Chapitre 2: Régression en grande dimension
Pénalisations Ridge et Lasso

2 Ingénierie des Données et Systèmes Décisionnels

Ecole Nationale d'Electronique et des Télécommunications de Sfax

07 Septembre 2022

Plan

Régression linéaire en grande dimension

- On observe $(Y_i, x_i)_{1 \leq i \leq n}$ avec $Y_i \in \mathbb{R}$ et $x_i \in \mathbb{R}^{d+1}$. Pour tout i,

$$Y_i = x_i^t \beta^* + \varepsilon_i,$$

avec

  - β^* paramètre inconnu,
  - $(\varepsilon_i)_i$ i.i.d tque $\mathbb{E}(\varepsilon_i) = 0$
  - $(x_i)_{1 \leq i \leq n}$ déterminsite.

- Écriture matricielle

$$Y = X\beta^* + \varepsilon,$$

avec

  - $Y = (Y_1, Y_2, \ldots, Y_n)^t$,
  - $\beta^* \in \mathbb{R}^{d+1}$,
  - X matrice non aléatoire de taille $n \times (d+1)$
  - $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

- Estimateur des moindres carrés est défini par

$$\hat{\beta}^{MC} \in \arg \min_{\beta \in \mathbb{R}^{d+1}} \| Y - X\beta^* \|^2.$$

- Si X de rang plein, alors X^tX est une matrice d'ordre $p+1$ inversible et l'estimateur des moindres carrés existe et est unique

$$\hat{\beta}^{MC} = (X^tX)^{-1}X^tY.$$

Remarques

- Si X^tX est une matrice inversible $(d+1 \leq n)$, alors
  - $\mathbb{E}(\hat{\beta}^{MC}) = \beta^*$.
  - $\mathbb{E}(\| \hat{\beta}^{MC} - \beta^* \|^2) = \sigma^2 tr((X^tX)^{-1})$.
  - En paraticulier, si $X^tX = nI_{d+1}$, alors

$$\mathbb{E}(\| \hat{\beta}^{MC} - \beta^* \|^2) = \frac{(d+1)\sigma^2}{n}.$$

Remarque

On veut connaître l'impacte de la dimension d sur l'erreur d'estimation.
En fait, si les colonnes de X sont orthonormales, alors l'application $d \longmapsto tr((X^tX)^{-1})$ est croissante.
$\Longrightarrow$ L'erreur augmente avec la dimension d.
$\Longrightarrow$ On veut essayer d'introduire un nouvel estimateur qui fonctionne en grande dimension.

Plan

Définition

L'estimateur Ridge dans le modèle $Y = X\beta^* + \varepsilon$ est défini par

$$\hat{\beta}_\lambda^R \in \arg \min_{\beta \in \mathbb{R}^{d+1}} \left\{ \| Y - X\beta \|^2 + \lambda \| \beta \|^2 \right\},$$

où $\lambda > 0$ à choisir.

Remarque

$\lambda > 0$ est un paramètre (coefficient de pénalité) qui permet de contrôler l'impact de la pénalité $\lambda \| \beta \|^2$. Il est à fixer.

Explication

L'idée est d'ajouter une contrainte sur les coefficients β_i lors de la modélisation pour maîtriser l'amplitude de leurs valeurs (i.e., pour éviter qu'elles partent dans tous les sens).

$$\min_{\beta} \{ \| Y - X\beta \|^2 \} = \min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} \beta_j x_i^j)^2 \right\},$$

sous la contrainte

$$\| \beta \|_2^2 = \sum_{j=0}^{p} \beta_j^2 \leq \tau,$$

où $\tau \geq 0$ est un un paramètre à fixer et x_i^j est la valeur de la variable X_j pour la $i$ème observation.

Remarques

- On parle de shrinkage (rétrécissement): on rétrécit les plages de valeurs que peuvent prendre les paramètres estimés.
- Les variables x_j doivent être centrées et réduites pour éviter que les variables à forte variance aient trop d'influence.
- Si $\tau \longrightarrow 0$, alors $\beta_j \longrightarrow 0$: Les variances des coefficients estimés sont nulles.
- Si $\tau \longrightarrow +\infty$, alors $\hat{\beta}^R = \hat{\beta}^{MC}$.

Définition

L'estimateur Ridge dans le modèle $Y = X\beta^* + \varepsilon$ est défini par

$$\hat{\beta}_\lambda^R \in \arg\min_{\beta \in \mathbb{R}^{d+1}} \left\{ \| Y - X\beta \|^2 + \lambda \| \beta \|^2 \right\},$$

où $\lambda > 0$ à choisir.

Questions

1. Est ce que ce problème de minimisation admet une solution?
2. La solution est-elle unique?

Proposition 1

L'estimateur de Ridge s'écrit sous la forme

$$\hat{\beta}_\lambda^R = (X^t X + \lambda I_{d+1})^{-1} X^t Y,$$

où I_{d+1} est la matrice unité d'ordre $d+1$.

Démonstration de la Proposition 1 en TD

Remarques

- $\hat{\beta}_\lambda^R$ existe et est unique.
- Pour $\lambda = 0$, $\hat{\beta}_\lambda^R = \hat{\beta}^{MC}$.
- L'estimateur Ridge est toujours défini. En effet, la matrice $X^t X + \lambda I_{d+1}$ (avec $\lambda > 0$) est toujours inversible puisqu'elle est définie positive. Pour tout $u \in \mathbb{R}^{d+1} \backslash \{0_{\mathbb{R}^{d+1}}\}$,

$$u^t(X^t X + \lambda I_{d+1})u = u^t X^t X u + \lambda u^t u = \| Xu \|^2 + \| \lambda u \|^2 > 0.$$

Proposition 2

$\hat{\beta}_\lambda^R$ est l'unique solution du problème d'optimisation sous contrainte

$$\hat{\beta}_\lambda^R = \arg \min_{\beta \in \mathbb{R}^{d+1},\ \|\beta\|^2 \leq M_\lambda} \left\{ \| Y - X\beta \|^2 \right\},$$

où

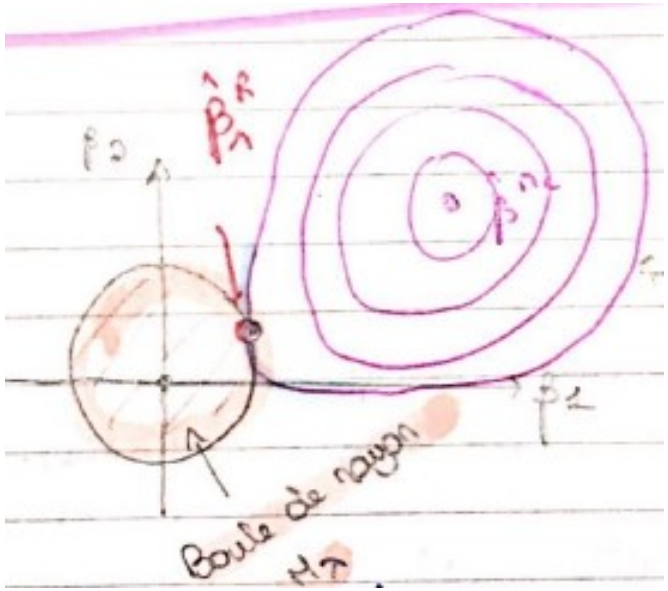
$$M_\lambda = \| (X^t X + \lambda I_{d+1})^{-1} X^t Y \|^2 .$$

Démonstration de la Proposition 2 en TD

Remarques

- On ne minimize pas le critère sur $\mathbb{R}^{d+1}$ mais sur la boule de rayon M_λ.
- On obtient l'estimateur de Ridge sur les bords de cettte boule.
- On trace les lignes de niveaux i.e., $\{\beta;\ \| Y - X\beta \| = L\}$.
- Ce contour a pour minimum l'estimateur des moindre carrés.
- L'intersection de la boule et les lignes de niveaux donne $\hat{\beta}_\lambda^R$.

Exemple $d = 2$



Proposition (Biais de $\hat{\beta}_\lambda^R$)

$$\mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* = -\lambda(X^tX + \lambda I_{d+1})^{-1}\beta^*. \quad (1)$$

Démonstration

On a

$$\hat{\beta}_\lambda^R = (X^tX + \lambda I_{d+1})^{-1}X^tY,$$

donc

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_\lambda^R) &= \mathbb{E}((X^tX + \lambda I_{d+1})^{-1}X^tY) \\
&= (X^tX + \lambda I_{d+1})^{-1}X^t\mathbb{E}(Y) \\
&= (X^tX + \lambda I_{d+1})^{-1}X^tX\beta^* (\text{car } \mathbb{E}(Y) = \mathbb{E}(X\beta^* + \varepsilon) = X\beta^*) \\
&= (X^tX + \lambda I_{d+1})^{-1}(X^tX + \lambda I_{d+1})\beta^* - \lambda(X^tX + \lambda I_{d+1})^{-1}\beta^* \\
&= \beta^* - \lambda(X^tX + \lambda I_{d+1})^{-1}\beta^*.
\end{aligned}$$

D'où le résultat.

Remarques

- Comme $\lambda > 0$, alors $\mathbb{E}(\hat{\beta}_\lambda^R) \neq \beta^* \Longrightarrow \hat{\beta}_\lambda^R$ est un estimateur biaisé.
- Si $\lambda \longrightarrow 0$, alors $\mathbb{E}(\hat{\beta}_\lambda^R) = \beta^* \Longrightarrow \hat{\beta}_\lambda^R$ est un estimateur sans biais.

Biais de $\hat{\beta}_\lambda^R$

$$\begin{aligned} \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|^2 &= \| \beta^* - \lambda(X^t X + \lambda I_{d+1})^{-1}\beta^* - \beta^* \|^2 \\ &= \lambda^2 \| (X^t X + \lambda I_{d+1})^{-1}\beta^* \|^2 . \end{aligned}$$

Remarque (Cas particulier)

Si $X^tX = nI_{d+1}$, alors

$$\| \textit{Biais} \|^2 = \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|^2 = \frac{\lambda^2}{(n+\lambda)^2} \| \beta^* \|^2 .$$

Pour avoir un bon biais il faut prendre λ petit. En effet,

$$\lim_{\lambda \longrightarrow 0} \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|^2 = 0 \Longrightarrow \text{ Overfitting}$$

et si λ est grand on aura

$$\lim_{\lambda \longrightarrow +\infty} \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|^2 = \| \beta^* \|^2 \Longrightarrow \text{ Underfitting.}$$

Théorème (Variance de $\hat{\beta}_\lambda^R$)

$$\mathbb{V}(\hat{\beta}_\lambda^R) = \sigma^2 (X^tX + \lambda I_{d+1})^{-2} XX^t. \tag{2}$$

Démonstration (Variance de $\hat{\beta}_\lambda^R$)

On a

$$\hat{\beta}_\lambda^R = (X^t X + \lambda I_{d+1})^{-1} X^t Y,$$

donc

$$\begin{aligned} \mathbb{V}(\hat{\beta}_\lambda^R) &= \mathbb{V}((X^t X + \lambda I_{d+1})^{-1} X^t Y) \\ &= (X^t X + \lambda I_{d+1})^{-1} X^t \mathbb{V}(Y) ((X^t X + \lambda I_{d+1})^{-1} X^t)^t \\ &= \sigma^2 (X^t X + \lambda I_{d+1})^{-1} X^t X (X^t X + \lambda I_{d+1})^{-1} \end{aligned} \tag{3}$$

Comme la matrice $X^t X$ est symétrique positive, alors $X^t X$ est diagonalisable. Ainsi

- Il existe une matrice P orthogonale ($P^{-1} = P^t$)
- Il existe une matrice D diagonale $D = diago(\alpha_1, \alpha_2, \ldots, \alpha_{d+1})$

$$X^t X = PDP^t.$$

Démonstration (Variance de $\hat{\beta}_\lambda^R$)

Il s'en suit que

$$X^tX + \lambda I_{d+1} = PDP^t + \lambda I_{d+1} = PDP^t + \lambda PP^t = P(D + \lambda I_{d+1})P^t$$

Ceci avec (7) et le fait que $P^t = P^{-1}$ donne

$$\begin{aligned}
\mathbb{V}(\hat{\beta}_\lambda^R) &= \sigma^2(P(D+\lambda I_{d+1})P^t)^{-1}PDP^t(P(D+\lambda I_{d+1})P^t)^{-1} \\
&= \sigma^2(P^t)^{-1}(D+\lambda I_{d+1})^{-1}P^{-1}PDP^t(P^t)^{-1}(D+\lambda I_{d+1})^{-1}P^{-1} \\
&= \sigma^2 P(D+\lambda I_{d+1})^{-1}D(D+\lambda I_{d+1})^{-1}P^t \\
&= \sigma^2 P(D+\lambda I_{d+1})^{-2}DP^t \text{ car } D \text{ est diagonale} \\
&= \sigma^2 P(D+\lambda I_{d+1})^{-2}P^tPDP^t \\
&= \sigma^2(X^tX+\lambda I_{d+1})^{-2}XX^t.
\end{aligned}$$

Proposition

$$\mathbb{E}\left(\| \hat{\beta}_\lambda^R - \mathbb{E}(\hat{\beta}_\lambda^R) \|^2\right) = tr\left(\mathbb{V}\left(\hat{\beta}_\lambda^R\right)\right).$$

Démonstration

On a par définition

$$\begin{aligned}\mathbb{E}(\| \hat{\beta}_\lambda^R - \mathbb{E}(\hat{\beta}_\lambda^R) \|^2) &= \sum_{i=1}^{d+1} \mathbb{E}\left(\left(\left(\hat{\beta}_\lambda^R\right)_i - \left(\mathbb{E}\left(\hat{\beta}_\lambda^R\right)\right)_i\right)^2\right) \\ &= \sum_{i=1}^{d+1} \mathbb{V}((\hat{\beta}_\lambda^R)_i) = tr\left(\mathbb{V}\left(\hat{\beta}_\lambda^R\right)\right).\end{aligned}$$

Evaluation de l'erreur quadratique

$$MSE = \mathbb{E}\left(\| \hat{\beta}_\lambda^R - \beta^* \|^2\right) = \mathbb{E}\left(\| \hat{\beta}_\lambda^R - \mathbb{E}(\hat{\beta}_\lambda^R) \|^2\right) + \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|^2 .$$

$$= tr\left(\mathbb{V}\left(\hat{\beta}_\lambda^R\right)\right) + \| \underbrace{\mathbb{E}(\hat{\beta}_\lambda^R) - \beta^*}_{Biais} \|^2 . \quad (4)$$

$\Longrightarrow$ Oublier le réflexe

$$Erreur = variance + Biais^2$$

(cas vectoriel).

En insérant (6) et (8) dans (4), on obtient

$$\mathbb{E}(\| \hat{\beta}_\lambda^R - \beta^* \|^2) = \sigma^2 tr((X^t X + \lambda I_{d+1})^{-2} X X^t) + \lambda^2 \| (X^t X + \lambda I_{d+1})^{-1} \beta^*$$

Remarque (cas particulier)

Si $X^tX = nI_{d+1}$, alors $X^tX + \lambda I_{d+1} = (n+\lambda)I_{d+1}$ et

$$\begin{aligned}\mathbb{E}(\| \hat{\beta}_\lambda^R - \beta^* \|^2) &= \sigma^2 tr((n+\lambda)^{-2} nI_{d+1}) + \lambda^2 \| ((n+\lambda)I_{d+1})^{-1}\beta^* \|^2 \\ &= \underbrace{\sigma^2 \frac{n}{(n+\lambda)^2}(d+1)}_{Variance} + \underbrace{\frac{\lambda^2}{(n+\lambda)^2} \| \beta^* \|^2}_{Biais^2}.\end{aligned}$$

Il s'en suit que

$$\lim_{\lambda \longrightarrow +\infty} (Variance) = 0 \text{ et } \lim_{\lambda \longrightarrow +\infty} (Biais^2) = \| \beta^* \|^2 .$$

$\Longrightarrow$ Variance petite et Biais grand (Underfitting).

$$\lim_{\lambda \longrightarrow 0} (Variance) = \frac{\sigma^2}{n}(d+1) \text{ et } \lim_{\lambda \longrightarrow +\infty} (Biais^2) = 0$$

$\Longrightarrow$ Variance grande et Biais petit (Overfitting)

Remarque (cas particulier)

- Il faut prendre λ ni trop grand ni trop petit.
- On a ajouté λ pour garantir que la matrice soit inversible, cependant il faut faire le bon choix de λ.

Exercice en TD

$$D = diago(\alpha_1, \alpha_2, \ldots, \alpha_{d+1}) \text{ et } X^tX = PDP^t.$$

1. Montrer que $\mathbb{E}\left(\| \mathbb{E}(\hat{\beta}_\lambda^R) - \hat{\beta}_\lambda^R \|_2^2\right) = \sigma^2 \sum_{j=1}^{d+1} \frac{\alpha_j}{(\lambda + \alpha_j)^2}$.
2. Montrer que $\| Biais \|_2^2 = \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|_2^2 = \sum_{j=1}^{d+1} \frac{\lambda^2}{(\lambda + \alpha_j)^2} (P^t\beta^*)_j^2$.

Remarque (Variance)

- On pose

$$V_\lambda^{(R)} = \mathbb{E}\left(\| \mathbb{E}(\hat{\beta}_\lambda^R) - \hat{\beta}_\lambda^R \|_2^2 \right) = \sigma^2 \sum_{j=1}^{d+1} \frac{\alpha_j}{(\lambda + \alpha_j)^2}.$$

Alors

$$\lim_{\lambda \longrightarrow +\infty} V_\lambda^{(R)} = 0.$$

De plus

$$\sum_{j=1}^{d+1} \frac{\alpha_j}{(\lambda + \alpha_j)^2} = \sum_{j=1}^{d+1} \frac{\alpha_j}{(\lambda + \alpha_j)^2} \mathbf{1}_{\{\alpha_j > 0\}} + \underbrace{\sum_{j=1}^{d+1} \frac{\alpha_j}{(\lambda + \alpha_j)^2} \mathbf{1}_{\{\alpha_j < 0\}}}_{=0}.$$

Ainsi

$$\lim_{\lambda \longrightarrow 0} V_\lambda^{(R)} = \sigma^2 \sum_{j=1}^{d+1} \frac{1}{\alpha_j} \mathbf{1}_{\{\alpha_j > 0\}} \neq 0.$$

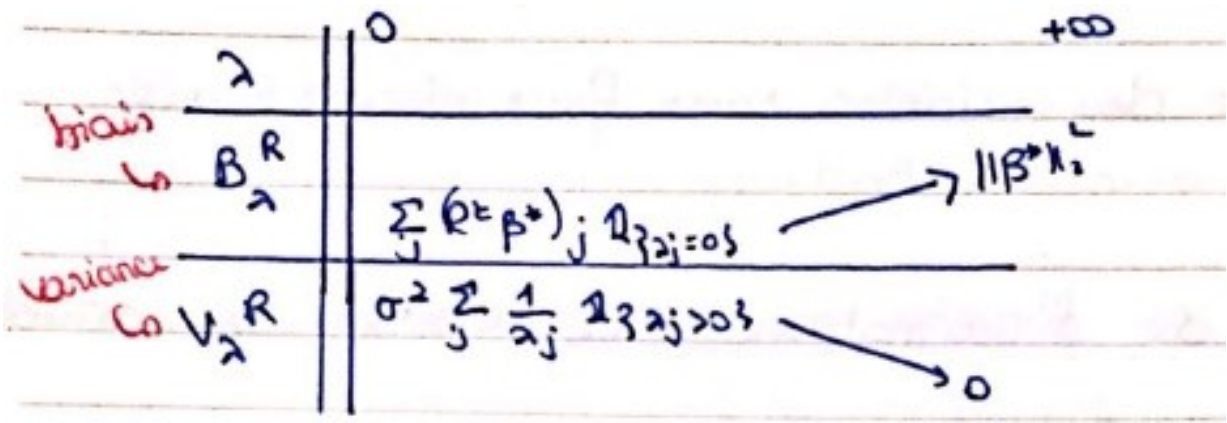
$\Longrightarrow \lambda \longmapsto V_\lambda^{(R)}$ est décroissante

$\Longrightarrow$ On n'a pas intérêt à choisir λ assez petit car dans ce cas, la variance converge vers une constante.

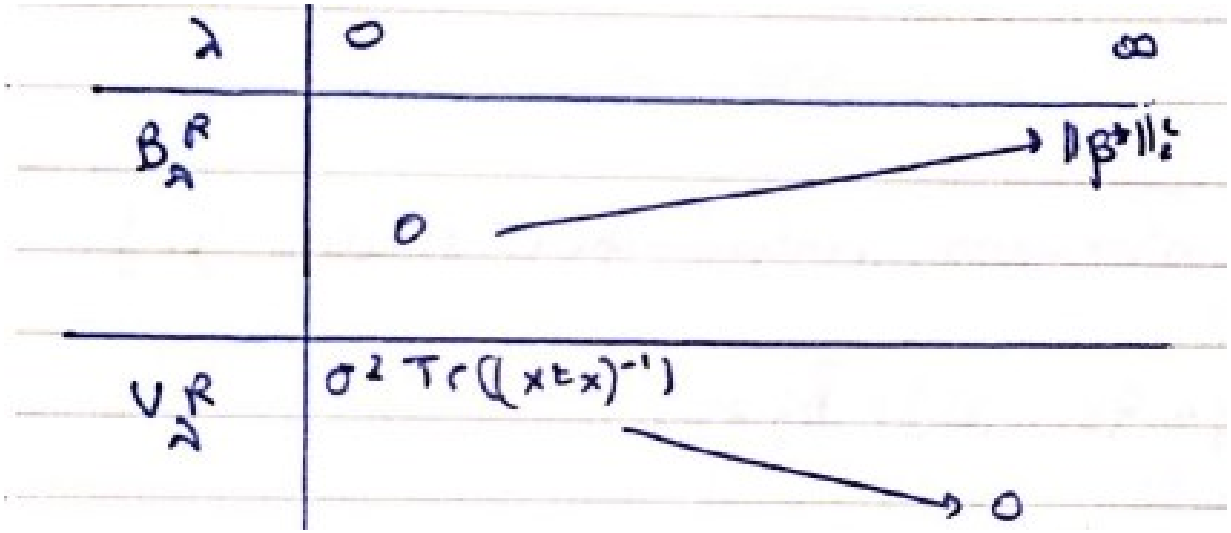
Remarques (Le Biais)

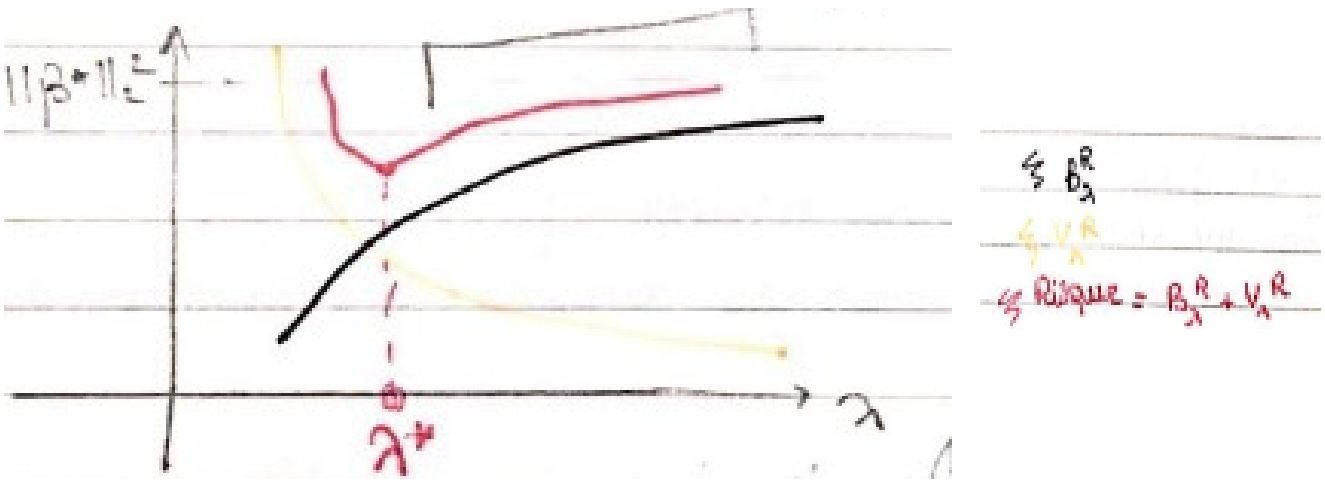
- $$B_\lambda^R = \| \mathbb{E}(\hat{\beta}_\lambda^R) - \beta^* \|_2^2 = \sum_{j=1}^{d+1} \frac{\lambda^2}{(\lambda+\alpha_j)^2}(P^t\beta^*)_j^2.$$
$$= \sum_{j=1}^{d+1} \frac{\lambda^2}{(\lambda+\alpha_j)^2}(P^t\beta^*)_j^2 \mathbf{1}_{\{\alpha_j=0\}} + \sum_{j=1}^{d+1} \frac{\lambda^2}{(\lambda+\alpha_j)^2}(P^t\beta^*)_j^2 \mathbf{1}_{\{\alpha_j>0\}}.$$
- $$\lim_{\lambda \longrightarrow 0} B_\lambda^R = \sum_{j=1}^{d+1}(P^t\beta^*)_j^2 \mathbf{1}_{\{\alpha_j=0\}}.$$
- $$\lim_{\lambda \longrightarrow +\infty} B_\lambda^R = \lim_{\lambda \longrightarrow +\infty} \sum_{j=1}^{d+1} \frac{\lambda^2}{(\lambda+\alpha_j)^2}(P^t\beta^*)_j^2 = \sum_{j=1}^{d+1}(P^t\beta^*)_j^2$$
$$= \| P^t\beta^* \|_2^2 = (P^t\beta^*)^t P^t\beta^* = (\beta^*)^t\beta^* = \| \beta^* \|_2^2 .$$

Cas où X^tX est non inversible.



Cas où X^tX est inversible.





- *Risque* $=$ *Variance* $+$ *Biais*^{2}.
- λ^* : La valeur de λ qui minimise le risque, alors

$$\lambda^* \in \arg\min_{\lambda>0} \left(\mathbb{E} \left(\| \hat{\beta}_\lambda^R - \beta^* \|^2 \right) \right).$$

- λ^* dépend de σ^2 qui est inconnu.
- Si $\lambda < \lambda^* \Longrightarrow$ Sur-apprentissage (Ovrefitting).
- Si $\lambda > \lambda^* \Longrightarrow$ Sous-apprentissage (Underfitting).

Remarques

- On désigne par $\Sigma_\lambda = \mathbb{V}(\hat{\beta}_\lambda^R)$.
  Si pour tout j, $(\Sigma_\lambda)_{jj} \neq 0$, alors $\mathbb{V}((\hat{\beta}_\lambda^R)_j) \neq 0$. D'où $(\hat{\beta}_\lambda^R)_j \neq 0$ p.s.
- En général, l'estimateur Ridge sélectionne toutes les variables (ce qui n'est pas bien).

Question

Comment faire pour sélectionner des variables sans faire des tests statistiques par variable? i.e.,
Soit le modèle linéaire suivant

$$Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \ldots + \beta_d X^d + \varepsilon. \tag{5}$$

Pour tester la significativité de la variable X^j pour ce modèle , on considère le test

$$\mathcal{H}_0 : \ \beta_j = 0 \text{ contre } \mathcal{H}_1 \ \beta_j \neq 0.$$

$\Longrightarrow$ C'est l'objectif de la régression LASSO: Sélection des variables.

Plan

Introduction

En statistique, le LASSO est une méthode de contraction des coefficients de la régression développée par Robert Tibshirani en 1996 "Regression shrinkage and selection via the lasso". Le nom est un acronyme anglais: Least Absolute Shrinkage and Selection Operator.

- La sélection des variables est faite en mettant à 0 certaines coordonnées $\hat{\beta}_j$ dans l'équation (5).
- On dira que la variable est sélectionnée si $\hat{\beta}_j \neq 0$ et qu'elle est non sélectionnée sinon.

$\Longrightarrow$ Si $\hat{\beta}_j = 0$, alors la variable X^j n'est pas utilisée pour prédire la valeur de la variable de sortie Y.

Notation

On note

- $supp(\hat{\beta}) = \left\{ j \in \{1, 2, \ldots, d\};\ \hat{\beta}_j \neq 0 \right\}$.
- $J(\hat{\beta}) = card\{supp(\hat{\beta})\}$: Nombre de paramètres non nul.

Première idée

Minimiser $\| Y - X\beta \|^2$ sous la contrainte $J(\beta) < s$, où s est le nombre maximal de variables qu'on souhaite sélectionner.

Problème

La fonction $\beta \longmapsto J(\beta)$ n'est pas continue et n'est pas convexe $\Longrightarrow$ Problème de recherche de minimum.

Idée solution

Dans la contrainte $J(\beta) < s$ ($J(\beta) = \sum_j \mathbf{1}_{\{\beta_j \neq 0\}}$), on remplace $J(\beta)$ par $\| \beta \|_1 = \sum_j | \beta_j |$.

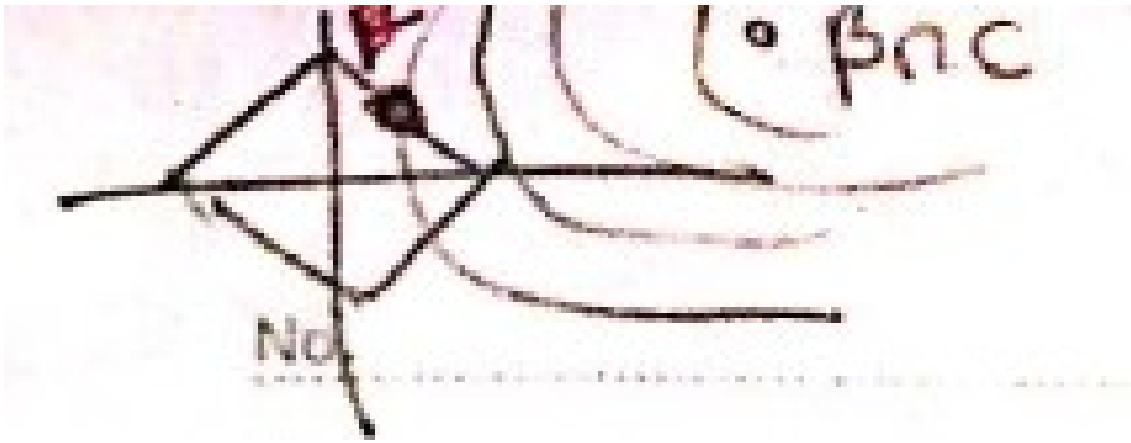
Définition

On appelle estimateur LASSO toute solution du problème

$$\arg \min_{\beta;\ \|\beta\|_1 \leq M} \left(\| Y - X\beta \|^2 \right) = \arg \min_{\beta;\ \|\beta\|_2 \leq M_\lambda} \left(\left(\sum_i Y_i - x_i^t \beta \right)^2 \right).$$

Remarque (Exemple graphique)

- $\beta \longmapsto \| \beta \|_1 = | \beta_1 | + | \beta_2 |$.
- On trace les lignes de niveaux L $\{\beta;\ \| Y - X\beta \|^2 = L\}$



Le problème $\arg \min_{\beta;\ \|\beta\|_1 \leq M} \left(\| Y - X\beta \|^2 \right)$ admet au moins une solution (mais pas unique).

Proposition 3.

L'estimateur LASSO est solution du critère pénalisé

$$\arg \min_{\beta \in \mathbb{R}^{d+1}} (\| Y - X\beta \|^2 + \lambda \| \beta \|_1),$$

où $\| \beta \|_1 = \sum_{j=1}^{d+1} | \beta_j |$ et $\lambda > 0$.

Démonstration de la Proposition 3 en TD

Remarque

L'estimateur LASSO est défini par

$$\hat{\beta}_\lambda^L \in \arg \min_{\beta \in \mathbb{R}^{d+1}} (\| Y - X\beta \|^2 + \lambda \| \beta \|_1). \quad (6)$$

Remarque

Comme l'application $\beta \longmapsto \sum_{j=1}^{n}(Y_j - X_j^t\beta)^2$ est continue sur le compacte $\{\beta \in \mathbb{R}^{d+1};\ \| \beta \|_1 \leq M\}$, alors l'estimateur LASSO existe toujours

Définition (compacte)

Soit (E, d) un espace métrique. Une partie K de E est dite compacte si, de toute suite $(u_n)_n$ d'éléments de K, on peut extraire une sous-suite convergente vers un élément de K.

Remarques

- Toute partie compacte de E est fermée et bornée.
- Un segment $[a, b]$ est une partie compacte de $\mathbb{R}$.
- En particulier, les parties compactes de $\mathbb{R}$ ou de $\mathbb{C}$ sont les parties fermées et bornées.

Problème

On rappelle que $\| \beta \|_1 = \sum_{j=1}^{d+1} | \beta_j |$. La fonction $x \longmapsto | x |$ non dérivable en 0, pour countourner ce problème, on définit la sous-différentielle.

Définition (sous-différentielle)

Soit $x_0 \in \mathbb{R}^n$, on définit

$$\partial N_1(x_0) = \left\{ x \in \mathbb{R}^n ;\ x_j = \left\{ \begin{array}{ccc} signe(x_{0j}) & si & x_{0j} \neq 0 \\ [-1;1] & si & x_{0j} = 0 \end{array} \right\} \right.,$$

où x_j désigne la $j^{\text{ième}}$ composante du vecteur x.

Remarque

L'ensemble $\partial N_1(x_0)$ est le sous-différentiel de la fonction $x \longmapsto \| x \|_1$ au point $x_0 \in \mathbb{R}^n$.

Remarque

- Une sous-dérivée d'une fonction convexe $f : I \longrightarrow \mathbb{R}$ en un point x_0 de l'intervalle ouvert I est un nombre réel s vérifiant pour tout x dans I,

$$f(x) \geq f(x_0) + s(x - x_0).$$

- Si x_0 est dans l'intérieur de I, l'ensemble des sous-dérivées en x_0 est un intervalle fermé non vide, donc de la forme $[a, b]$, avec

$$a = \lim_{x \uparrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

et

$$b = \lim_{x \downarrow x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

- a et b sont finis et vérifient $a \leq b$.
- L'ensemble $[a, b]$ de toutes les sous-dérivées est dit le sous-différentiel de la fonction f en x_0.

Exemple

Considérons la fonction $f(x) = \mid x \mid$.

- Le sous-différentiel à l'origine est l'intervalle $[-1, 1]$.
- Le sous-différentiel en tout point $x_0 < 0$ est le singleton $\{-1\}$.
- Le sous-différentiel en tout point $x_0 > 0$ est le singleton $\{1\}$.

Exemple (n=2)

On pose $x_0 = \begin{pmatrix} -10 \\ 0 \end{pmatrix}$,

$$\partial N_1(x_0) = \left\{ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} ;\ x_1 = -1 \text{ et } x_2 \in [-1; 1] = \{-1\} \times [-1; 1] \right\}$$

Proposition 4.

Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ convexe et dérivable sur un ouvert U de $\mathbb{R}^n$ et soit $x_0 \in \arg\min_{x \in \mathbb{R}^n} \{f(x) + \lambda \parallel x \parallel_1\}$ alors il existe $\delta \in \partial N_1(x_0)$ tel que

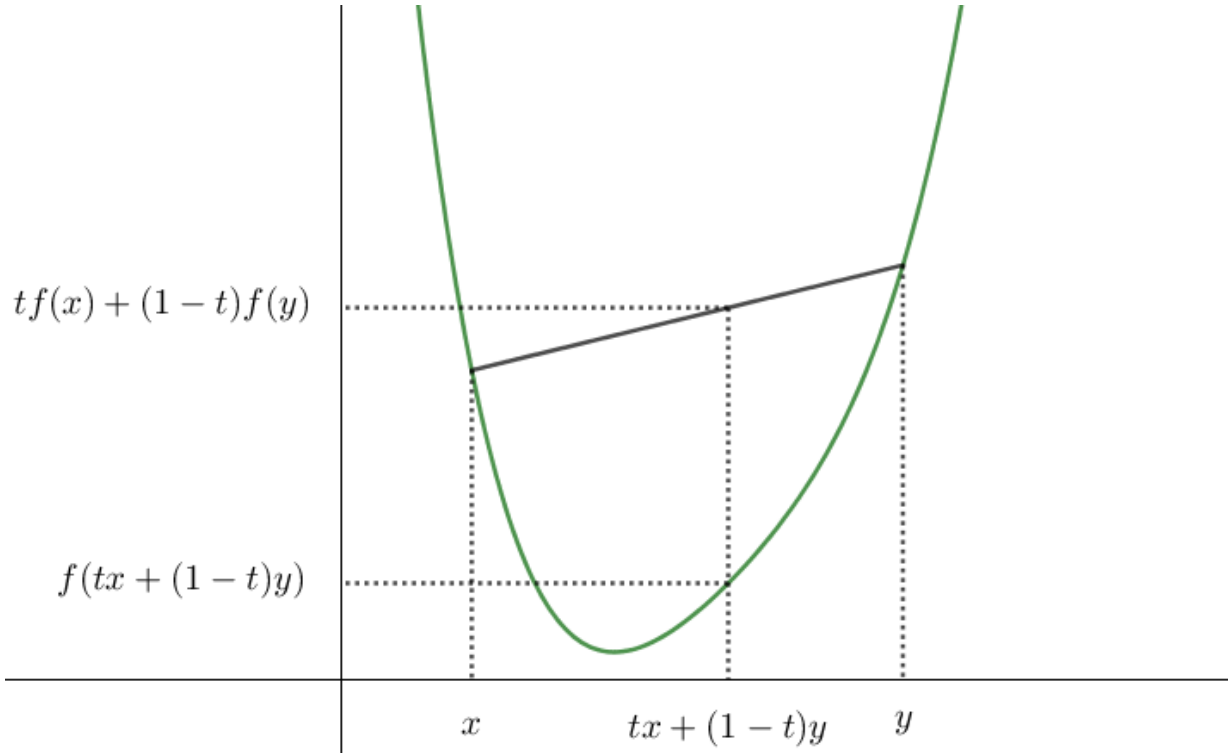
$$\nabla f(x_0) + \lambda\delta = 0.$$

Définition (fonction convexe)

Une fonction f d'un intervalle réel I vers $\mathbb{R}$ est dite convexe si pour tous x et y de I et tout t dans $[0;1]$ on a :

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Graphiquement (fonction convexe)



Remarque

Une fonction est convexe si et seulement si sa courbe est au-dessous de ses cordes.

Exemples (fonction convexe)

- $x \longmapsto x^2$
- $x \longmapsto e^x$

Proposition

- Si f est dérivable sur I, alors f est convexe si et seulement si f' est croissante.
- Si f est deux fois dérivable sur I, alors f est convexe si et seulement si $f'' \geq 0$.

Proposition

Il existe $\hat{\delta} \in \partial N_1(\hat{\beta}_\lambda^L)$ tel que l'estimateur LASSO $\hat{\beta}_\lambda^L$ défini par (6) vérifie l'équation

$$X^t X \hat{\beta}_\lambda^L = X^t Y - \frac{\lambda}{2}\hat{\delta}.$$

Démonstartion

On applique la Proposition 4, pour la fonction $f(\beta) = \sum_{i=1}^{n}(Y_i - x_i^t\beta)^2$, on obtient

$$\nabla f(\beta) = -2X^t(Y - X\beta). \tag{7}$$

Comme

$$\hat{\beta}_\lambda^L = \in \arg\min_{\beta\in\mathbb{R}^{d+1}}(f(\beta) + \lambda \parallel \beta \parallel_1),$$

signifie qu'il existe $\hat{\delta} \in \partial N_1(\hat{\beta}_\lambda^L)$ tel que

$$\nabla f(\hat{\beta}_\lambda^L) + \lambda\hat{\delta} = 0.$$

Ceci avec (7), donne

$$\begin{aligned} & 2X^tY - 2X^tX\hat{\beta}_\lambda^L = \lambda\hat{\delta}. \\ \iff \quad & 2X^tX\hat{\beta}_\lambda^L = 2X^tY - \lambda\hat{\delta}. \\ \iff \quad & X^tX\hat{\beta}_\lambda^L = X^tY - \frac{\lambda\hat{\delta}}{2}. \end{aligned}$$

Remarques

- Si X^tX est inversible, alors l'estimateur LASSO est unique et vérifie

$$\hat{\beta}_\lambda^L = (X^tX)^{-1}(X^tY - \frac{\lambda\hat{\delta}}{2}).$$

- La pénalisation LASSO fait que plusieurs coefficients soient nuls $\Longrightarrow$ On utilise LASSO pour la sélection des variables.

Proposition 5 (Calcul de $\hat{\beta}_\lambda^L$ lorsque $X^tX = nI_{d+1}$)

Si $X^tX = nI_{d+1}$ et $(\hat{\beta}_\lambda^L)_j \neq 0$, alors

$$(\hat{\beta}_\lambda^L)_j = \frac{1}{n}(X^tX)_j \left(1 - \frac{\lambda}{2} \frac{1}{|\,(X^tX)_j\,|}\right) \mathbf{1}_{\{|(X^tX)_j| > \frac{\lambda}{2}\}}.$$

Démonstration

On vient de voir que

$$X^tX\hat{\beta}_\lambda^L = X^tY - \frac{\lambda\hat{\delta}}{2}.$$

Comme $X^tX = nI_{d+1}$, alors

$$\hat{\beta}_\lambda^L = \frac{X^tY}{n} - \frac{\lambda}{2n}\hat{\delta}.$$

Ainsi pour tout $j \in \{0, 1, \ldots, d\}$,

$$(\hat{\beta}_\lambda^L)_j = \frac{1}{n}(X^tX)_j - \frac{\lambda}{2n}\hat{\delta}_j. \tag{8}$$

Démonstration

Si $(\hat{\beta}_\lambda^L)_j \neq 0$, alors

$$\hat{\delta}_j = signe((\hat{\beta}_\lambda^L)_j) = \begin{cases} 1 & \text{si} \quad (\hat{\beta}_\lambda^L)_j > 0 \\ -1 & \text{si} \quad (\hat{\beta}_\lambda^L)_j < 0. \end{cases}$$

Ainsi (8) donne

$$(\hat{\beta}_\lambda^L)_j + \frac{\lambda}{2n} signe((\hat{\beta}_\lambda^L)_j) = \frac{1}{n}(X^t X)_j.$$

On déduit que le signe de $(\hat{\beta}_\lambda^L)_j$ est le même que le signe de $(X^t X)_j$.

Démonstration

Il s'en suit que

$$
\begin{aligned}
(\hat{\beta}_\lambda^L)_j &= \frac{1}{n}(X^tX)_j - \frac{\lambda}{2n} signe((X^tX)_j) \\
&= \begin{cases} \frac{1}{n}(X^tX)_j - \frac{\lambda}{2n} & \text{si} \quad (X^tX)_j > 0 \\ \frac{1}{n}(X^tX)_j + \frac{\lambda}{2n} & \text{si} \quad (X^tX)_j < 0 \end{cases}
\end{aligned}
$$

Remarques

En conclusion

- Si $(\hat{\beta}_\lambda^L)_j \neq 0$, alors forcément $\mid (X^tX)_j \mid > \frac{\lambda}{2}$.
- Si $\mid (X^tX)_j \mid \leq \frac{\lambda}{2}$, alors $(\hat{\beta}_\lambda^L)_j = 0$ (par contraposé).

Démonstration

Finalement

$$
\begin{aligned}
(\hat{\beta}_\lambda^L)_j &= \left(\frac{1}{n}(X^tX)_j - \frac{\lambda}{2n} signe((X^tX)_j)\right) \mathbf{1}_{\{|(X^tX)_j| > \frac{\lambda}{2}\}} \\
&= \frac{1}{n}(X^tX)_j \left(1 - \frac{\lambda}{2} \frac{signe((X^tX)_j)}{(X^tX)_j}\right) \mathbf{1}_{\{|(X^tX)_j| > \frac{\lambda}{2}\}} \\
&= \frac{1}{n}(X^tX)_j \left(1 - \frac{\lambda}{2} \frac{1}{|\,(X^tX)_j\,|}\right) \mathbf{1}_{\{|(X^tX)_j| > \frac{\lambda}{2}\}}.
\end{aligned}
$$

Corollaire

Sous la condition $X^tX = nI_{d+1}$, on

$$(\hat{\beta}_\lambda^L)_j = \frac{1}{n}(X^tX)_j \max\left(\left(1 - \frac{\lambda}{2}\frac{1}{|\,(X^tX)_j\,|}\right); 0\right).$$

Remarque

- Dans ce cas, $(\hat{\beta}_\lambda^L)_j = 0$ lorsque $|\,(X^tX)_j\,| \leq \frac{\lambda}{2}$.
- Les variables sélectionnées sont celles pour lesquelles $|\,(X^tX)_j\,| > \frac{\lambda}{2}$.

Rappel (AIC)

Le critère d'information d'Akaike s'écrit comme suit

$$AIC = 2k - 2\ln(L),$$

où k désigne le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

Rappel (BIC)

Il existe de nombreux critères d'informations inspirés du critère d'Akaike. Le critère d'information bayésien (BIC) est l'un des plus populaires.

$$BIC = -2\ln(L) + \ln(n)k,$$

où n est le nombre d'observations dans l'échantillon et k est le nombre de paramètres.

Remarque (Choix de λ selon le critère BIC)

$$BIC_\lambda = \frac{1}{n\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - x_i^t\hat{\beta}_\lambda^L)^2 + \ln(n) \times card\{j;\ (\hat{\beta}_\lambda^L)_j \neq 0\},$$

où

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i^t\hat{\beta}_{\overline{\lambda}}^L)^2,$$

avec $\overline{\lambda}$ arbitraire.

Choix de λ

$$\hat{\lambda}^{BIC} = \arg\min_{\lambda}(BIC_\lambda).$$