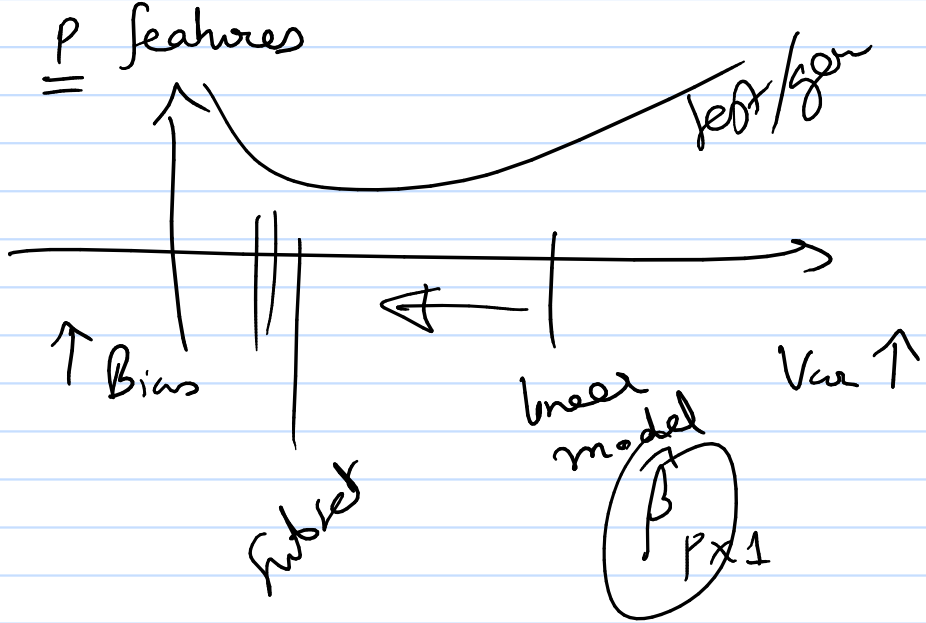
① Subset selection : why?



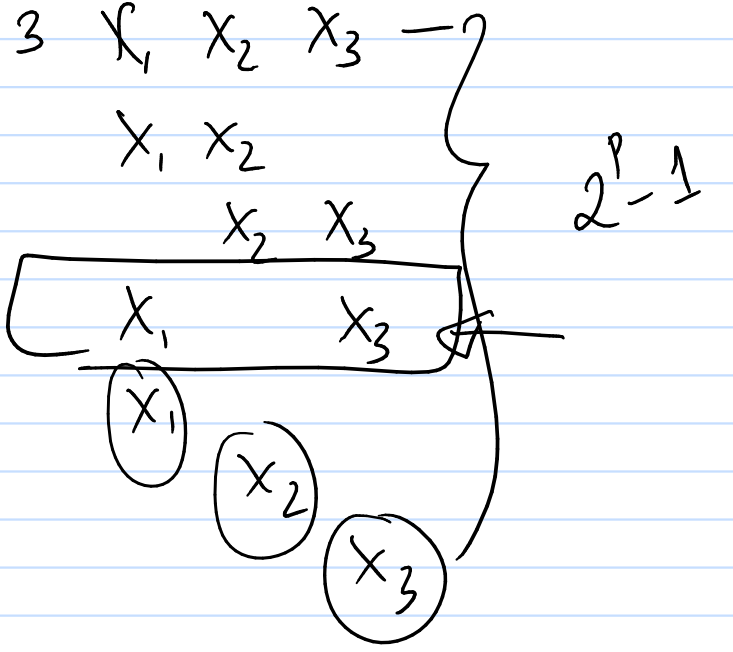
p features

1, 2, 3, ... p

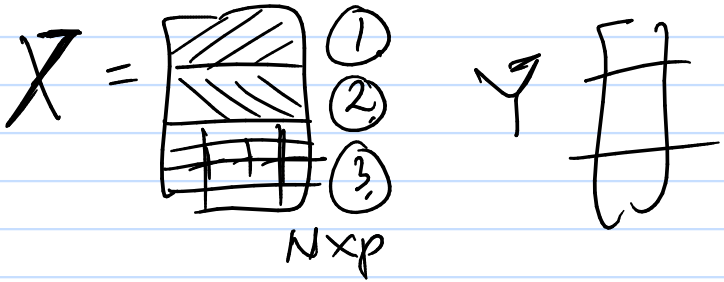
↑

p $\binom{p}{2}$ 1

How to find the best subset



Recall



for each choice:

| for each fold i:
| get a model using remaing folds
| error(fold i)

estimated EPE = Sum (errors) across folds $\times \frac{1}{3}$

① LASSO

Criteria: $RSS(\beta) + \lambda \cdot \sum_{j=1}^{p} |\beta_j|$

$10^{-2} = 0.01$

$\|\beta\|_1$

② Ridge regression

10^{3}

$RSS(\beta) + \lambda_1 \sum_{j=1}^{p} \beta_j^2$

$\|\beta\|_2^2$

mimic subset selection

$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$

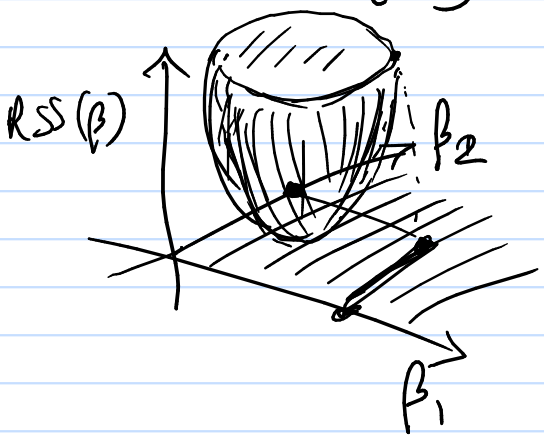
$\beta_3 \to 0$, $\beta_4 \to 0$

LASSO

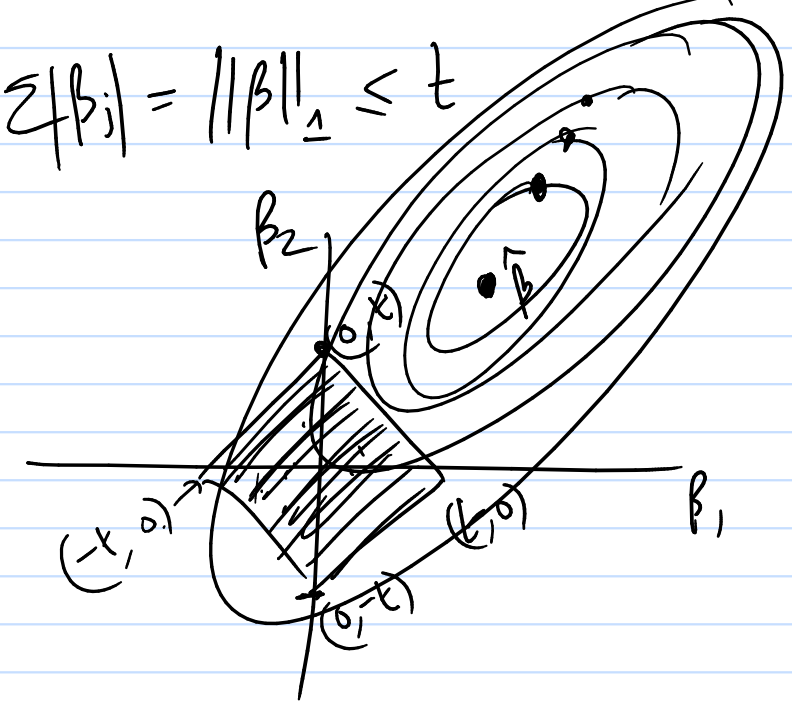
$$\min_{\beta} RSS(\beta) \quad \text{st} \quad \|\beta\|_1 \leq t$$

RR $\min_{\beta} RSS(\beta)$ st. $\|\beta\|_2^2 \leq t$

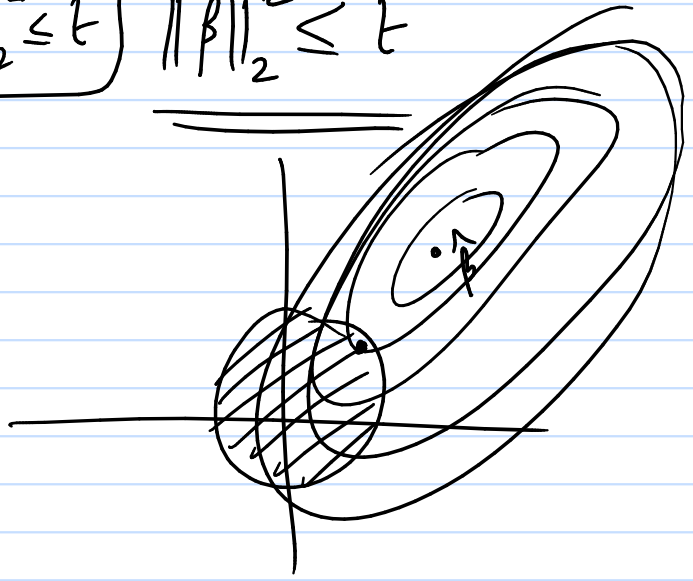
β_1 β_2 $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$ $p = 2$

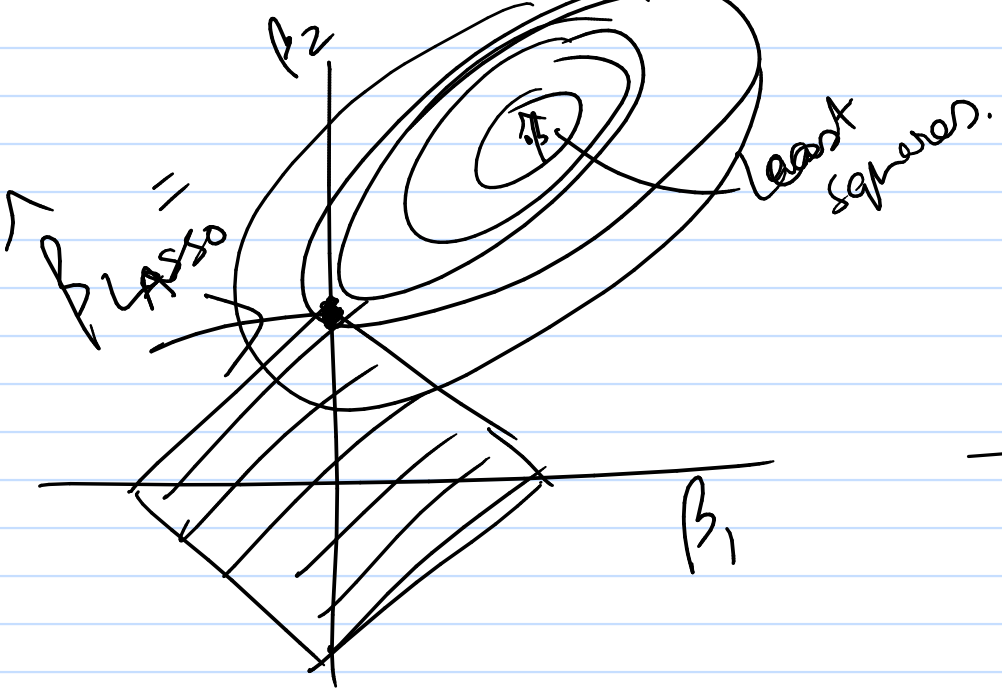


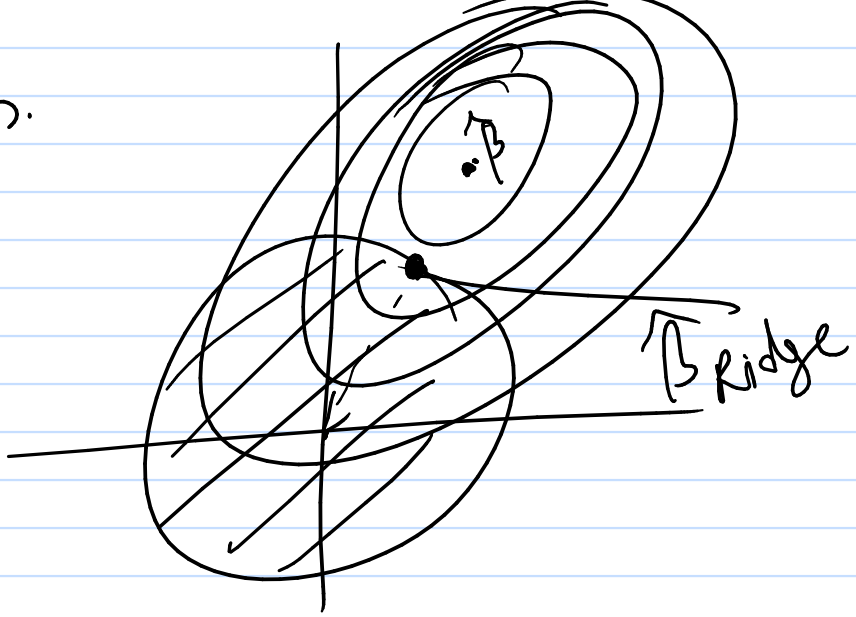
$\sum |\beta_j| = \|\beta\|_1 \leq t$



$\beta_1^2 + \beta_2^2 \leq t$ $\|\beta\|_2^2 \leq t$




β_2
$\hat{\beta}$
least squares.
$\hat{\beta}_{Lasso}$
β_1


$\hat{\beta}$
$\hat{\beta}_{Ridge}$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \longrightarrow P(Z|W) = \frac{P(Z,W)}{P(W)}$$

$$P(Z=z \mid W=w) = \frac{P(W=w, Z=z)}{P(W=w)}$$

Bayes rule:

$$P(Z|W) \cdot P(W) = P(W|Z) \cdot P(Z) = P(Z,W)$$

$$P(Z|W) = \frac{P(W|Z)\,P(Z)}{P(W)}$$

$\beta \sim RV$ Prior: $\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}\right)$ X fixed

Likelihood: $Y|\beta; X \sim \mathcal{N}(\beta^T X, \tau^2)$

Then

$$P(Y_1=y_1, Y_2=y_2 \dots Y_N=y_N | \beta, X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(y_i-\beta^T x_i)^2}{2\tau^2}\right)$$

data

$P(\beta | Y_1=y_1, Y_2=y_2 \dots Y_N=y_N)$ ← Normal

Ridge: $\min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T x_i)^2 + \lambda \cdot \sum_{j=1}^{p} \beta_j^2$. (MAP)

$P(\beta \mid data) \propto P(data \mid \beta) \cdot P(\beta)$: Maximum a-posteriori estimate

$\max_{\beta} P(data \mid \beta) \cdot P(\beta)$

$\min_{\beta} -\log P(data \mid \beta) - \log P(\beta)$

$\log \frac{1}{(\sqrt{2\pi}\beta)^2} \exp\left(\frac{-1}{2\beta^2}(\beta_1^2 + \beta_2^2)\right)$

$= \lambda \sum_{j=1}^{p} \beta_j^2 + \dots$

$$\sum_{i=1}^{N} (y_i - \beta^T x_i)^2 \propto -\log P(\text{data}|\beta)$$

$$RSS(\beta) \longleftrightarrow \propto \text{maximizing likelihood}$$

$$\min -\log \text{likelihood}$$

$$-\log P(\text{data}|\beta) = -\log \left[P(Y_1 = y_1|\beta) \times P(Y_2 = y_2|\beta) \cdots \times P(Y_N = y_N|\beta) \right]$$

$$\log(a \times b) = \log a + \log b$$

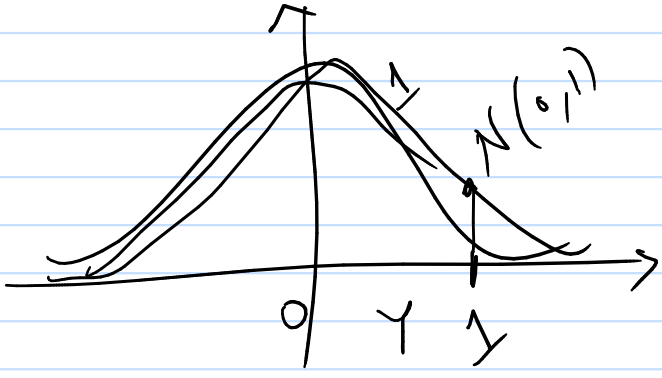
$$= -\log P(Y_1 = y_1 | \beta) - \log P(Y_2 = y_2 | \beta) \dots$$

$$= \sum_{i=1}^{N} -\log P(Y_i = y_i | \beta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(y_i - \beta^T x_i)^2}{2\tau^2} \right)$$

$$\propto \sum_{i=1}^{N} \frac{(y_i - \beta^T x_i)^2}{2\tau^2}$$

$$\log \frac{1}{\sqrt{2\pi}\tau} - \frac{(y_i - \beta^T x_i)^2}{2\tau^2}$$

a b



$$\frac{1}{\sqrt{2\pi}1}\exp\left(\frac{-y^2}{2}\right) = P(Y=y)$$

$$P(Y=1$$
$$=$$