1 Ridge Theory

Assume the typical linear model setup, that is, $Y_i = X_i^\top \beta + \epsilon_i$, for $i = 1, \ldots, n$, where

- $X_i \in \mathbb{R}^p$
- $\mathbb{E}\epsilon_i = 0$ and $\mathbb{E}\epsilon\epsilon^\top = I_n$ (w.l.o.g. $\sigma^2 = 1$)
- $\mathbb{X}$ is the design matrix, and $\text{rank}(\mathbb{X}) = p$

We will consider consistent estimates of the parameter vector β and doing good predictions.

1.1 Low Dimensional Prediction

Consider the ridge regression solution to the normal equations, $\hat{\beta}_\lambda = (\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Y$. Then, to get L^2 consistency, we need to show that the risk

$$R(\hat{\beta}_\lambda) = \mathbb{E}_{\mathcal{D}}||\hat{\beta}_\lambda - \beta||_2^2$$

goes to zero.

As in the past, we will decompose the risk into bias and variance components by,

$$\begin{aligned}
R(\hat{\beta}) &= \mathbb{E}||\hat{\beta} - \mathbb{E}\hat{\beta}||_2^2 + ||\mathbb{E}\hat{\beta} - \beta||_2^2 && (1)\\
&= \text{trace}\mathbb{V}\hat{\beta} + \sum_{j=1}^{p}(\mathbb{E}\hat{\beta}_j - \beta_j)^2 && (2)\\
&= \text{trace}\mathbb{V}\hat{\beta}_\lambda + ||\mathbb{E}\hat{\beta}_\lambda - \beta||_2^2 && (3)\\
&= \text{trace}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1} + && (4)\\
&\quad + ||((\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top\mathbb{X} - I)\beta||_2^2 && (5)\\
&= \text{variance} + \text{bias}^2 && (6)
\end{aligned}$$

Let's address each of these terms separately

For the bias, we use the Woodbury matrix inversion lemma

$$(A - BC^{-1}E)^{-1} = A^{-1} + A^{-1}B(C - EA^{-1}B)^{-1}EA^{-1}$$

with $A = -I, B = I, C = \lambda I$ and $E = \mathbb{X}^\top \mathbb{X}$
See Henderson, Searle (1980), equation (12) for a statement and discussion. [1]

$$\begin{aligned}
\text{bias}^2 &= ||((\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top \mathbb{X} - I)\beta||_2^2 && (7)\\
&= ||(I + (\mathbb{X}^\top \mathbb{X})\lambda^{-1})^{-1}\beta||_2^2 && (8)\\
&= \lambda^2 ||(\lambda I + \mathbb{X}^\top \mathbb{X})^{-1}\beta||_2^2 && (9)\\
&= \lambda^2 ||(\lambda I + V D^2 V^\top)^{-1}\beta||_2^2 && (10)\\
&= \lambda^2 ||(V(\lambda V^\top V + D^2)V^\top)^{-1}\beta||_2^2 && (11)\\
&= \lambda^2 ||(\lambda I + D^2)^{-1}\theta||_2^2 && (12)\\
&= \lambda^2 \sum_{j=1}^{p} \frac{\theta_j^2}{(\lambda + d_j^2)^2} && (13)
\end{aligned}$$

where $\theta = V^\top \beta$

Likewise,

$$\begin{aligned}
\text{variance} &= \text{trace}\left((\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1} \mathbb{X}^\top \mathbb{X} (\mathbb{X}^\top \mathbb{X} + \lambda I)^{-1}\right) && (14)\\
&= \text{trace}\left(D^2 (D^2 + \lambda I)^{-2}\right) && (15)\\
&= \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2} && (16)
\end{aligned}$$

Putting them together:

$$R(\hat{\beta}) = \sum_{j=1}^{p} \left(\frac{\lambda^2 \theta_j^2 + d_j^2}{(\lambda + d_j^2)^2} \right)$$

Now we can attempt to select a value for λ that minimizes this risk.

$$\begin{aligned}
R(\hat{\beta}) &= \sum_{j=1}^{p} \left(\frac{\lambda^2 \theta_j^2 + d_j^2}{(\lambda + d_j^2)^2} \right) && (17)\\
\Rightarrow \frac{\partial \hat{R}(\hat{\beta})}{\partial \lambda} &= \sum_{j=1}^{n} \frac{2 d_j^2 (\lambda \theta_j^2 - 1)}{(\lambda + d_j^2)^3} && (18)
\end{aligned}$$

This suggests taking $\hat{\lambda} = 1/\theta_{\max}^2$. Observe

$$R(\hat{\beta}_{\hat{\lambda}}) = \sum_{j=1}^{p} \left(\frac{\theta_j^2/\theta_{\max}^4 + d_j^2}{(1/\theta_{\max}^2 + d_j^2)^2} \right) \leq \sum_{j=1}^{p} \left(\frac{1/\theta_{\max}^2 + d_j^2}{(1/\theta_{\max}^2 + d_j^2)^2} \right) = \sum_{j=1}^{p} \left(\frac{1}{(1/\theta_{\max}^2 + d_j^2)} \right) < \sum_{j=1}^{p} \left(\frac{1}{d_j^2} \right)$$

as long as $\theta_{\max}^2 < \infty$.

The point is that the risk for estimation of the parameter vector in a linear model can always be improved with ridge regression with an appropriate selection of the tuning parameter λ.

1.2 High Dimensional Prediction

In the case where $p > n$, there is an issue as the parameter vector is non-identified. That is,

$$Y = \mathbb{X}(\beta + b) + \epsilon$$

for any b in the null space of $\mathbb{X}$.

Example 1.1. *For example, if,*

$$X = \begin{pmatrix} 1 & 2 & 1 & 4 \\ 0 & 4 & 7 & 1 \\ 2 & 4 & 2 & 8 \end{pmatrix}, \beta = \begin{pmatrix} 2 \\ 1 \\ 3 \\ 4 \end{pmatrix}$$

then,

$$X\beta = X\left(\beta + \begin{pmatrix} 0.146 \\ -0.842 \\ 0.442 \\ 0.274 \end{pmatrix}\right) = \begin{pmatrix} 23 \\ 29 \\ 46 \end{pmatrix}$$

If we let $(\mathbb{X}) = r$ (and hence $r < p$), then the SVD matrices have dimension,

- $U \in \mathbb{R}^{n\times r}$
- $D \in \mathbb{R}^{r\times r}$
- $V \in \mathbb{R}^{p\times r}$
- $V_\perp \in \mathbb{R}^{p\times(p-r)}$ be orthonormal and $V^\top V_\perp = 0$

Lemma 1.2. *β is identifiable if and only if there exists a known function ϕ from $\mathcal{R}^r$ to $\mathcal{R}^{p-r}$ such that*

$$\mathbf{B} = \{\beta : \beta = V\xi + V_\perp\phi(\xi), \xi \in \mathcal{R}^r\} \tag{19}$$

Thus, identifiable β's must be in a set having a one-to-one correspondence with the range space of $\mathbb{X}$.

Typically, the rank of X is equal to the sample size.

Let $\theta = \mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^\dagger\mathbb{X}\beta = VV^\top\beta$ be the projection of β onto the range space of $\mathbb{X}$. Then, $\theta \in \mathbb{R}^p$ and $Y = \mathbb{X}\theta + \epsilon$ and hence estimating θ is enough for predictions.

Now, we form

$$\hat{\theta} = (\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top Y$$

Again, using the woodbury matrix inversion lemma and defining $\Gamma = [V, V_\perp]$ we can write the bias as,

$$\begin{aligned}
\text{Bias}(\hat{\theta}) &= \mathbb{E}\hat{\theta} - \theta && (20)\\
&= -(\lambda^{-1}\mathbb{X}^\top\mathbb{X} + I)^{-1}\theta && (21)\\
&= -\Gamma(\lambda^{-1}\Gamma^\top\mathbb{X}^\top\mathbb{X}\Gamma + I)^{-1}\Gamma^\top VV^\top\theta && (22)\\
&= -[V, V_\perp]\begin{bmatrix} (\lambda^{-1}D^2 + I_r)^{-1} & 0 \\ 0 & I_{p-r} \end{bmatrix}\begin{bmatrix} V^\top \\ V_\perp^\top \end{bmatrix} VV^\top\theta && (23)\\
&= -[V(\lambda^{-1}D^2 + I_r)^{-1}, V_\perp]\begin{bmatrix} V^\top\theta \\ 0 \end{bmatrix} && (24)\\
&= -V(\lambda^{-1}D^2 + I_r)^{-1}V^\top\theta && (25)
\end{aligned}$$

Γ is such that $\Gamma\Gamma^\top = \Gamma^\top\Gamma = I$

We can make a somewhat simple bound on the variance:

$$\mathbb{V}\hat{\theta} = (\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top\mathbb{X}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1} \tag{26}$$
$$\leq (\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1} \tag{27}$$
$$\leq \lambda^{-1} I \tag{28}$$

where $A \leq B$ means $B - A$ is nonnegative definite. Using the bias representation and variance bound we can bound the risk using the following theorem.

Theorem 1.3. *Shao and Deng, 2012, Theorem 1*
There exists a constant C such that for n large enough

$$n^{-1}\mathbb{E}||\mathbb{X}(\hat{\theta} - \theta)||_2^2 \leq C\left(\frac{r}{n} + \lambda^2 n^{-(1+\eta-2\tau)}\right)$$

where $d_{\min}^{-2} \leq n^{-\eta}$ and $||\theta||_2 \leq n^\tau$

Note that
$$||\beta||_2^2 \geq ||VV^\top\beta||_2^2 = ||\theta||_2^2$$

Proof.
$$\mathbb{E}||\mathbb{X}(\hat{\theta} - \theta)||_2^2 = \mathrm{trace}(\mathbb{X}\mathbb{V}[\hat{\theta}]\mathbb{X}^\top) + ||\mathbb{X}\mathrm{bias}(\hat{\theta})||_2^2.$$
Using the variance bound
$$\mathbb{X}\mathbb{V}[\hat{\theta}]\mathbb{X}^\top \leq \mathbb{X}(\mathbb{X}^\top\mathbb{X} + \lambda I)^{-1}\mathbb{X}^\top \leq UU^\top$$
Since $UU^\top$ is a rank r projection, we get,
$$\mathrm{trace}(\mathbb{X}\mathbb{V}[\hat{\theta}]\mathbb{X}^\top) \leq \mathrm{trace}(UU^\top) = r$$

Also, from the alternative bias representation we have,

$$||\mathbb{X}\mathrm{bias}(\hat{\theta})||_2^2 = ||UD(\lambda^{-1}D^2 + I_r)^{-1}V^\top\theta||_2^2 \tag{29}$$
$$\leq ||D(\lambda^{-1}D^2 + I_r)^{-1}||_2^2||\theta||_2^2 \tag{30}$$
$$\leq (\max_j \frac{\lambda^2}{d_j^2})^2||\theta||_2^2 = \lambda^2 d_{\min}^{-2}||\theta||_2^2 \tag{31}$$

■

So,
$$n^{-1}\mathbb{E}||\mathbb{X}(\hat{\theta} - \theta)||_2^2 \leq \frac{r}{n} + \frac{\lambda^2 d_{\min}^{-2}||\theta||_2^2}{n}$$

1.3 Thresholding for a Faster Rate

Shao and Deng (2012) [2] introduce a strongly thresholded estimator $\widetilde{\theta}$ defined as $\widetilde{\theta} = \hat{\theta}$ if $|\hat{\theta}| > a_n$ and $\widetilde{\theta} = 0$ if $|\hat{\theta}| \leq a_n$ where
$$a_n = C_1 n^{-\alpha}, \qquad 0 < \alpha \leq 1/2, C_1 > 0$$
The result is a faster rate of convergence and smaller risk as shown in the plots below.

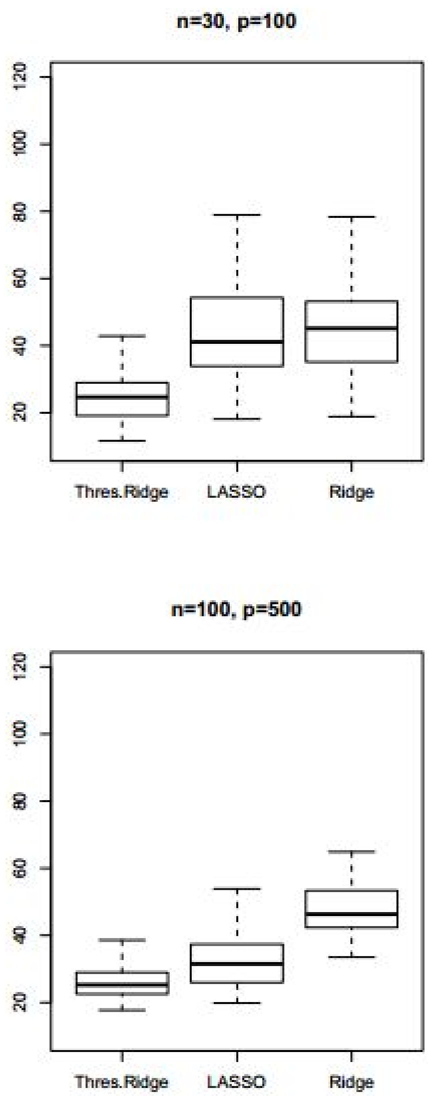

Figure 1:

References

[1] Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23:53–60.

[2] Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40(2):812–831.