Ridge Regression

Ridge penalizes the sq. err. to avoid "large" β:

$$\underbrace{\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|_2^2}_{} + \underbrace{\lambda \|\beta\|_2^2}_{}$$

$\lambda \geq 0$

By adding $\lambda\|\beta\|_2^2$ if the entries of β become large so does this penalty

λ = penalty strength

$\lambda = 0$ gives OLS $\hat{\beta}$

$\lambda \to \infty$ we get $\hat{\beta}^{(ridge)} \to 0$

Typically, we don't include β_0 in the

Typically, we don't include β_0 in the penalty

Often, we standardize vars before ridge

Also, typically choose λ via x-val.

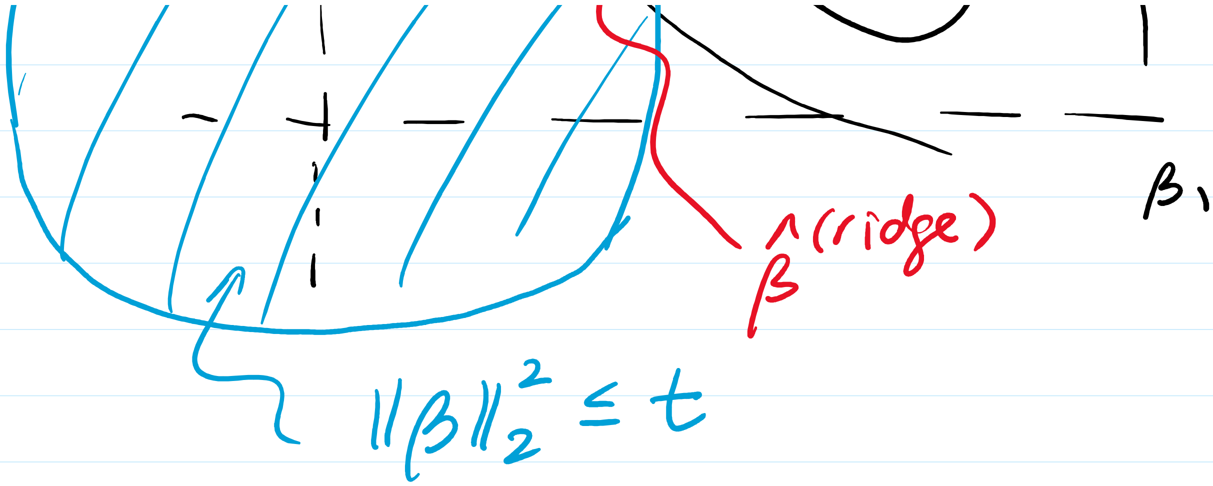
Second interpretation: ridge is equivalent to

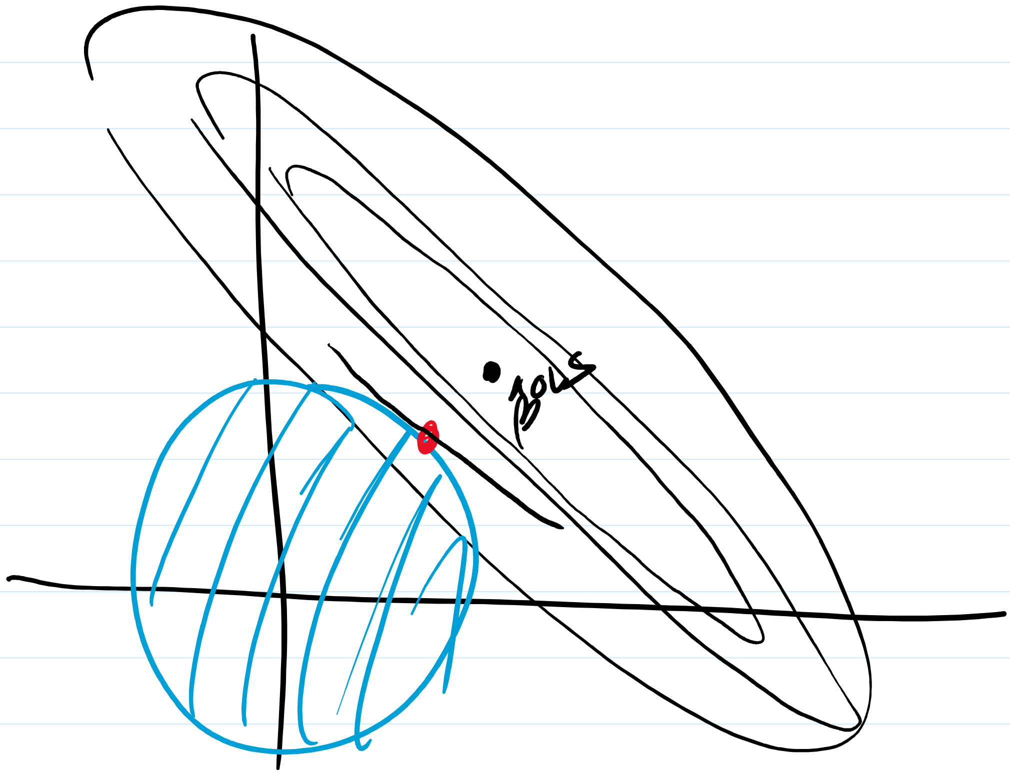
$$\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|_2^2$$

$$\text{s.t. } \|\beta\|_2^2 \le t$$

1-1 corresp w/ λ

Ex,





How do we get $\hat{\beta}^{(ridge)}$?

Because $\lambda\|\beta\|^2$ is quadratic and so is

$\Big(\quad \sum^{p} \beta_j^2$

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

$\|Y - X\beta\|_2^2$ then there is a closed form soln for $\hat{\beta}^{(ridge)}$.

<u>OLS</u>: $\frac{\partial L}{\partial \beta} = 0 \Rightarrow$ Solve $(X^TX)\beta = X^TY$

<u>Ridge</u>: $\frac{\partial L}{\partial \beta} = 0 \Rightarrow$ Solve $(X^TX + \lambda I)\beta = X^TY$

for $\lambda > 0$ $X^TX + \lambda I$ is invertible

so $\boxed{\hat{\beta}^{(ridge)} = (X^TX + \lambda I)^{-1} X^TY.}$

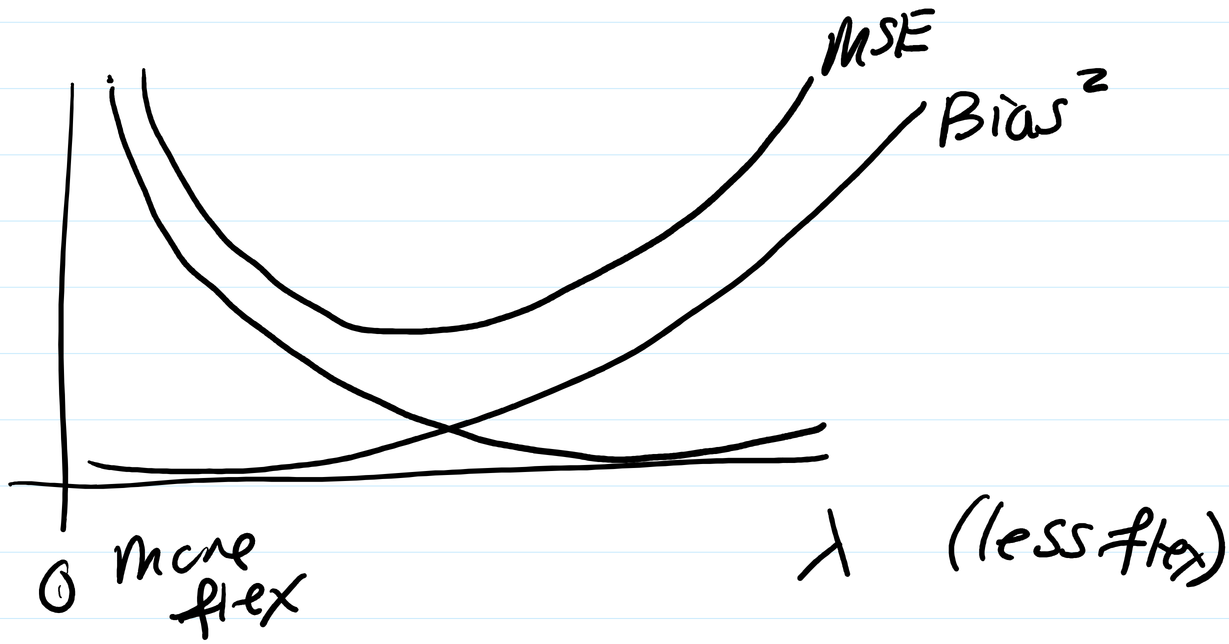
For OLS the sens. of $\hat{\beta}^{(OLS)}$ depended on

$$K(X^TX)$$

For ridge the sens. of $\hat{\beta}^{(ridge)}$ deps.

$$K(X^TX + \lambda I)$$

$$K(X^TX+\lambda I)$$



For OLS we showed that if

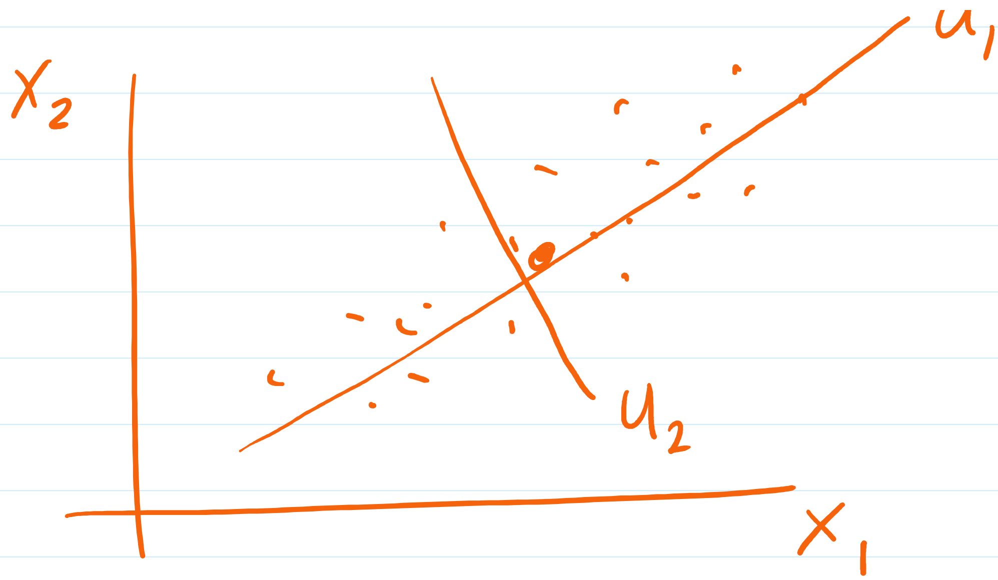
$$X = UDV^T \quad \text{(full rank)}$$

then

$$\hat{Y} = X\hat{\beta}^{(OLS)} = \sum_{j=1}^{p} u_j u_j^T Y$$

(1) proj. Y onto u_j

(2) Sum up these contribs.



For ridge: can show that

$$\hat{Y} = X\hat{\beta}^{(ridge)} = \sum_{j=1}^{p} \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda} \right) u_j u_j^T Y$$

① project Y onto u_j

② rescale each by $\dfrac{\sigma_j^2}{\sigma_j^2+\lambda} \leq 1$

③ Sum up contribs

scale comps assoc. w/ smaller σ_j more towards

σ_j more towards zero

Degrees of Freedom

For OLS : $df = P$ if $\text{rank}(X) = P$

For ridge: $df = \sum_{j=1}^{P} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \leq P$

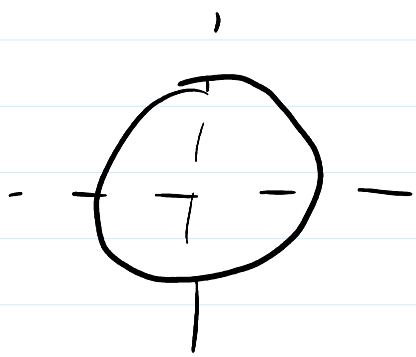
$$df \to 0 \text{ as } \lambda \to \infty$$
$$df \to P \text{ as } \lambda \to 0$$

Norms

Euclidean Norm:

$$\|x\|_2 = \sqrt{\sum_{j=1}^{p} x_j^2}$$

Consider: $\{x : \|x\|_2 = 1\}$



Can generalize:

q-norm : $\|x\|_q = \left(\sum_{j=1}^{p} |x_j|^q \right)^{1/q}$

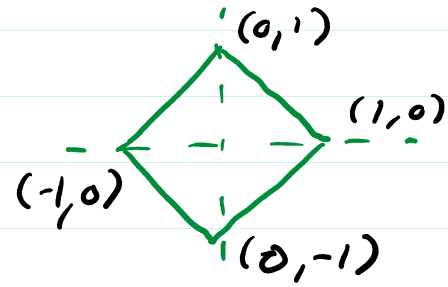
q-norm : $\|x\|_q = \left(\sum_{j=1} |x_j|^q \right)$

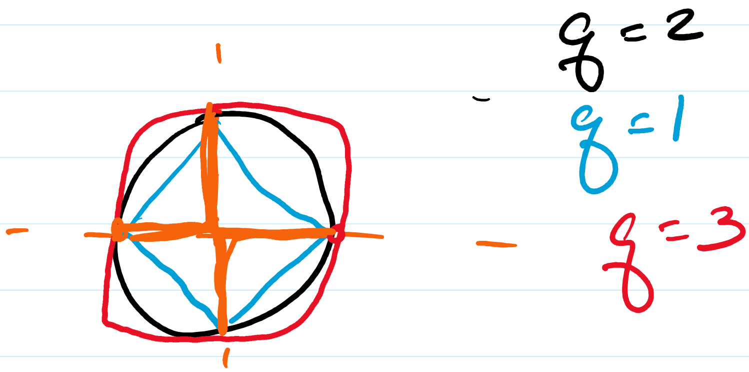
When $q=2$, I get euclidean norm (L2 norm)

If $q=1$, get L1 norm

$$\|x\|_1 = \sum_{j=1}^{p} |x_j|$$

Consider $\{x : \|x\|_1 = 1\}$





as $q \to \infty$ I get $\|x\|_q \to \max_j |x_j| = \|x\|_\infty$

$q \to 0$ I get $\|x\|_0 \to$ # of non-zero

$q \to 0$ I get $\|X\|_q \to$ # of non-zero elements

$= \|X\|_0$

Variable selection is like zeroing out some of my $\hat{\beta}$s:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots$$

set $\hat{\beta}_2 = 0$

now: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots$
