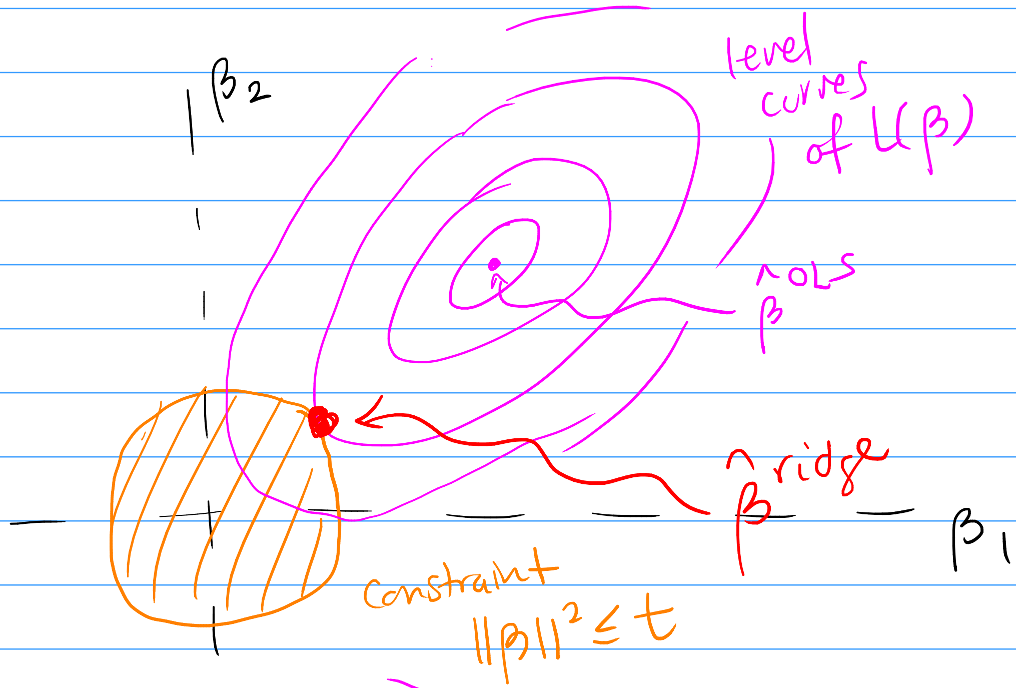
Lecture 13 : Ridge Regression

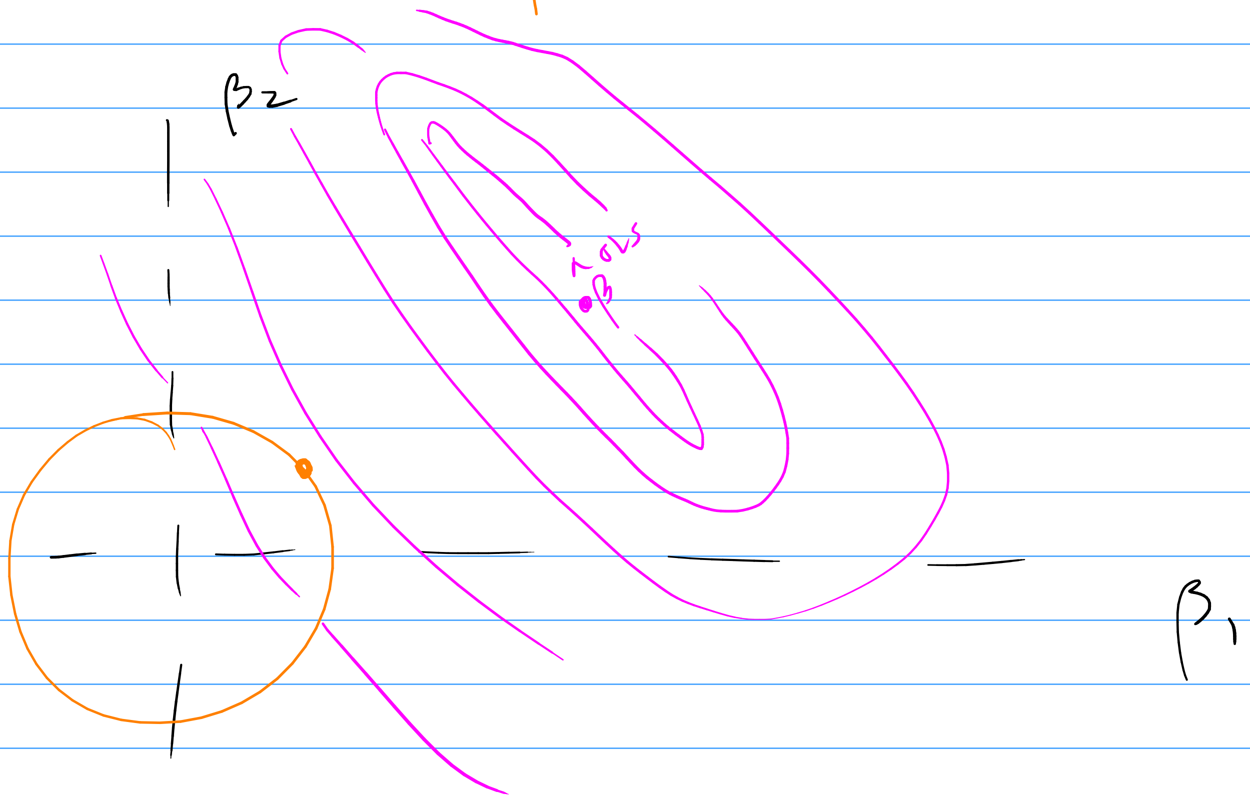
Second Interpretation: ridge is equivalent to

$$\hat{\beta}^{(ridge)} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \quad \text{s.t. } \|\beta\|^2 \leq t$$

1 – 1 corresp. w/ λ

Ex.





How do I actually get $\hat{\beta}^{ridge}$?

Because $\lambda\|\beta\|^2$ is quadratic, and so is $L(\beta)$

there is a closed-form soln

OLS: $\frac{\partial L}{\partial \beta} = 0 \Rightarrow$ solve $(X^T X)\beta = X^T Y$

Ridge: $\frac{\partial(L + \lambda\|\beta\|^2)}{\partial \beta} \Rightarrow$ Solve $(X^T X + \lambda I)\beta = X^T Y$

for $\lambda > 0$ $X^T X + \lambda I$ is always invertible.

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

For OLS that sensitivity of $\hat{\beta}^{OLS}$ depended on

$K(X^T X)$

For ridge sensitivity depends on

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

OLS: $\hat{\beta}^{OLS} = (X^TX)^{-1}X^TY$ $\quad$ ($\text{rank}(X) = \#\text{cols} = p$)

$X = UDV^T$

$X^TX = VD^TU^TUDV^T = VD^TDV^T$

$D_* = \text{diag}(\sigma_i)$

$$D = \left[\begin{array}{c|c} \sigma_1 \cdots \sigma_p & 0 \\ \hline 0 & 0 \end{array}\right]_{n\times p} = \left[\begin{array}{c} D_* \\ \hline 0 \end{array}\right]$$

$= V D_*^T D_* V^T$

$= V D_*^2 V^T$

$$D^TD = \left[\begin{array}{c|c} D_*^T & 0 \end{array}\right]\left[\begin{array}{c} D_* \\ \hline 0 \end{array}\right] = D_*^TD_* = D_*^2$$

So $(X^TX)^{-1} = V D_*^{-2} V^T$ $\quad$ $D_*^{-2} = \begin{pmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_p^2 \end{pmatrix}$

Consider training est.

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

$= UD\cancel{V^TV}D_*^{-2}\cancel{V^TV}D^TU^TY$

$$= U\underbrace{DD_*^{-2}D^T}_{\left[\begin{array}{c|c} I_p & 0 \\ \hline 0 & 0 \end{array}\right]}U^TY$$

If I ignore $DD_*^{-2}D^T$ then I would get

$\hat{Y}^{OLS} = UU^TY$

$$UU^T = \sum_{j=1}^{N} u_j u_j^T \qquad u_j = j^{th} \text{ col of } U$$

Actually have

$$\hat{Y}_1^{OLS} = U\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]U^T Y$$

$$= \left(\sum_{j=1}^{p} u_j u_j^T\right) Y$$

$$\boxed{\hat{Y}^{OLS} = \sum_{j=1}^{p} \underbrace{u_j u_j^T Y}_{\text{proj. onto } u_j}}$$

proj. onto u_j.

<u>OLS Preds:</u> 1 proj. Y onto u_j $(u_j u_j^T Y)$

2 Sum up these projs.

Re-do w/ ridge $\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y$

$$\hat{Y}^{ridge} = X\hat{\beta}^{ridge} = UDV^T(VD_*^2V^T + \lambda I)^{-1} VD^T U^T Y$$

$$= UD\left(V^T(VD_*^2V^T + \lambda I)V\right)^{-1} D^T U^T Y$$

$$= UD\left(\cancel{V^T V} D_*^2 \cancel{V^T V} + \lambda \underbrace{\cancel{V^T V}}_{I}\right)^{-1} D^T U^T Y$$

$$= UD\underbrace{(D_*^2 + \lambda I)^{-1}}_{\begin{pmatrix} \frac{1}{\sigma_1^2+\lambda} & & \\ & \ddots & \\ & & \frac{1}{\sigma_p^2+\lambda} \end{pmatrix}} D^T U^T Y$$

$$\underbrace{D(D_*^2+\lambda I)^{-1}D^T}_{\begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & & \\ & \frac{\sigma_2^2}{\sigma_2^2+\lambda} & & 0 \\ & & \ddots & \\ & 0 & & 0 \end{bmatrix}}$$

$$\hat{y}^{ridge} = U \; U^T y$$

$$= \sum_{j=1}^{p} \left(\frac{\sigma_j^2}{\sigma_j^2+\lambda} \right) u_j u_j^T y$$

Ridge:

(1) proj. y onto u_j

(2) re-scale by $\frac{\sigma_j^2}{\sigma_j^2+\lambda} < 1$

(3) Sum up components

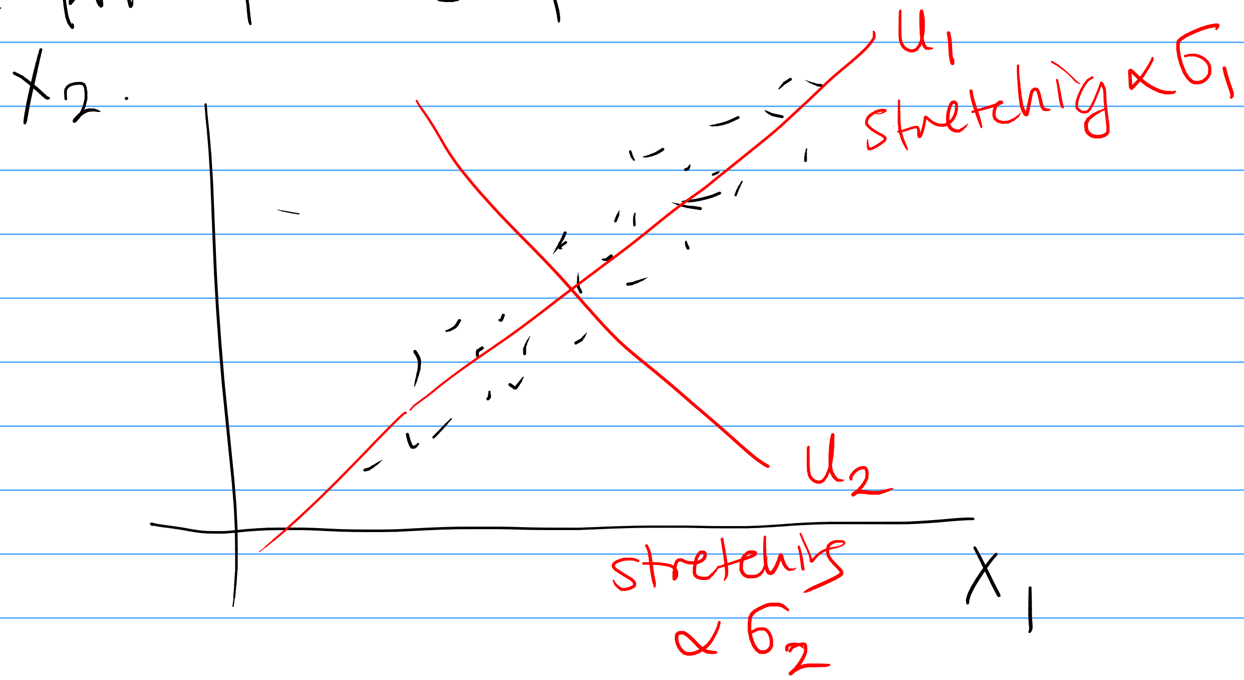
---

(1) If $\lambda = 0$ we get OLS b/c

$$\frac{\sigma_j^2}{\sigma_j^2+\lambda} = 1$$

(2) Large λ we really shrink components assoc. w/ smaller σ_j

$u_i \approx$ principal components



3 Degrees of freedom:

For OLS : $df = p$ $(\text{rank}\,X = p)$

For ridge : $df = \sum_{j=1}^{\text{rank}(X)} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \leq p$

$df \to 0$ as $\lambda \to \infty$

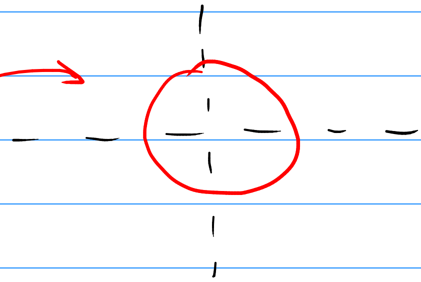
$df \to p$ as $\lambda \to 0$

Norms

Euclidean Norm:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$$

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



Can generalize:

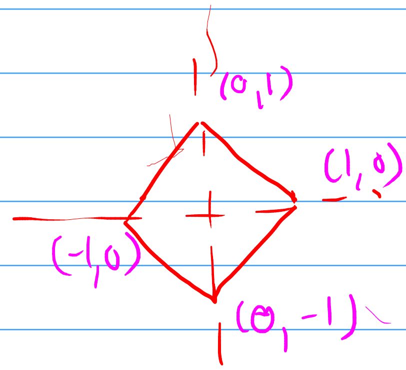
q-norm : $$\|x\|_q = \left(\sum_{i=1}^{p} |x_i|^q\right)^{1/q}$$

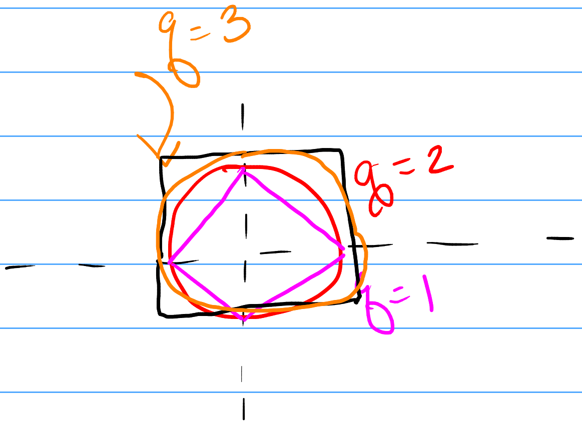
when $q = 2$ I get $\|\cdot\|_2$

when $q = 1$ I get the L1-norm

$$\|x\|_1 = \sum_{i=1}^{p} |x_i|$$

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





$\{x \mid \|x\|_q = 1\}$

as $q \to \infty$ I get $\|x\|_q = \max_i |x_i|$

$q \to 0$ I get $\|x\|_q = \#$ of non-zero element in x