TP : Introduction Supervised Learning

Jeanne Malécot

1 OLS

Since $\tilde{\beta}$ is unbiased :

$$\begin{align}
\beta &= \mathbb{E}(\tilde{\beta}) \tag{1}\\
&= \mathbb{E}(Hy) + \mathbb{E}(Dy) \tag{2}\\
&= \mathbb{E}(\beta^*) + D\mathbb{E}(y) \tag{3}\\
&= \beta + D\mathbb{E}(y) \tag{4}
\end{align}$$

Then we have $DX\beta = 0$, and finally :

$$\boxed{DX = 0}$$

Now we compute :

$$\begin{align}
Var(\tilde{\beta}) &= Var(Cy) \tag{5}\\
&= \sigma^2(H + D)(H^T + D^T) \tag{6}\\
&= \sigma^2 HH^T + \sigma^2 DD^T + \sigma^2 DH^T + \sigma^2 HD^T \tag{7}\\
&= \sigma^2 HH^T + \sigma^2 DD^T \tag{8}
\end{align}$$

And $Var(\beta^*) = \sigma^2 HH^T$. So we have

$$Var(\tilde{\beta}) = Var(\beta^*) + \sigma^2 DD^T$$

And finally :

$$\boxed{Var(\tilde{\beta}) > Var(\beta^*)}$$

2 Ridge Regression

We have $\beta^*_{ridge} = (\lambda I_d + X^T X)^{-1} X^T Y$.

$$\begin{align}
\mathbb{E}[\beta^*_{ridge}] &= \mathbb{E}[(\lambda I_d + X^T X)^{-1} X^T y] \tag{9}\\
&= (\lambda I_d + X^T X)^{-1} X^T X\theta \tag{10}\\
&= (\lambda I_d + X^T X)^{-1}(\lambda I_d + X^T X)\theta - \lambda I_d(\lambda I_d + X^T X)^{-1}\theta \tag{11}\\
&= \theta - \underbrace{\lambda I_d(\lambda I_d + X^T X)^{-1}\theta}_{bias} \tag{12}
\end{align}$$

$X = UDV^T$, so :

$$\begin{aligned}
\beta^*_{ridge} &= (\lambda I_d + X^T X)^{-1} X^T \qquad (13)\\
&= (\lambda I_d + V D^2 V^T)^{-1} V D U^T \qquad (14)\\
&= (V(\lambda + D^2) V^T)^{-1} V D U^T \qquad (15)\\
&= V(D^2 + \lambda)^{-1} D U^T \qquad (16)
\end{aligned}$$

$$\boxed{\beta^*_{ridge} = V(D^2 + \lambda)^{-1} D U^T}$$

$$\begin{aligned}
Var(\beta^*_{ridge}) &= \mathbb{E}[(V(D^2 + \lambda I_d)^{-1} D U^T \epsilon)(V(D^2 + \lambda I_d)^{-1} D U^T \epsilon)^T] \qquad (17)\\
&= \sigma^2 V (D^2 + \lambda I_d)^{-1} D^2 (D^2 + \lambda I_d)^{-1} V^T \qquad (18)
\end{aligned}$$

With $D = diag(d_i)$:

$$Var(\beta^*_{ridge}) = \sigma^2 V diag(\frac{d_i^2}{(d_i^2 + \lambda)^2}) V^T$$

So : $Var(\beta^*_{ridge}) - Var(\beta^*_{OLS}) = \sigma^2 V diag(\frac{1}{d_i^2}(\frac{1}{(1+\frac{\lambda}{d_i^2})^2} - 1)) V^T$.

If $\lambda \geq 0$:

$$\frac{1}{d_i^2}(\frac{1}{(1 + \frac{\lambda}{d_i^2})^2} - 1) \leq 0$$

And then :

$$\boxed{Var(\beta^*_{ridge}) \leq Var(\beta^*_{OLS})}$$

We have :

$$\lim_{\lambda \to +\infty} Var(\beta^*_{ridge}) = 0$$

$$\lim_{\lambda \to 0} Var(\beta^*_{ridge}) = \sigma^2 (X^T X)^{-1}$$

Since $\mathbb{E}[\beta^*_{ridge}] = V diag(\frac{d_i^2}{(d_i^2 + \lambda)}) V^T \theta$:

$$\boxed{\lim_{\lambda \to +\infty} \mathbb{E}[\beta^*_{ridge}] = 0}$$

$$\boxed{\lim_{\lambda \to 0} \mathbb{E}[\beta^*_{ridge}] = \theta}$$