PCA and Ridge Regression

$$\mathbb{E}Y = \underset{n\times p}{X}\beta = XUU^T\beta$$

where $UU^T = U^TU = I$.

$\mathbb{E}Y = Z\gamma$, $Z = XU$ & $\gamma = U^T\beta$

$(z_1, \ldots, z_p) = \left(\sum_{i=1}^{p} u_{i1}X_i, \ldots, \sum_{i=1}^{p} u_{ip}X_i\right)$

Each z_i is a linear combination of X_i's. Pick u s.t.

$$u^TX^TXu = D = \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{pmatrix}$$

$$\underbrace{X^TX}_{p\times p} = UDU^T$$

Collinearity, $\lambda_{r+1}, \lambda_{r+2}, \ldots, \lambda_p = 0$.

Then $D = \begin{pmatrix}\Lambda_r & 0 \\ 0 & 0\end{pmatrix}$, $\Lambda_r = \begin{pmatrix}\lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_r\end{pmatrix}$

$U = \{u_1 | u_2\}$

$$\begin{pmatrix}u_1^T \\ u_2^T\end{pmatrix} X^TX (u_1\ u_2) = \begin{pmatrix}\Lambda_r & 0 \\ 0 & 0\end{pmatrix}$$

$$u_1^TX^TXu_1 = \Lambda_r$$

$$(Xu_2)^TXu_2 = u_2^TX^TXu_2 = 0 \Rightarrow$$

$$Xu_2 = 0.$$

$$\mathbb{E}Y = X\beta = X(u_1\ u_2)\begin{pmatrix}u_1^T\beta \\ u_2^T\beta\end{pmatrix}$$

$$= Xu_1\gamma_1$$

Xu_1 are the p.c. of X.

1^{st} p.c. of X:

$z_1 = Xu_1$ & $X^TXu_1 = \lambda_1 u_1$

$z_r = Xu_r$ & $X^TXu_r = \lambda_r u_r$

So Reg y on $z_1 \cdots z_r$

note $z_i^Tz_j = u_i^TX^TXu_j$

$$= \begin{cases}\lambda_i & j=i \\ 0 & j\neq i\end{cases}$$

Scree plot: $\sum_{i=1}^{j}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$

Ridge Regression

Recall $Var\,\hat{\beta}_i = \frac{\sigma^2}{\|X_i^\perp\|_2^2}$. $\hat{\beta}$ = LS estimate for β.

$$\hat{\beta}^K = (X^TX + KI)^{-1}X^Ty$$

Ridge Regression estimate for β

MSE of $\tilde{\beta}$ of β:

$$mse(\hat{\beta}) = \mathbb{E}\underbrace{(\hat{\beta}-\beta)}_{p\times 1}\underbrace{(\hat{\beta}-\beta)^T}_{1\times p}$$

$$= Cov(\hat{\beta}) + bb^T, \quad b = \mathbb{E}(\hat{\beta}-\beta).$$

Clearly $b(\hat{\beta}) = 0$. $mse(\hat{\beta}) = Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$

$b_K = (X^TX + KI)^{-1}X^Ty$, $K > 0$.

$= G_KX^Ty$

$$\mathbb{E}b_K = G_KX^TX\beta = (X^TX + KI)^{-1}X^TX\beta$$

$$= (X^TX + K(X^TX)(X^TX)^{-1})^{-1}X^TX\beta$$

$$= (I + K(X^TX)^{-1})^{-1}\beta$$

250A: $(I-AB)^{-1} = I - A(I-BA)^{-1}B$

$A = -KI$, $B = (X^TX)^{-1}$

$\therefore bias(b_K) = -KI(I + (X^TX)^{-1}KI)^{-1}(X^TX)^{-1}\beta$.

$= -k((X^TX)^{-1}(X^TX) + (X^TX)^{-1}kI)^{-1}(X^TX)^{-1}\beta$

$= -k(X^TX + kI)^{-1}\beta$

$= -k G_k \beta$

$Var\, b_k = Var\, (X^TX + kI)^{-1}X^Ty$

$= G_k X^T(\sigma^2 I)X G_k$

$= \sigma^2 G_k X^TX G_k$

$mse(b_k) = \sigma^2 G_k X^TX G_k + k^2 G_k \beta\beta^T G_k$

$= G_k(\sigma^2 X^TX + k^2\beta\beta^T)G_k$

Show

$mse(\hat\beta) - mse(b_k) \geq 0$ for some k

Now show $Tr(mse(\hat\beta)) - Tr(mse(b_k)) \geq 0$ for some k.

$[mse(b_k)]$

$= (G_k\beta\beta^T G_k k^2 + G_k X^TX G_k \sigma^2)$

$= (TDT^T + kI)^{-1}(TDT^T\sigma^2 + k^2\beta\beta^T)(TDT^T + kI)^{-1}$

where $T^TX^TXT = D = \begin{pmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_p\end{pmatrix}$

$X^TX = TDT^T$ & $G_k = (X^TX + kI)^{-1}$

$= T(D+kI)^{-1}(\sigma^2 D + \beta\beta^T k^2)(D+kI)T^T \quad (\triangle)$

$Tr(\triangle) = Tr\left\{\begin{pmatrix}\frac{1}{\lambda_1+k} & & \\ & \ddots & \\ & & \frac{1}{\lambda_p+k}\end{pmatrix}(D\sigma^2 + \beta\beta^T k^2)\begin{pmatrix}\frac{1}{\lambda_1+k} & & 0\\ & \ddots & \\ 0 & & \frac{1}{\lambda_p+k}\end{pmatrix}\right\}$

$= Tr\begin{pmatrix}\frac{\lambda_1\sigma^2}{(\lambda_1+k)^2} & & \\ & \ddots & \\ & & \frac{\lambda_p\sigma^2}{(\lambda_p+k)^2}\end{pmatrix} + k^2\beta^T\begin{pmatrix}\frac{1}{(\lambda_1+k)^2} & & \\ & \ddots & \\ & & \frac{1}{(\lambda_p+k)^2}\end{pmatrix}\beta$

$= \sum_{i=1}^{p}\frac{\lambda_i\sigma^2 + k^2\beta_i^2}{(\lambda_i+k)^2}$

$= g(k)$ if σ^2, X, β.

$g(0) = \sum_{i=1}^{p}\frac{\sigma^2}{\lambda_i} = Tr(Cov(\hat\beta)) = Tr(mse(\hat\beta))$

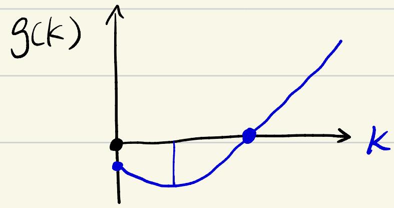
Let $g(k) - g(0) = Tr(mse(b_k)) - Tr(mse(\hat\beta))$

$= \sum_{i=1}^{p}\frac{\lambda_i\sigma^2 + k^2\beta_i^2}{(\lambda_i+k)^2} - \sum_{i=1}^{p}\frac{\sigma^2}{\lambda_i^2}$

Verify:

$g'(k) = 0 = -\sum_{i=1}^{p}\frac{2\lambda_i(k\beta_i^2 - \sigma^2)}{(\lambda_i+k)^3}$!!!

Further, $g'(k) = \begin{cases}\geq 0 & \text{if } k \text{ small}\\ \leq 0 & \text{if } k \text{ Large}\end{cases}$



Can impose unrealistic assumptions:

$k < \frac{2\sigma^2}{\beta^T\beta}$

Use SVD thm to view Ridge Reg est.

$\underset{n\times p}{X} = \underset{n\times n}{U}\,\underset{n\times p}{D}\,\underset{p\times p}{V^T}$

Verify

$\hat\theta = X\hat\beta = X(X^TX)^{-1}X^Ty = \Sigma_i u_i u_i^T y$

$Xb_k = X(X^TX + kI)^{-1}X^Ty$

$= \Sigma_i \frac{\lambda_i^2}{\lambda_i^2 + k} u_i u_i^T y$

Since $\frac{\lambda_i^2}{\lambda_i^2 + k} < 1$ $\therefore$ Xb_k shrinks $X\hat\beta$.