Lecture 15:

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

$X = UDV^T$ where U, V are orthogonal
$U^TU = I$, $V^TV = I$

$$D = \left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$

$$\boxed{(AB)^T = B^TA^T}$$

$$X^TX = (UDV^T)^T UDV^T = VD^TU^TUDV^T$$
$$= VD^TDV^T$$

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

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

$$(X^TX)^{-1} = V(D^TD)^{-1}V^T$$

assume rank $X = P = \#$ cols
then $D^TD = D_*^2$

$$= VD_*^{-2}V^T$$

no zero blocks

$D_* = \text{diag}(\sigma_1, \dots, \sigma_r)$
$= D_*^T$

$$= \left[\begin{array}{c|c} D_*^2 & 0 \\ \hline 0 & 0 \end{array}\right]_{N \times P}$$

$$D_* = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$$

then $D_*^2 = \begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_r^2 \end{bmatrix}$ and $D_*^{-2} = \begin{bmatrix} 1/\sigma_1^2 & & \\ & \ddots & \\ & & 1/\sigma_r^2 \end{bmatrix}$

$$\hat{Y}^{OLS} = X\hat{\beta}^{OLS} = \overbrace{UDV^T}^{X}\overbrace{VD_*^{-2}V^T}^{(X^TX)^{-1}}\overbrace{VD^TU^TY}^{X^T}$$

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

If I ignore $DD_*^{-2}D$ then $\hat{Y} = UU^TY$

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

I can't actually ignore $DD_*^{-2}D$ so instead I get

$$U\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]U^T = \sum_{j=1}^{P} u_j u_j^T \qquad (P \leq N)$$

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

① project Y onto u_j : $u_j u_j^T Y$

② Sum up these components

Ridge:

$X = UDV^T$

$(AB)^{-1} = B^{-1}A^{-1}$

$$\hat{y}^{ridge} = X\hat{\beta}^{ridge}$$
$$= X(X^TX + \lambda I)^{-1}X^TY$$
$$= UDV^T(VD_*^2V^T + \lambda I)^{-1}VD^TU^TY$$
$$= UD(V^T(VD_*^2V^T + \lambda I)V)^{-1}D^TU^TY$$
$$= UD(\cancel{V^TV}D_*^2\cancel{V^TV} + \lambda \cancel{V^TV})^{-1}D^TU^TY$$
$$= UD\underbrace{(D_*^2 + \lambda I)}^{-1}D^TU^TY$$

$$D_*^2 + \lambda I = \mathrm{diag}(\sigma_i^2 + \lambda)$$
$$(D_*^2 + \lambda I)^{-1} = \mathrm{diag}\left(\frac{1}{\sigma_i^2 + \lambda}\right)$$
$$D(D_*^2 + \lambda I)^{-1}D^T = \left[\begin{array}{ccc|c} \sigma_1^2/\sigma_1^2+\lambda & & 0 & \\ 0 & \ddots & \sigma_r^2/\sigma_r^2+\lambda & 0 \\ \hline & 0 & & 0 \end{array}\right]$$

and so similar to previously

$$\boxed{\hat{y}^{ridge} = \sum_{j=1}^{r}\left(\frac{\sigma_i^2}{\sigma_i^2+\lambda}\right)u_j u_j^T Y}$$

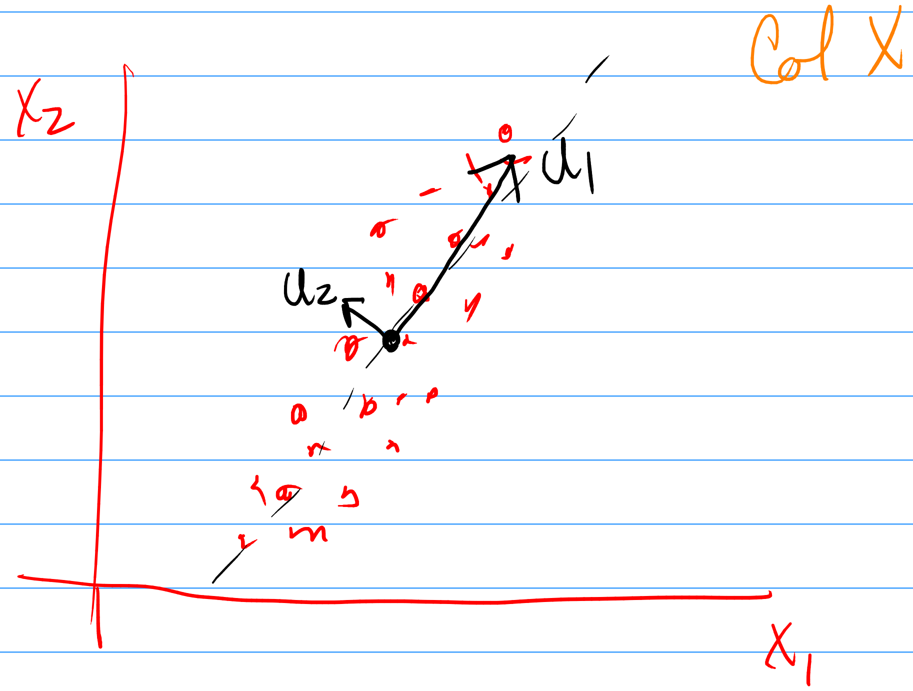
1. proj. Y onto u_j : $u_j u_j^T Y$
2. weighting by $\sigma_i^2/\sigma_i^2+\lambda < 1$
3. Sum up these components

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

b/c $\sigma_i^2 / \sigma_i^2 + \lambda = 1$

(2) Large λ really "shrinks" the contribution of the j^{th} component

(3) Shrinks contrib. of smaller σ_i faster



Degrees of Freedom

For OLS : $df = P$

For Ridge : $df = \sum_{j=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} < P$

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

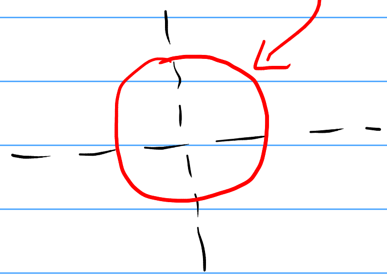
As we increase λ, flexibility decreases, my bias increases, var dec.

Aside: Norms

Ex. Euclidean Norm:

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

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



Can generalize

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

notice when $q=2$ I get Euclidean Norm.

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

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

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