

MS BGD

MDI 720 : Statistiques

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

ACP

- Définition

- Interprétation

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

ACP

Définition

Interprétation

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 :

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

De plus $\forall i \in \llbracket 1, n \rrbracket$, $S\mathbf{u}_i = \lambda_i \mathbf{u}_i$

Rappel : une matrice orthogonale $U \in \mathbb{R}^n$ est une matrice telle que $U^\top U = UU^\top = \operatorname{Id}_n$ ou $\forall i, j = 1, \dots, n$, $\mathbf{u}_i^\top \mathbf{u}_j = \langle \mathbf{u}_i, \mathbf{u}_j \rangle = \delta_{i,j}$

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

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

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

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 $\alpha = 2\lambda_1\lambda_2$, 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

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

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: $r = \operatorname{rg}(X)$ nombre de valeurs singulières (non-nulles)

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

Rem: 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

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

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., $r = \text{rg}(X)$). Si $k < r$ et si $X_k = \sum_{i=1}^k s_i \mathbf{u}_i \mathbf{v}_i^\top$ alors

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

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

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

ACP

Définition

Interprétation

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 :

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

Rem: 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 $X^+ = X^{-1}$

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 toutes bibliothèques 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

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

ACP

Définition

Interprétation

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

Retour sur les moindres carrés (suite)

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

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

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

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

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

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

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

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

ACP

Définition

Interprétation

Le biais dans le cas général

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

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

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 :

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

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

ACP

Définition

Interprétation

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 :

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

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

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

Variance dans le cas général

Matrice de variance/covariance des moindres carrés

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

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

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

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

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 :

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

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

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

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 :

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

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

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

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 :

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

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

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

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscedastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscedastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscedastique :

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

Preuve (début identique) :

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

Risque de prédiction

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

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

Sous l'hypothèse de modèle homoscedastique :

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

Preuve (début identique) :

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

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

ACP

Définition

Interprétation

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

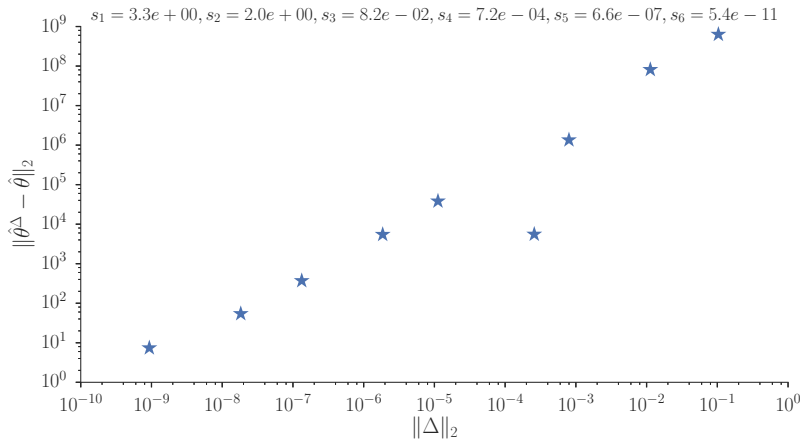
$$\hat{\boldsymbol{\theta}}^\Delta = X^+(\mathbf{y} + \Delta)$$

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

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

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 $\hat{\theta}$ à n'être pas trop grands

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

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

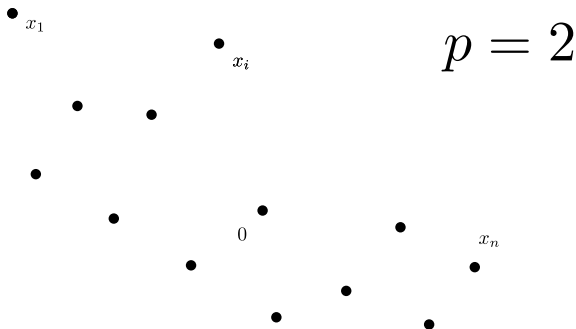
ACP

Définition

Interprétation

ACP

On observe n points x_1, \dots, x_n dans \mathbb{R}^p , ainsi on crée une matrice $X = [x_1, \dots, x_n]^\top$ matrice $n \times p$: n observations (lignes), p features (colonnes)



Rem: on suppose que l'on a recentré les points pour qu'ils aient une moyenne nulle $X \leftarrow [x_1 - \bar{x}_n, \dots, x_n - \bar{x}_n]^\top = X - \mathbf{1}_n \bar{x}_n^\top$ (il est aussi conseillé de mettre à l'échelle et d'avoir un écart-type

Analyse en Composante Principale, ACP

( : *Principal Component Analysis, PCA*)

Paramètre k : nombre d'axes pour représenter un nuage de n points (x_1, \dots, x_n) , représentés par les lignes de $X \in \mathbb{R}^{n \times p}$
Cette méthode **compresse** le nuage de points de dimension p en un nuage de dimension k

L'ACP (de niveau k) consiste à effectuer la SVD de X , et à ne garder que les k axes principaux pour représenter le nuage.

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

On appelle **axes principaux** les k vecteurs $\mathbf{v}_1, \dots, \mathbf{v}_k$, et en général $k \ll p$ (voir $k = 2$ la visualisation)

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

ACP

Définition

Interprétation

ACP : Interprétation

Construction récursive : définir les axes principaux de manière récursive

On peut voir que l'axe principal (normalisé) \mathbf{v}_1 est la solution du problème :

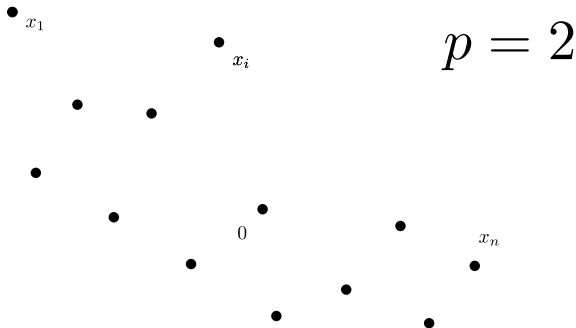
$$\mathbf{v}_1 = \arg \max_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \mathbf{v}^\top X^\top X \mathbf{v} = \arg \max_{\mathbf{v} \in \mathbb{R}^p, \|\mathbf{v}\|=1} \|X \mathbf{v}\|^2$$

Puis par récurrence, on définit le k^{e} axe pour qu'il soit orthogonal aux axes principaux précédents :

$$\mathbf{v}_k = \arg \max_{\mathbf{v} \in \mathbb{R}^p, \mathbf{v}^\top \mathbf{v}_1 = \dots = \mathbf{v}^\top \mathbf{v}_{k-1} = 0, \|\mathbf{v}\|=1} \|X \mathbf{v}\|^2$$

ACP

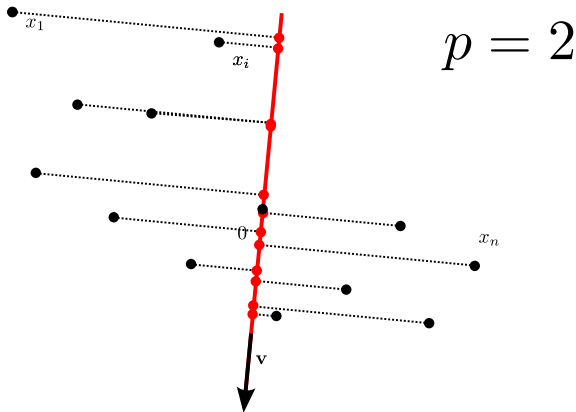
X matrice $n \times p$: n observations, p dimensions (*features*)



- ▶ le premier axe maximise la variance des données projetées sur l'axe porté par ce vecteur
- ▶ le deuxième axe est celui orthogonal au premier, de variance projetée maximale
- ▶ etc.

ACP

X matrice $n \times p$: n observations, p dimensions (*features*)

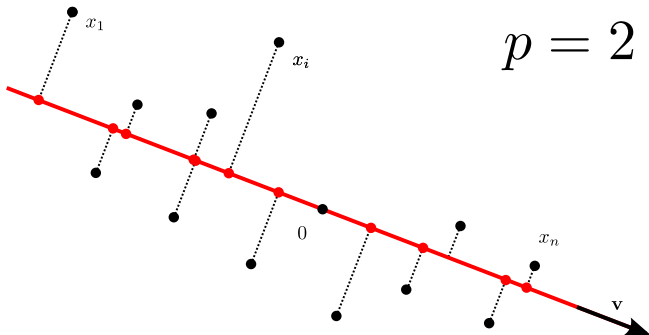


- ▶ le premier axe maximise la variance des données projetées sur l'axe porté par ce vecteur
- ▶ le deuxième axe est celui orthogonal au premier, de variance projetée maximale
- ▶ etc.

ACP

X matrice $n \times p$: n observations, p dimensions (*features*)

$$p = 2$$



- ▶ le premier axe maximise la variance des données projetées sur l'axe porté par ce vecteur
- ▶ le deuxième axe est celui orthogonal au premier, de variance projetée maximale
- ▶ etc.

Premier axe principal

Maximiser la fonction objectif suivante en \mathbf{v} :

$$\mathcal{L}(\mathbf{v}, \lambda) = (X\mathbf{v})^\top (X\mathbf{v}) - \lambda(\mathbf{v}^\top \mathbf{v} - 1) = \mathbf{v}^\top X^\top X\mathbf{v} - \lambda(\mathbf{v}^\top \mathbf{v} - 1)$$

λ : multiplicateur de Lagrange

Conditions d'optimalité du premier ordre en un extremum

$$\frac{\partial \mathcal{L}(\mathbf{v}_1, \lambda)}{\partial \mathbf{v}} = 0 \Leftrightarrow X^\top X\mathbf{v}_1 = \lambda\mathbf{v}_1$$

La matrice de Gram $X^\top X$ est diagonalisable (symétrique) donc si \mathbf{v}_1 est un extremum alors c'est un vecteur propre.

Rem: si l'on normalise \mathbf{v}_1 pour que $\|\mathbf{v}_1\| = 1$, alors $\lambda = \mathbf{v}_1^\top X^\top X\mathbf{v}_1$ et donc \mathbf{v}_1 est le vecteur propre de valeur propre λ maximale

Nouvelle représentation des données

- ▶ Les axes (de direction) $\mathbf{v}_1, \dots, \mathbf{v}_p \in \mathbb{R}^p$ sont appelés **axes principaux** ou **axes factoriels**, les nouvelles variables $\mathbf{c}_j = X\mathbf{v}_j, j = 1, \dots, p$ sont appelées **composantes principales**

Nouvelle représentation :

- ▶ La matrice XV_k (avec $V_k = [\mathbf{v}_1, \dots, \mathbf{v}_k]$) est la matrice représentant les données dans la base des k premiers vecteurs propres

Reconstruction dans l'espace original (débruiter) :

- ▶ Reconstruction "parfaite" pour $\mathbf{x} \in \mathbb{R}^p$: $\mathbf{x} = \sum_{j=1}^p (\mathbf{x}^\top \mathbf{v}_j) \mathbf{v}_j$
- ▶ Reconstruction avec perte d'information : $\hat{\mathbf{x}} = \sum_{j=1}^k (\mathbf{x}^\top \mathbf{v}_j) \mathbf{v}_j$

Références I

- ▶ G. H. Golub and C. F. van Loan.

Matrix computations.

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