Ridge regression $\hat{\beta}^R_\lambda = (X^TX + \lambda I)^{-1}X^TY$

Theorem 1 Suppose rank$(X) = p$ and $\lambda > 0$ is suff. small (depending on β^0, σ^2)

Have $\mathbb{E}(\hat{\beta}^{OLS} - \beta^0)(\hat{\beta}^{OLS} - \beta^0)^T - \mathbb{E}(\hat{\beta}^R_\lambda - \beta^0)(\hat{\beta}^R_\lambda - \beta^0)^T$ (*) is pos. def

Proof Bias $\mathbb{E}\hat{\beta} - \beta^0 = (X^TX + \lambda I)^{-1}X^TX\beta^0 - \beta^0$

$\hat{\beta}^R_\lambda$

$$= (X^TX + \lambda I)^{-1}\{X^TX\beta^0 - X^TX\beta^0 - \lambda\beta^0\}$$

$$= -\lambda(X^TX + \lambda I)^{-1}\beta^0$$

$$\text{Var}(\hat{\beta}) = \sigma^2(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}$$

$$(X^TX)^{-1} = (X^TX + \lambda I)^{-1}\underbrace{(X^TX + \lambda I)(X^TX)^{-1}(X^TX + \lambda I)}_{X^TX + 2\lambda I + \lambda^2(X^TX)^{-1}}(X^TX + \lambda I)^{-1}$$

(*) pos def iff

$$\sigma^2(X^TX + 2\lambda I + \lambda^2(X^TX)^{-1}) - \sigma^2X^TX - \lambda^2\beta^0\beta^{0T} \quad \text{is pos def}$$

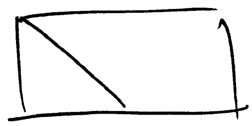
$$\lambda(2\sigma^2 I + \sigma^2\lambda(X^TX)^{-1} - \lambda\beta^0\beta^{0T})$$

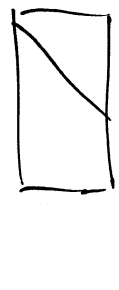
But this is pos def for $0 < \lambda < \frac{2\sigma^2}{\|\beta^0\|_2^2}$ □

The singular value decomposition (SVD)

Can factorise any $X \in \mathbb{R}^{n\times p}$ into its SVD

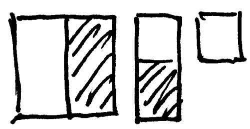
$$X = \underset{\substack{n\times n\\ \text{orthogonal}}}{U}\ \underset{n\times p}{D}\ \underset{\substack{p\times p\\ \text{orthogonal}}}{V^T}$$

$D =$ 



D has $D_{11} \geq D_{22} \geq \dots \geq D_{mm} \geq 0$ where $m = \min(n,p)$ and all other entries of D are 0.

Alternative form (when $n > p$)



Replace U by its first p cols

D by its first p rows

Have $X = \underset{n\times p}{U}\ \underset{p\times p}{D}\ \underset{p\times p}{V^T}$ U has orthonormal cols so $U^TU = I$ but $UU^T \neq I$.

Suppose $n > p$. Fitted vals of ridge regression

$$X\hat{\beta}_\lambda^R = X(X^TX+\lambda I)^{-1}X^TY$$
$$= UDV^T(VD^2V^T+\lambda I)^{-1}VDU^TY$$

Now $VD^2V^T + \lambda I = V(D^2+\lambda I)V^T$

and $\{V(D^2+\lambda I)V^T\}^{-1} = V(D^2+\lambda I)^{-1}V^T$

$$X\hat{\beta}_\lambda^R = UD(D^2+\lambda I)^{-1}DU^TY$$
$$= \sum_{j=1}^p U_j \frac{D_{jj}^2}{D_{jj}^2+\lambda} U_j^T Y$$

OLS fitted vals

$$X\hat{\beta}^{OLS} = \sum_{j=1}^p U_j U_j^T Y \qquad (\text{provided } D_{pp} > 0 \text{ so } X \text{ has full col rank})$$

Both OLS and ridge regression compute coords of Y w.r.t. the cols of U.
Ridge shrinks these coords by factor $D_{jj}^2/(D_{jj}^2+\lambda)$ whilst OLS leaves these unchanged.

Cols of U correspond to the principal components of X.

Take $u \in \mathbb{R}^p$, $\|u\|_2 = 1$. Sample variance $Xu \in \mathbb{R}^n$ is $\frac{1}{n}\|Xu\|_2^2 = \frac{1}{n}u^TX^TXu$
$= \frac{1}{n}u^TVD^2V^Tu$

Let $W = V^Tu$. Then $\|w\|_2 = 1$

$\frac{1}{n}u^TVD^2V^Tu = \frac{1}{n}W^TD^2W = \frac{1}{n}\sum_j w_j^2 D_{jj}^2 \leq \frac{1}{n}D_{11}^2$

with equality when $W = \begin{pmatrix}1\\0\\0\\\vdots\end{pmatrix}$, so $u = V_1$.

Thus V_1 gives the coeffs of the linear combination of cols of X with largest sample variance, when coeffs are constrained to have l_2-norm 1.

$$\underset{UDV^TV_1}{\underset{\parallel}{X V_1}} = U_1 D_{11} \text{ is the 1st principal component of } X.$$

Subsequent principal components obey the same optimality conditions subject to being normal to all prior principal components. Can show rth principal comp is $D_{rr}U_r$.

Conclusion:

Ridge regression works best when most of the signal EY is in the direction of the large principal compts of X.