

MS BGD

MDI 720 : SVD

Joseph Salmon

<http://josephsalmon.eu>

Télécom Paristech, Institut Mines-Télécom

Plan

Algèbre linéaire

SVD

Pseudo-inverse

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

Sommaire

Algèbre linéaire

SVD

Pseudo-inverse

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

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 λ_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

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

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

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

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 $X^\top X$ **Golub et Van Loan (1996)**

SVD : visualisation

$$\begin{array}{c} \text{Diagram illustrating the Singular Value Decomposition (SVD) of a matrix } X. \\ \text{The matrix } X \text{ is shown as a rectangle with width } p \text{ and height } n. \\ \text{The decomposition is: } X = U \cdot \Sigma \cdot V^\top \\ \text{Matrix } U \text{ is } n \times n. \\ \text{Matrix } \Sigma \text{ is } n \times p, \text{ diagonal with entries } s_1, \dots, s_n. \\ \text{Matrix } V^\top \text{ is } p \times p. \end{array}$$

SVD : visualisation

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

The diagram illustrates the Singular Value Decomposition (SVD) of a matrix X . The matrix X is represented as a rectangle with height n and width p . It is shown to be equal to the product of three matrices: U , Σ , and V^\top . Matrix U is a square matrix of size $n \times n$. Matrix Σ is a rectangular diagonal matrix of size $n \times p$, representing the singular values s_1, \dots, s_p . Matrix V^\top is a square matrix of size $p \times p$.

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

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

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

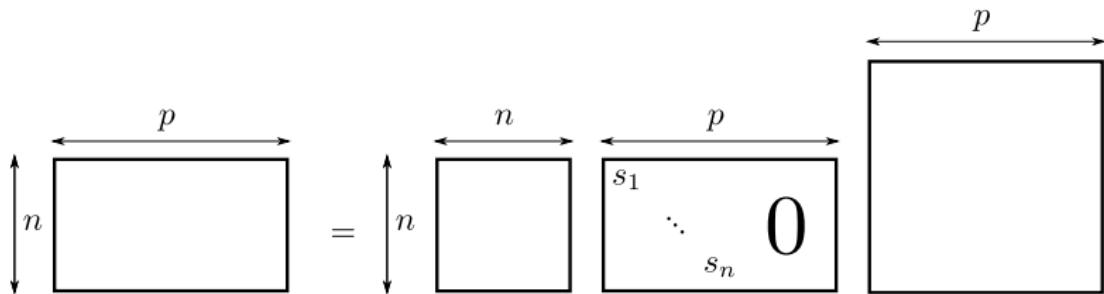
Rem: Quand on en garde que les $r = \operatorname{rang}(X)$ valeurs singulières non-nulles, on parle alors de **SVD compacte**.

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

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

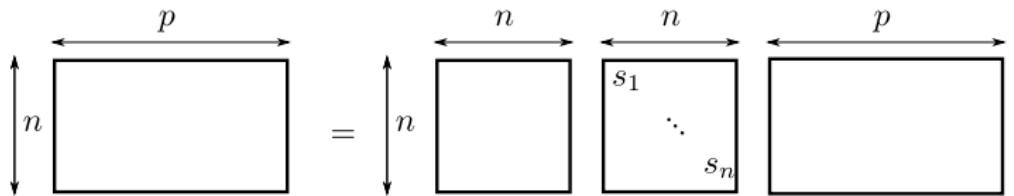
SVD réduite



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

SVD

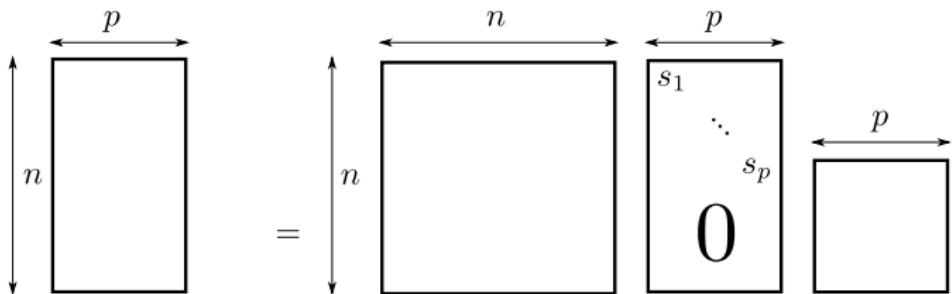
SVD réduite



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

SVD réduite

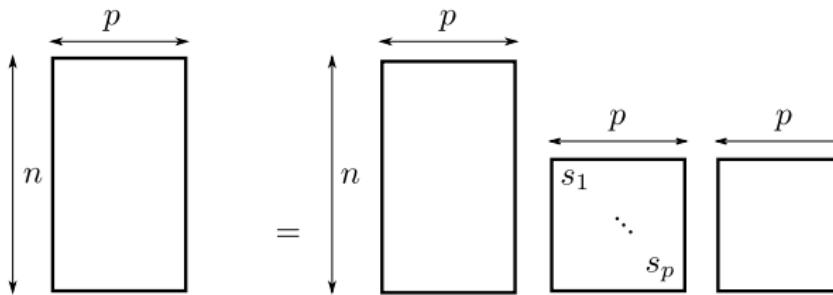
SVD réduite



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

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 $k \in \llbracket 1, r \rrbracket$. 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)

Sommaire

Algèbre linéaire

SVD

Pseudo-inverse

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

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

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

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

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 :

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

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 :

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

SVD et numérique

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

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

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

Exo: 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 9×6 , pour $k = 3$)

Sommaire

Algèbre linéaire

SVD

Pseudo-inverse

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 ,

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

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

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

Retour sur les moindres carrés

Partons de la SVD de X ,

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

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

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

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

Retour sur les moindres carrés

Partons de la SVD de X ,

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

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

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

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

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

Retour sur les moindres carrés

Partons de la SVD de X ,

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

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

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

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

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

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

Retour sur les moindres carrés

Partons de la SVD de X ,

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

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

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

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

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

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

Retour sur les moindres carrés (suite)

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

avec égalité si $\theta = \sum_{i=1}^r \frac{\mathbf{y}^\top \mathbf{u}_i}{s_i} \mathbf{v}_i$, 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 :

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

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

Sommaire

Algèbre linéaire

SVD

Pseudo-inverse

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

Le biais dans le cas général

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

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

- Π_l : projecteur sur l'espace des lignes de X

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

- Π_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

Sommaire

Algèbre linéaire

SVD

Pseudo-inverse

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

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 :

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 X^+ (X^+)^T$$

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

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

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 :

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 X^+ (X^+)^T$$

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

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

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 :

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 X^+ (X^+)^T$$

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

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

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 :

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 X^+ (X^+)^T$$

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

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

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

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 :

$$\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 X^+ (X^+)^T$$

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

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

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

Risque de prédition

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

Risque (quadratique) de prédition $\mathbb{E}\|X\boldsymbol{\theta}^* - X\hat{\boldsymbol{\theta}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Sommaire

Algèbre linéaire

SVD

Pseudo-inverse

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

Quelques mots de stabilité numérique

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

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

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

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

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

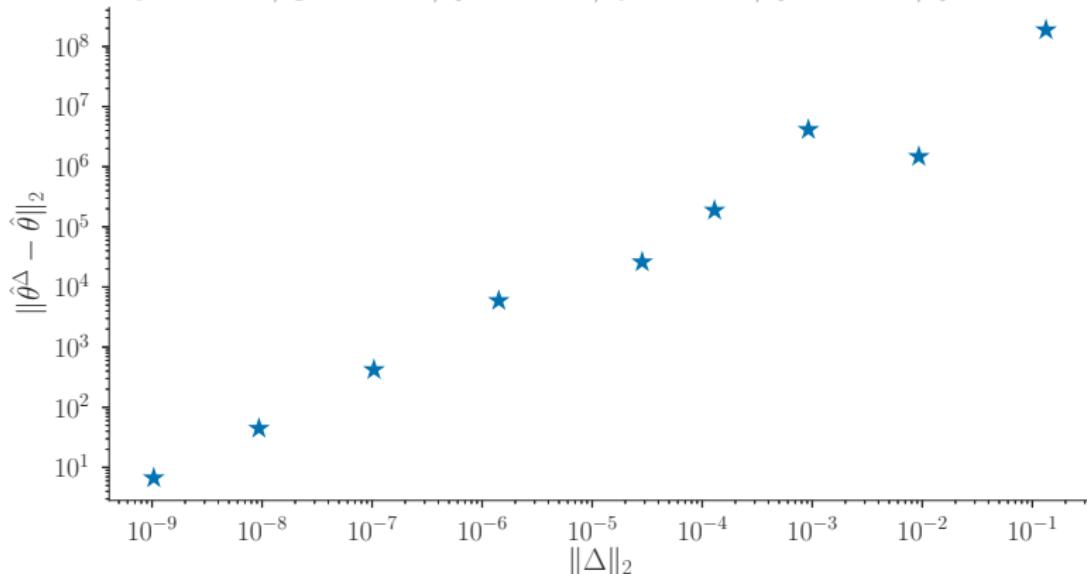
$$\hat{\theta}^\Delta = \hat{\theta} + \sum_{i=1}^r \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \Delta$$

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 :

$$s_1 = 3.3e + 00, s_2 = 2.0e + 00, s_3 = 8.2e - 02, s_4 = 7.2e - 04, s_5 = 6.6e - 07, s_6 = 5.4e - 11$$



Amplification des erreurs

Prochains cours :

Remèdes possibles contre les mauvais “conditionnements”

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

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

Références I

- ▶ G. H. Golub and C. F. van Loan.

Matrix computations.

Johns Hopkins University Press, Baltimore, MD, third edition, 1996.