Lecture September 10

Shrinkage Methods: Ridge and Lasso,

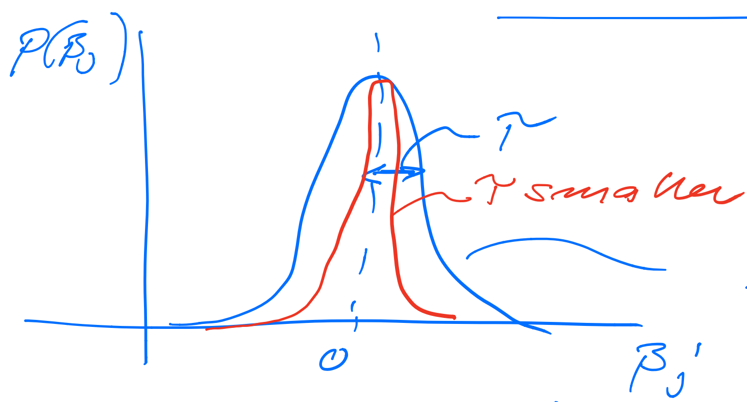
Intuitive understanding:

Bayes's theorem:

$$P(\beta|X y) \propto P(y|X\beta)\, P(\beta)$$

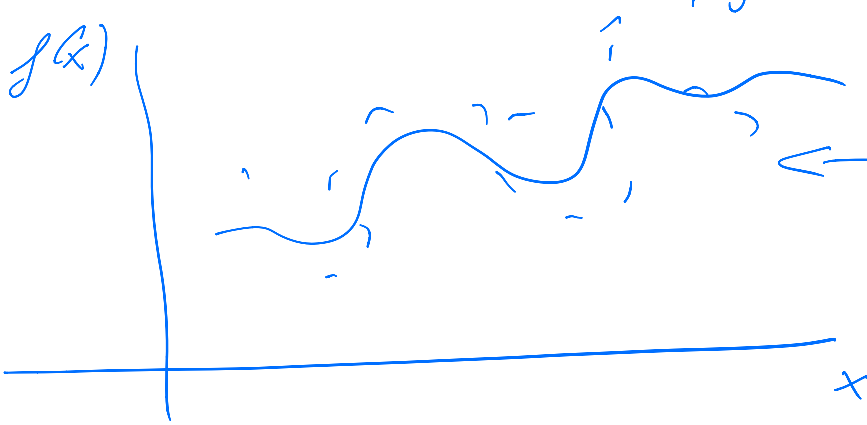
$$\prod_{i=0}^{n-1} N(y_i|x_i^T\beta)$$

$\boxed{\text{prior}}$



$$P(\beta) = \prod_{j=0}^{p-1} N(0,\tau^2)$$

$$e^{-\beta_j^2/2\tau^2}$$



potential for large variance of β

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

Can we shrink this variance!

$$\hat{\beta}^{Ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; \frac{1}{n}\sum_{i=0}^{n-1}(y_i - x_{i*}\beta)^2 + \lambda \sum_{j=0}^{p-1}\beta_j^2$$

L2-norm Regularization

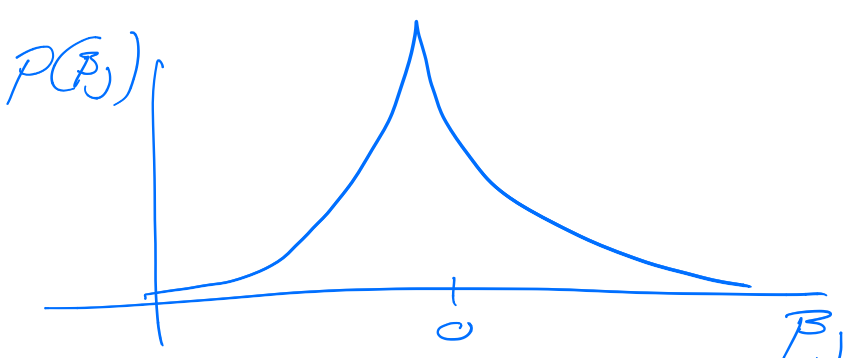
$\lambda \sim \frac{1}{2\tau^2}$ $\qquad \lambda \geq 0$

$$\Rightarrow \hat{\beta}^{Ridge} = \underbrace{(X^T X + \lambda I)}_{\in \mathbb{R}^{p\times p}}^{-1} X^T y$$

Lasso: Least absolute shrinkage and selection operator.

$P(\beta)$ = double exponential (Laplace) distribution

$$P(\beta_j) = e^{-|\beta_j|/\tau} \quad \left(e^{-|\beta_j - \mu|/\tau}\right)$$



$$P(\beta | xy) \propto \prod_{i=0}^{n-1} P(y_i | x_{i*}\beta) \prod_{j=0}^{p-1} e^{-|\beta|/\lambda}$$

minimize $-\log P(\beta|xy)$ to find $\hat{\beta}_{optimal}$

$$\hat{\beta}^{Lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=0}^{n-1} (y_i - X_{i*}\beta)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j|$$

L1-norm regularization

$$\hat{\beta}^{Lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \|(y - X\beta)\|_2^2 + \lambda \|\beta\|_1$$

Ridge $\lambda \geq 0$ given $\sum_{j=0}^{p-1} \beta_j^2 \leq s$

Lasso $\lambda \geq 0$ $\sum_{j=0}^{p-1} |\beta_j| \leq t$

Simple example to illustrate differences between $\hat{\beta}^{OLS}$, $\hat{\beta}^{Ridge}$ and $\hat{\beta}^{Lasso}$

X is a diagonal matrix

$$X = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & \ddots\, 1 \end{bmatrix}$$

Skip $1/n$ in MSE, $n = p$

<u>OLS</u> : $\sum_{i=0}^{n-1} (y_i - \beta_i)^2$

$$\hat{\beta}^{OLS} \Rightarrow \underline{\hat{\beta}_j^{OLS} = y_j}$$

<u>Ridge</u> : $\sum_{i=0}^{n-1} (y_i - \beta_i)^2 + \lambda \sum_{j=0}^{n-1} \beta_j^2$

$$\hat{\beta}_i^{Ridge} = \boxed{\frac{y_i}{1+\lambda}}$$

$\lambda = 0$

$\beta_i^{Ridge} = \beta_i^{OLS}$

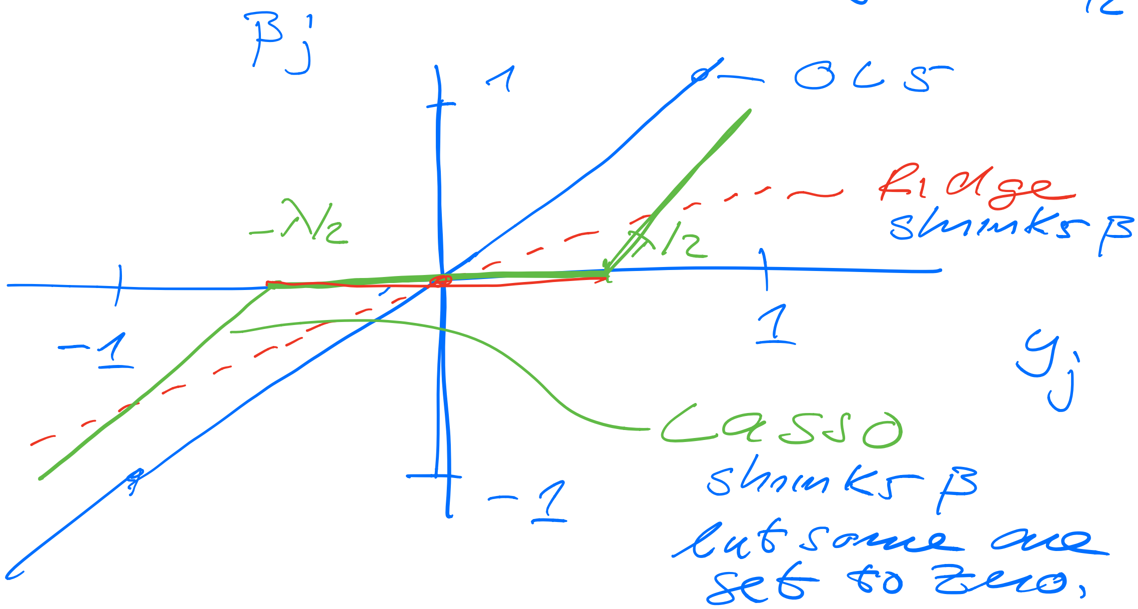
$$\lim_{\lambda \to \infty} \hat{\beta}_i^{Ridge} = 0$$

<u>Lasso</u>

$$\frac{\partial}{\partial \beta_i}\left(\sum_{i=0}^{n-1} (y_i - \beta_i)^2 + \lambda \sum_{i=0}^{n-1} |\beta_i| \right) = 0$$

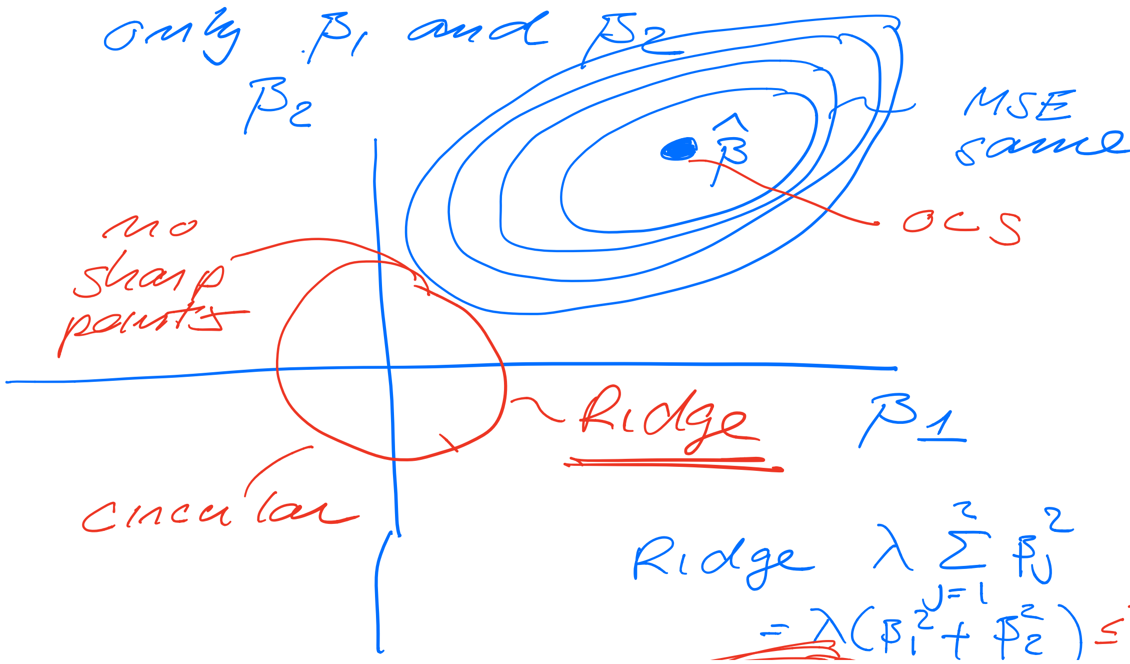
$$-2 \sum_i (y_i - \beta_i) + \lambda \sum_i \frac{\beta_i}{|\beta_i|} = 0$$

$$\beta_i^{Lasso} = \begin{cases} y_i - \lambda/2 & \text{if } y_i > \lambda/2 \\ y_i + \lambda/2 & \text{if } y_i < -\lambda/2 \\ 0 & \text{if } |y_i| \le \lambda/2 \end{cases}$$

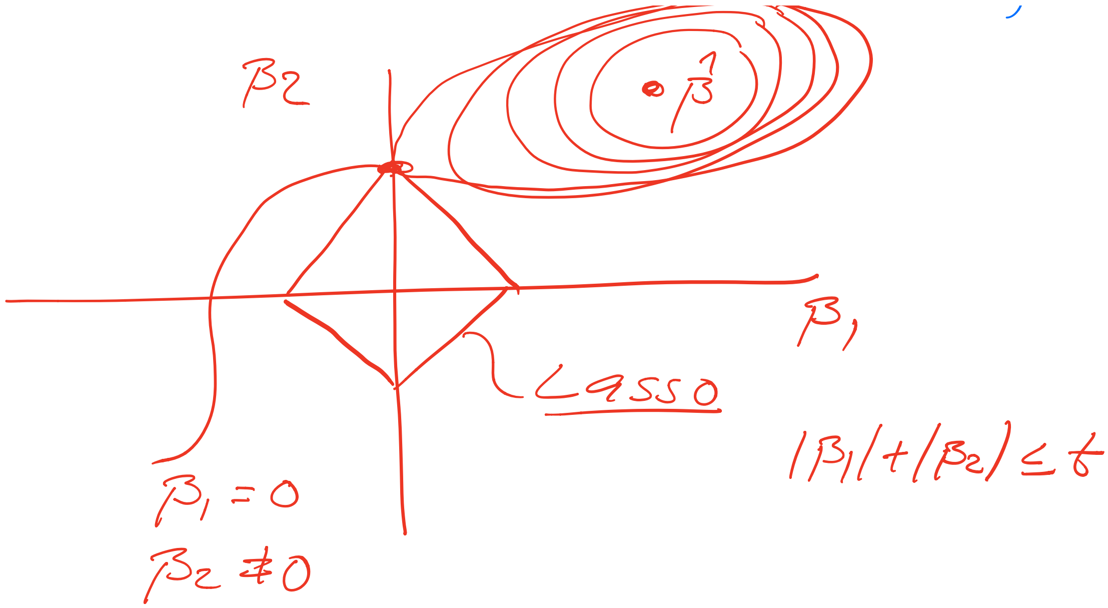


Example 2

only β_1 and β_2



Ridge $\lambda \sum_{j=1}^{2} \beta_j^2 = \lambda(\beta_1^2 + \beta_2^2) \le \lambda s$



Ridge analysis

- Singular value decomposition,
- covariance and correlation matrix.