V = \operatorname{diag}(s_1, \ldots, s_{\min(n,p)}) = \Sigma \in \mathbb{R}^{n \times p}$$

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

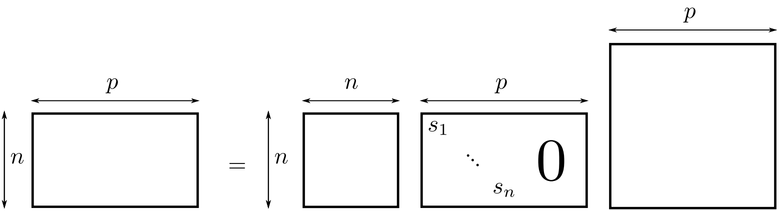
$$\boxed{X = U \Sigma V^\top}$$

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

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

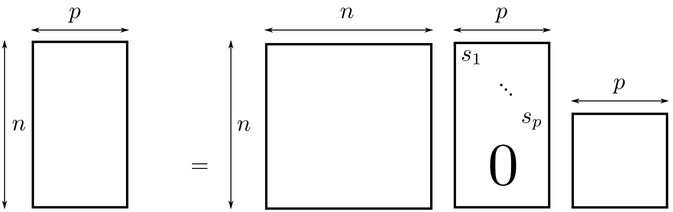
Démonstration : diagonaliser $X^\top X$ Golub et Van Loan (1996)

# SVD : visualisation




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

# SVD : visualisation




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

# 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**

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

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

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

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 $r = \min(n,p)$ :

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

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

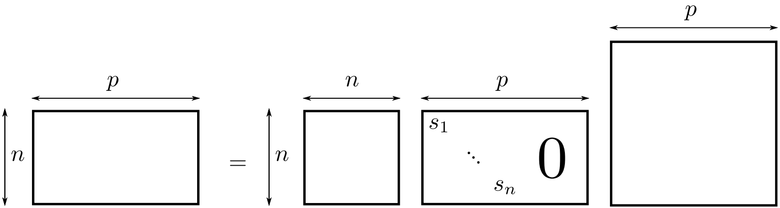
Rem: Quand on en garde que les $r = \operatorname{rang}(X)$ valeurs singulières non-nulles, on parle alors de **SVD compacte**.

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 $X$

$$\boxed{\operatorname{vect}(\mathbf{x}_1, \ldots, \mathbf{x}_p) = \operatorname{vect}(\mathbf{u}_1, \ldots, \mathbf{u}_r)}$$

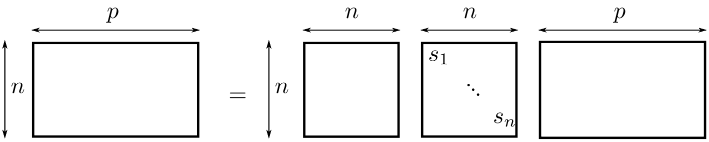
# SVD réduite




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

SVD

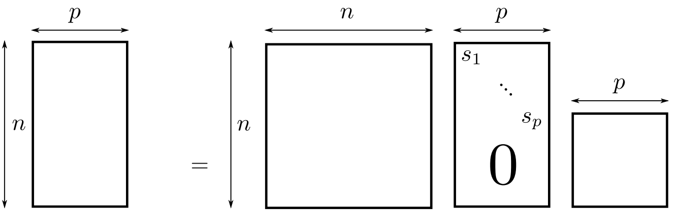
# SVD réduite




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

SVD réduite

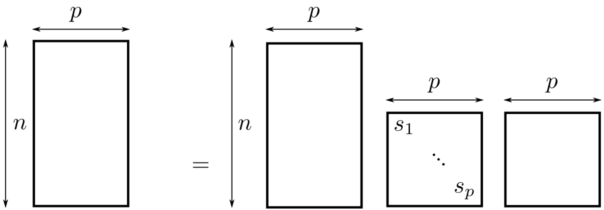
# SVD réduite




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

SVD

# SVD réduite




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

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 $k \in [\![1, r]\!]$. Ainsi,

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

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 $r = \mathrm{rang}(X)$, 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 : $XX^+ = \sum_{j=1}^{n} s_j \mathbf{u}_j \mathbf{v}_j^\top \sum_{i=1}^{n} \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top$

# 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 $r = \mathrm{rang}(X)$, 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 :

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

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

# 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 $r = \mathrm{rang}(X)$, 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 :

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

# 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 $r = \mathrm{rang}(X)$, alors sa **pseudo-inverse** $X^+ \in \mathbb{R}^{p\times n}$ est définie par :

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

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

<u>Démonstration</u> :

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

# SVD et numérique

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

- SVD : `U, s, V = np.linalg.svd(X)`

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

- Pseudo-inverse : `Xinv = np.linalg.pinv(X)`

---

**Exercise**: 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

# Retour sur les moindres carrés

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

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

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

# Retour sur les moindres carrés

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

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

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

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

# Retour sur les moindres carrés

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

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

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

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

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

# Retour sur les moindres carrés

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

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

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

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

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

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

# Retour sur les moindres carrés

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

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

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

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

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

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

# Retour sur les moindres carrés (suite)

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

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

Ainsi **<u>UNE</u>** solution des moindres carrés peut s'écrire :

$$\boxed{\hat{\boldsymbol{\beta}} = X^+ \mathbf{y}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|X\boldsymbol{\beta} - \mathbf{y}\|^2$$

<u>Rem</u>: l'ensemble de toutes les solutions est :

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

<u>Rem</u>: $X^+\mathbf{y}$ est **la** solution de norme $\|\cdot\|_2$ minimale

# Le biais dans le cas général

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

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

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

<u>Rem</u>: si $r = \operatorname{rang}(X) = n$ 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 ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Démonstration : notons $V = \operatorname{Cov}(\hat{\boldsymbol{\beta}})$ et $y = X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$

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

# Variance dans le cas général

**Matrice de variance/covariance des moindres carrés**

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

<u>Démonstration</u> : notons $V = \operatorname{Cov}(\hat{\boldsymbol{\beta}})$ et $y = X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$

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

# Variance dans le cas général

**Matrice de variance/covariance des moindres carrés**

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Démonstration : notons $V = \operatorname{Cov}(\hat{\boldsymbol{\beta}})$ et $y = X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$

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

# Variance dans le cas général

**Matrice de variance/covariance des moindres carrés**

Sous l'hypothèse de modèle homoscédastique $(\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n)$ :

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

<u>Démonstration</u> : notons $V = \operatorname{Cov}(\hat{\boldsymbol{\beta}})$ et $y = X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$

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

<u>Rem</u>: si $\operatorname{rg}(X) = n$ on retrouve $\operatorname{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(X^\top X)^{-1}$

# Variance dans le cas général

**Matrice de variance/covariance des moindres carrés**

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Démonstration : notons $V = \operatorname{Cov}(\hat{\boldsymbol{\beta}})$ et $y = X\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$

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

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

# Risque de prédiction

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

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Preuve (début identique) :

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

# Risque de prédiction

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

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Preuve (début identique) :

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

# Risque de prédiction

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

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Preuve (début identique) :

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

# Risque de prédiction

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

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Preuve (début identique) :

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

# Risque de prédiction

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

Sous l'hypothèse de modèle homoscédastique ($\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$) :

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

Preuve (début identique) :

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

# Quelques mots de stabilité numérique

Prenons $\hat{\boldsymbol{\beta}} = X^{+}\mathbf{y}$ 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

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

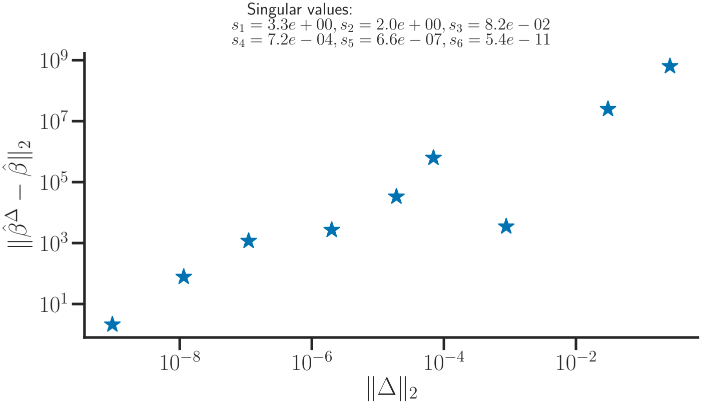
$$\hat{\boldsymbol{\beta}}^{\Delta} = \hat{\boldsymbol{\beta}} + X^{+}\Delta$$

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

<u>Rem</u>: 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 [(1)] :




Amplification des erreurs

(1). voir code `SVD.ipynb` associé

# 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

# Références


- GOLUB, G. H. et C. F. VAN LOAN. *Matrix computations*. Third. Baltimore, MD : Johns Hopkins University Press, 1996, p. xxx+698.