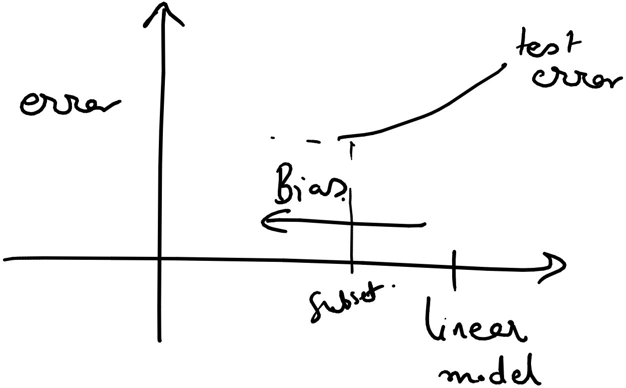
Beyond Subset Selection:

beyond retaining or excluding a coordinate

Ridge & LASSO



Ridge → loss should be low

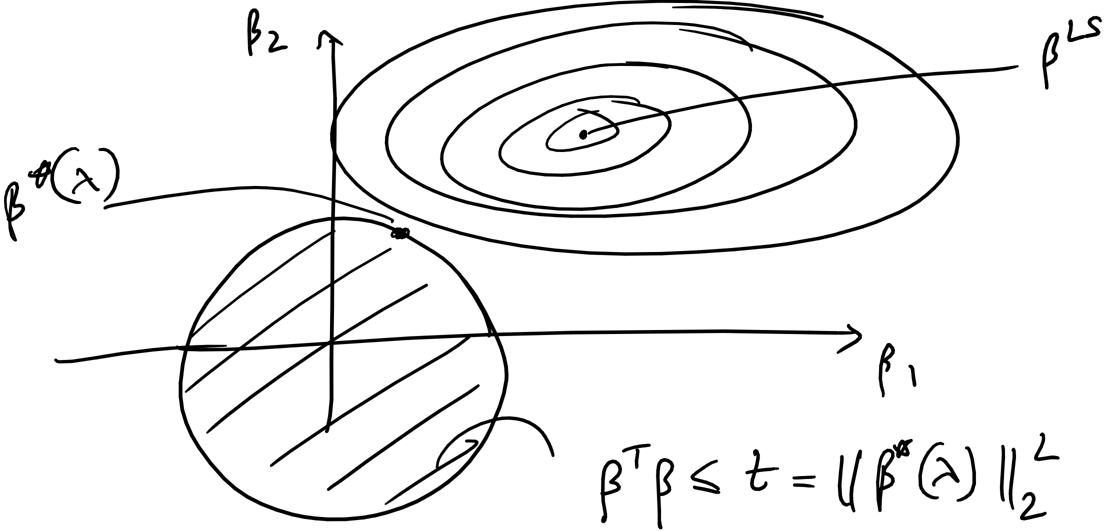
→ Coefficients ($\hat{\beta}_j$) be low in magnitude.

$$RSS(\lambda, \beta) = \underbrace{(Y - X\beta)^T (Y - X\beta)} + \underbrace{\lambda \beta^T \beta} \qquad \text{where } \lambda > 0$$

$\beta^*(\lambda)$

$$\min_{\beta} \; (Y - X\beta)^T (Y - X\beta)$$

$$\text{st} \quad \beta^T \beta \leq \underbrace{\|\beta^*(\lambda)\|_2^2} = t$$



$RSS(\theta, \beta)$

$\beta^T\beta \leq t = \|\beta^*(\lambda)\|_2^2$

$\begin{bmatrix} \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

$$\beta^{LS} = (X^T X)^{-1} X^T Y$$

$$\beta^{ridge} = (X^T X + \underline{\underline{\lambda I}})^{-1} X^T Y$$

$$\underbrace{\overset{Y^T Y}{a} - b - c + \beta^T X^T X \beta} + \beta^T (\lambda I) \beta$$

$$\beta^T (\underbrace{X^T X + \lambda I}_{D}) \beta$$

Goal: Relate MLE to Ridge regression $\longrightarrow$ Lec 6 $\square$

MLE eg $\quad z_1, \dots z_N \sim N(\mu, 1)$

Likelihood function $= P_\mu(Z_1 = z_1, Z_2 = z_2 \dots\dots, Z_N = z_N)$

$$= \prod_{i=1}^{N} P_\mu(Z_i = z_i)$$

$$LL = \sum_{i=1}^{N} \log P_\mu(Z_i = z_i)$$

$$P(Z_i = z_i) = \frac{1}{\sqrt{2\pi}\,1} \exp\left(-\frac{(z_i-\mu)^2}{2 \cdot 1^2}\right)$$

$$\log \text{''} = -\frac{(z_i-\mu)^2}{2} - \log\sqrt{2\pi}$$

$$\max_{\mu}(LL) = -\frac{1}{2}\sum_{i=1}^{N}(z_i-\mu)^2 - N\log\sqrt{2\pi}$$

$$\frac{\partial(\text{''})}{\partial\mu} = \sum_{i=1}^{N}(z_i-\mu) = 0$$

$$\boxed{\mu = \frac{1}{N}\sum_{i=1}^{N} z_i}$$

positive semi-definite matrix

$X^T X$ is if smallest eigenvalue is ≥ 0

$A_{p \times p}$ $\quad \underline{\underline{\lambda_1}} \ldots \underline{\underline{\lambda_p}}$

$\lambda = v^T A v$ for v being eigenvector.

≥ 0

$$v^T A \underline{\underline{v}} = \underbrace{\underline{\underline{v^T \lambda v}}}_{\geq 0} = \underbrace{\|v\|_2^2}_{\geq 0} \lambda \quad X^T X$$

$$\lambda = v^T X^T X v$$

$$= (Xv)^T (Xv)$$

$$= \|Xv\|_2^2 \geq 0$$

Interpreting Ridge Regression:

$y_i - \hat{\beta}^T x_i$

Fact: $\text{svd}(X) = \underset{N\times p}{U}\ \underset{p\times p}{D}\ \underset{p\times p}{(V^T)}$

$A = UDV^T$

$|d_1| > |d_2| > \cdots$

$$X\hat{\beta} = X(X^TX)^{-1}X^TY$$

$$= UDV^T(VD^2V^T)^{-1}VDU^TY$$

$$= UDD^{-2}DU^TY$$

$$X^TX = VD^2V^T$$

$$= VDU^TUDV^T$$

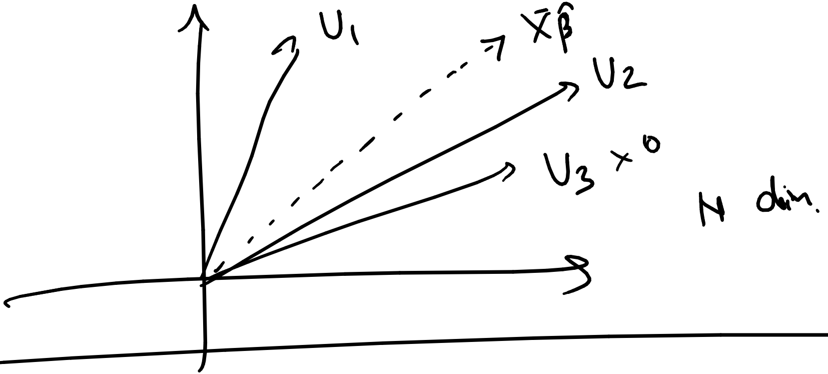
$$\underset{N\times 1}{X\hat{\beta}} = \underset{N\times p}{U} \cdot \underbrace{\underset{p\times N}{U^T} \underset{N\times 1}{Y}}$$

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} = c$$

$$= Uc$$

$$= \sum_{j=1}^{p} c_j \cdot u_j$$

$$U = \underset{N\times p}{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_p \\ | & & | \end{bmatrix}}$$



<u>Ridge</u>

$$X\beta^{ridge} = UD\underline{(D^2+\lambda I)}^{-1}DU^TY$$

$$= \sum_{j=1}^{p} \underline{\left(\frac{d_j^2}{d_j^2+\lambda}\right)(U_j^T Y)} \cdot U_j$$

What does d_j^2 mean?

$$X = U D V^T_{p \times p}$$

$$\underbrace{\frac{1}{N} X^T X_{p \times p}}$$

V : the columns are called pricipal component directions.

Centering : X

. take column means

. Subtract from column entries

$$\underset{p\times p}{X^T X} = \underset{p\times p}{V}\,\underset{p\times p}{D^2}\,\underset{p\times p}{V^T} \quad : \text{eigen decomposition.}$$

$$D^2 = \begin{bmatrix} d_1^2 & & 0 \\ & \ddots & \\ 0 & & d_p^2 \end{bmatrix}_{p\times p}$$

v_1

$$= \sum_{j=1}^{p} d_j^2 v_j v_j^T$$

Fact

$$d_1^2 = \mathrm{Var}(X v_1)$$

1 2 3 4 5

-2 -1 0 1 2

$$\underline{\underline{\mathrm{Var}(X v_1)}} = v_1^T X^T X v_1$$

$$= v_1^T \underline{\underline{\left(\sum_{j=1}^{p} d_j^2 v_j v_j^T \right)}} v_1$$

$$= \sum_{j=1}^{p} d_j^2 (v_1^T v_j)(v_j^T v_1)$$

$$= \underline{\underline{d_1^2}} \cdot 1 \cdot 1.$$

$$\underbrace{X v_1}_{N \times 1 \text{ dim}}$$

$$\mathrm{svd}(X^T X) = V D^2 V^T = \mathrm{pca}(X) = \text{eigen decom positions } (X^T X)$$

Missing argument for $\beta^*(\lambda)$:

1. We know $\beta^*(\lambda)$ is feasible. i.e, $\beta^*(\lambda)^T \beta^*(\lambda) \leq \|\beta^*(\lambda)\|_2^2$ by definition.

Also for any other β we have

$$(Y - X\beta^*(\lambda))^T (Y - X\beta^*(\lambda)) + \lambda \|\beta^*(\lambda)\|_2^2$$
$$\leq (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_2^2$$

$$(Y - X\beta^*(\lambda))^T (Y - X\beta^*(\lambda)) \leq (Y - X\beta)^T (Y - X\beta) + \lambda \underbrace{(\|\beta\|_2^2 - \|\beta^*(\lambda)\|_2^2)}_{<0}$$

$\therefore$ $\beta^*(\lambda)$ is the minimizer for

$$\min \; (Y - X\beta)^T (Y - X\beta)$$
$$\text{Subject to } \beta^T\beta \leq \|\beta^*(\lambda)\|_2^2.$$