MS BGD
MDI 720 : SVD

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

Sommaire

La décomposition spectrale

Théorème spectral

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

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

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

$$S = \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top$$

De plus $\forall i \in [\![1, n]\!], \quad S\mathbf{u}_i = \lambda_i \mathbf{u}_i$

<u>Rappel</u> : une matrice orthogonale $U \in \mathbb{R}^{n \times n}$ est une matrice telle que $U^\top U = UU^\top = \mathrm{Id}_n$ ou

$\forall i, j = 1, \ldots, n, \mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$

<u>Vocabulaire</u> : 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

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

avec $s_1 \geqslant s_2 \geqslant \ldots \geqslant s_{\min(n,p)} \geqslant 0$, 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 : $\begin{cases} \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}, & \forall i, j \in [\![1, n]\!] \\ \langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{i,j}, & \forall i, j \in [\![1, p]\!] \end{cases}$

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

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 $XX^\top$

La SVD toujours et encore

SVD compacte

On ne garde que les éléments non-nuls de la diagonale

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

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

<u>Rem</u>: $r = \operatorname{rg}(X)$ nombre de valeurs singulières (non-nulles)
<u>Rem</u>: les matrices $\mathbf{u}_i \mathbf{v}_i^\top$ sont toutes de rang 1
<u>Rem</u>: les vecteurs $\mathbf{u}_i$ (resp. les vecteurs $\mathbf{v}_i^\top$) sont des vecteurs orthonormaux qui engendrent le même espace que celui engendré par les colonnes (resp. les lignes) de X

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

SVD et meilleure approximation

Théorème (meilleure approximation de rang k)

Prenons la SVD de $X \in \mathbb{R}^{n \times p}$ donnée par $X = \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^\top$ (*i.e.*, $r = \mathrm{rg}(X)$). Si $k < r$ et si $X_k = \sum_{i=1}^k s_i \mathbf{u}_i \mathbf{v}_i^\top$ alors

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

Rem: la norme spectrale de X est définie par

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

Rem: ce théorème est aussi crucial pour l'analyse en composante principale (ACP)

Sommaire

Pseudo-inverse

Définition

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

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

<ins>Rem</ins>: Si $X \in \mathbb{R}^{n \times n}$ est inversible (*i.e.*, de rang n) alors $X = \sum_{i=1}^{n} s_i \mathbf{u}_i \mathbf{v}_i^\top$ et alors $X^+ = X^{-1}$

<ins>Démonstration</ins> :

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

SVD et numérique

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

- Pseudo-inverse : `U, s, V = np.linalg.svd(X)`
  Attention dans ce cas :
  `X=np.dot(U, np.dot(np.diag(S), V))`
  Il y a aussi plusieurs variantes matrice pleine ou non
  cf. `full_matrices=True/False`
- Pseudo-inverse : `Xinv = np.linalg.pinv(X)`

Exo: Vérifier numériquement le théorème de meilleure approximation de rang fixé pour une matrice tirée aléatoirement selon une loi gaussienne (*e.g.*, de taille 9×6, pour $k = 3$)

Sommaire

Retour sur les moindres carrés

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

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

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

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

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

<u>Rem</u>: $\boldsymbol{\theta} = \sum_{i=1}^{r} \frac{\langle \mathbf{u}_i, \mathbf{y} \rangle}{s_i} \mathbf{v}_i$ annule le premier terme du 2^d membre

Retour sur les moindres carrés (suite)

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

avec égalité quand $\boldsymbol{\theta} = \sum_{i=1}^{r} \frac{\langle \mathbf{u}_i, \mathbf{y} \rangle}{s_i} \mathbf{v}_i$

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

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

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

L'ensemble de toutes les solutions est l'ensemble :

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

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

Sommaire

Le biais dans le cas général

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

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

Projecteur sur l'espace des lignes de X :

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

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

Sommaire

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 :

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

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

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 :

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

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

Variance dans le cas général

Matrice de variance/covariance des moindres carrés

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

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

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

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

Variance dans le cas général

Matrice de variance/covariance des moindres carrés

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

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

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

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

Rem: si $r = n$ on retrouve le fait que $V = \sigma^2 (X^\top X)^{-1}$

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 :

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

<u>Démonstration</u> : notons $V = \mathrm{Cov}(\hat{\boldsymbol{\theta}})$

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

<u>Rem</u>: si $r = n$ on retrouve le fait que $V = \sigma^2 (X^\top X)^{-1}$

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscédastique :

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

Preuve (début identique) :

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

Sommaire

Quelques mots de stabilité numérique

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

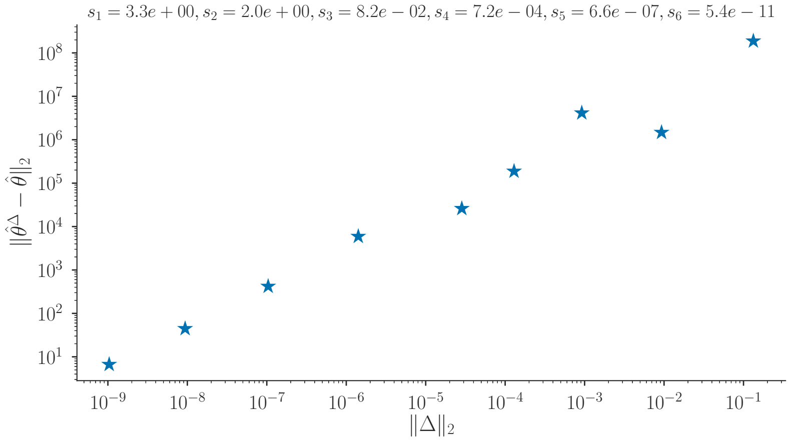
$$\hat{\boldsymbol{\theta}}^{\Delta} = X^+(\mathbf{y} + \Delta)$$

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

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

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

- Régulariser le spectre / les valeurs singulières
- Contraindre les coefficients de $\hat{\boldsymbol{\theta}}$ 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.