Régression linéaire

"Ridge" et "Lasso"

C. Preda

$Y, \{X_1, \dots X_p\}$ centrées

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \Sigma$$

$$\Sigma \sim N(0, \sigma^2)$$

n - observations

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ \vdots & & \vdots \\ X_{i1} & \dots & X_{ip} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \Sigma_1 \\ \vdots \\ \Sigma_i \\ \vdots \\ \Sigma_n \end{pmatrix}$$

En forme matricielle

$$Y = X\beta + \varepsilon$$

matrice du design.

Estimateur de MCO (ou de maximum de vraisemblance)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Si $X^T X$ est singulière

$$\det(X^T X) = 0$$

alors $(X^T X)^{-1}$ n'existe pas.

Si $\lambda_1, \dots \lambda_p$ sont v. p. de $(X^T X)$

alors

$$\det(X^T X) = 0 \iff \exists\, q \in [1:p] \text{ t.q.}$$

$$\lambda_q = \lambda_{q+1} = \dots \lambda_p = 0$$

Or, on a vu que (lecture 3, p. 47) :

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{n-p-1} \sum_{k=1}^{p} u_{kj}^2 \cdot \frac{1}{\lambda_k}$$

Donc si $\lambda_q \neq 0 \Rightarrow \mathrm{Var}(\hat{\beta}_j) \to \infty$

$\forall j = 1 \dots p.$

Solutions :

- PCR
- PLS
- Ridge
- Lasso

Idée de la régression Ridge :

Considérer au lieu de

$$\hat{\beta}_{MCO} = (X^T X)^{-1} X^T Y$$

l'estimateur

$$\boxed{\hat{\beta}_{ridge} = (X^T X + \lambda I_p)^{-1} X^T Y}$$

avec $\lambda > 0$ – paramètre de régularisation

et I_p – la matrice identité $p \times p$

$$I_p = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

Remarque :

$(X^T X + \lambda I_p)$ est toujours inversible

car si λ_i est valeur propre de $(X^T X)$

alors $\boxed{\mu_i = \lambda_i + \lambda \text{ est v.p de } X^T X + \lambda I_p}$

En effet, soit (λ_i, u_i) un couple vp-vecteur propre de X^TX.

$$X^TX u_i = \lambda_i u_i$$

Alors

$$\left(X^TX + \lambda I_p\right) \cdot u_i = X^TX u_i + \lambda I_p u_i$$

$$= \lambda_i \cdot u_i + \lambda \cdot u_i$$

$$= \underbrace{(\lambda_i + \lambda) \cdot u_i}_{\mu_i} = \mu_i \cdot u_i$$

Donc $X^TX + \lambda I_p$ a les mêmes vecteurs propres que X^TX associés aux valeurs propres

$$\mu_i = \lambda_i + \lambda.$$

Comme $\lambda_i \geq 0$ et $\lambda > 0$ alors

$$\mu_i > 0.$$

et donc $X^TX + \lambda I_p$ est inversible

On sait que $\hat{\beta}_{MCO}$ est sans biais

$$\mathbb{E}\left(\hat{\beta}_{MCO}\right) = \beta$$

Proposition : L'estimateur ridge est biaisé

En effet :

$$\hat{\beta}_{ridge} = \left(X^T X + \lambda I_p\right)^{-1} X^T Y$$

$$= \left(X^T X + \lambda I_p\right)^{-1} \underbrace{\left(X^T X\right)\left(X^T X\right)^{-1}}_{I_p} X^T Y$$

$$= \left(X^T X + \lambda I_p\right)^{-1} X^T X \cdot \underbrace{\hat{\beta}_{MCO}}$$

Donc

$$\mathbb{E}\left(\hat{\beta}_{ridge}\right) = \left(X^T X + \lambda I_p\right)^{-1}\left(X^T X\right) \cdot \beta$$

Biais : $\beta - \mathbb{E}(\beta) = \left(I - \left(X^T X + \lambda I_p\right)^{-1} X^T X\right) \beta$

Par contre, la variance des estimateurs ridge est plus petite que celle MCO

$$\mathrm{Var}(\hat{\beta}_{ridge,j}) = \frac{\sigma^2}{n-p-1} \sum_{k=1}^{p} u_{kj}^2 \cdot \frac{1}{\lambda_k + \lambda}$$

et, de manière générale,

$$V(\hat{\beta}_{ridge}) = \sigma^2 \cdot W X^T X W$$

avec $W = \left(X^T X + \lambda I_p\right)^{-1}$

et

$$\mathrm{Biais}(\hat{\beta}_{ridge}) = -\lambda W \beta$$

Remarque importante

① On peut montrer (calcul direct) que l'estimateur ridge est le minimiseur du problème :

$$\hat{\beta}_{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \underbrace{\sum_{i=1}^{n} \left(Y_i - (\beta_1 X_{i,1} + \dots + \beta_p X_{i,p}) \right)^2}_{\text{MCO}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{\text{pénalité}} \right\}$$

λ est donc un paramètre qui "joue" (arbitre) entre l'adéquation aux données (MCO) et la "grandeur" des paramètres (stabilité)

$$\lambda \to 0 \implies \hat{\beta}_{ridge} \to \hat{\beta}_{MCO}$$

$$\lambda \to \infty \implies \hat{\beta}_{ridge} \to 0$$

② En format matriciel

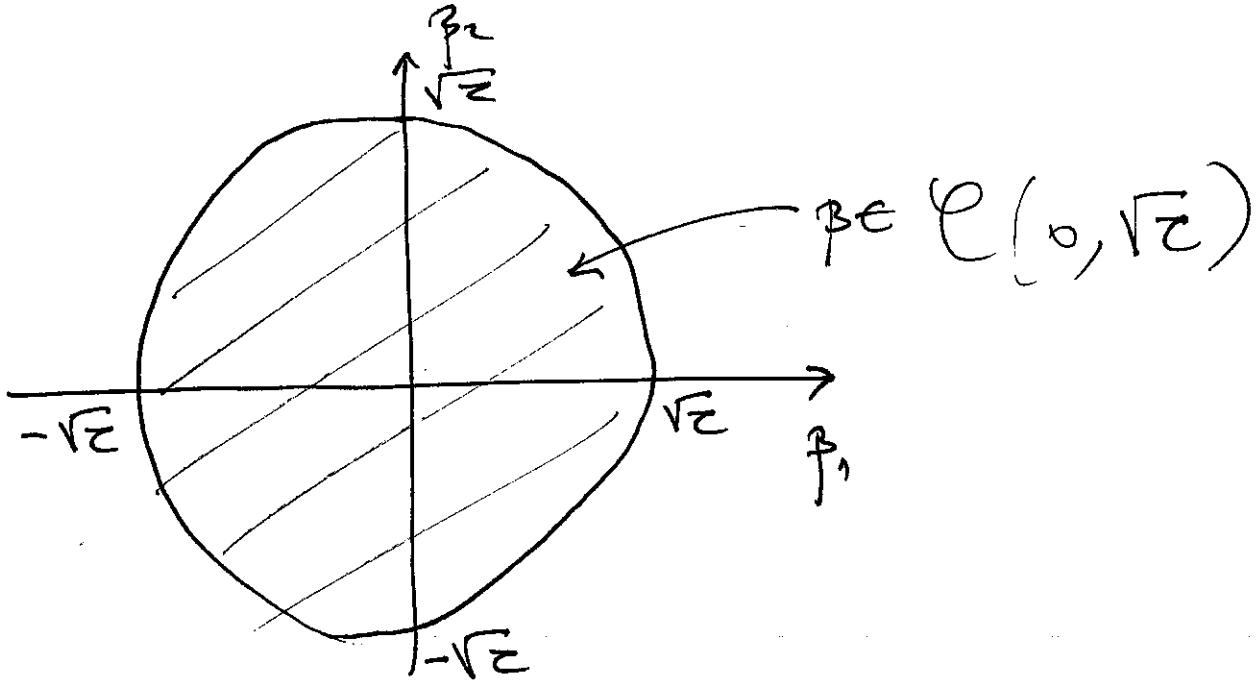
$$\hat{\beta}_{ridge} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|Y - X\beta\|^2 + \lambda \|\beta\|_2^2$$

Remarque importante (continuation)

(2) On peut montrer que $\hat{\beta}_{ridge}$ est solution du pb. de minimisation sous contrainte (multiplicateurs de Lagrange)

$$\hat{\beta}_{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \left(Y_i - (\beta_1 X_{i1} + \dots + \beta_p X_{ip}) \right)^2$$

$$\|\beta\|_2^2 < \tau \quad , \tau > 0.$$



λ est alors fonction de τ

$$\lambda = g(\tau)$$

Régression LASSO

$$\hat{\beta}_{Lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; \| y - X\beta \|_2^2 + \lambda \| \beta \|_1$$

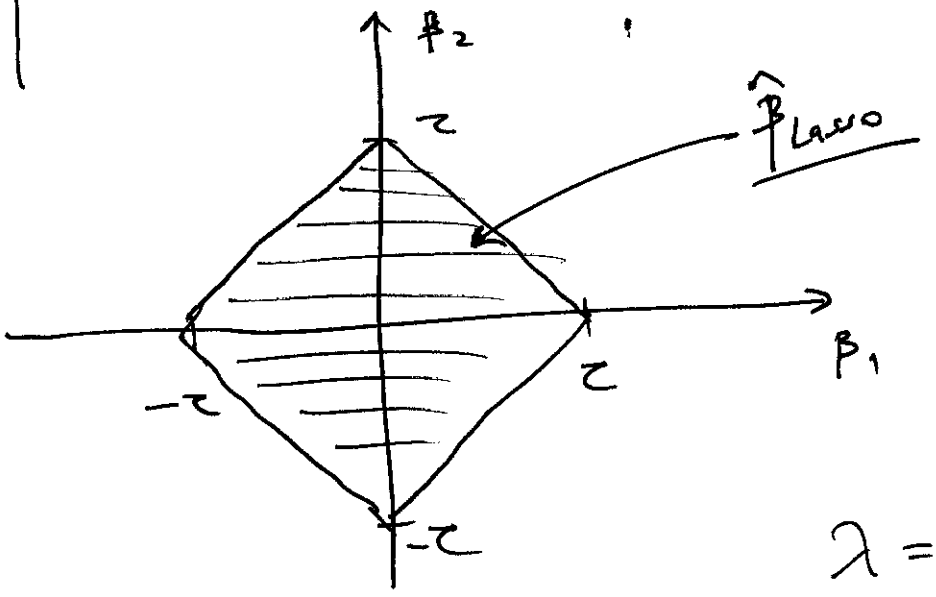
avec $$\| \beta \|_1 = |\beta_1| + |\beta_2| + \dots |\beta_p|$$

(la norme L_1)

Ce problème est équivalent à

$$\hat{\beta}_{Lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; \| y - X\beta \|_2^2$$

$$\| \beta \|_1 < \tau \quad , \quad \tau > 0$$



$$\lambda = g(\tau)$$

⚠ **Sparsité :**

La solution LASSO est telle que certains $\hat{\beta}_{lasso,j} = 0$

Donc Lasso régularise les coefficients et réalise également une sélection de variables

Elastic-net :

$$\hat{\beta}_{elastic} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; \|Y - X\beta\|_2^2 + \lambda\left[(1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1\right]$$

$\alpha = 0 \Rightarrow$ ridge

$\alpha = 1 \Rightarrow$ lasso

shrinkage (réduction de variance) + sélection des variables (lasso)