# STATS305A - Lecture 13

John Duchi
Scribed by Michael Howes

11/04/21

## Contents



## 1 Announcements

- Exercise 3 to be released today.
- An email announcement will be sent when exercise 3 is available.

## 2 Risk and model selection

Recall that we are now looking at model selection and prediction. We are starting with the more general model $y = f(x) + \varepsilon$ where $f(x) = \mathbb{E}[y|x]$ and we want to fit a predictor $h : \mathcal{X} \to \mathbb{R}$.

We can think of having a sequence of model spaces $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_k \subset \ldots$. We want to find $h$ such that the quantity

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2,$$

is small but $h$ is not too "complex". Typically we think of $\mathcal{H}_k$ as growing in complexity as $k$ increases. Thus we wish to choose $h \in \mathcal{H}_k$ minimizing

$$\sum_{i=1}^{n}(y_i - h(x_i))^2 + \text{penalty}(h),$$

or

$$\sum_{i=1}^{n}(y_i - h(x_i))^2 + \text{penalty}(\mathcal{H}_k).$$

Talking about this last time led to Mallows's $C_p$ statistic which was

$$C_p = \|y - X\beta\|_2^2 + 2\sigma^2 p,$$

where $\beta \in \mathbb{R}^p$ and $2\sigma^2 p$ is a complexity penalty. Mallows's method is to choose the model that minimizes $C_p$.

### 2.1 Bias variance tradeoff

For any $x \in \mathcal{X}$ and any model procedure $\widehat{h}$ we ahve

$$\begin{aligned}
\mathbb{E}[(\widehat{h}(x) - f(x))^2] &= \mathbb{E}\left[(\widehat{h}(x) - \mathbb{E}[\widehat{h}(x)] + \mathbb{E}[\widehat{h}(x)] - f(x))^2\right] \\
&= \Big(\mathbb{E}[\widehat{h}(x)] - f(x)\Big)^2 + \mathbb{E}[\Big(\widehat{h}(x) - \mathbb{E}[\widehat{h}(x)]\Big)^2] \\
&= \text{Bias}(\widehat{h}(x))^2 + \text{Var}(\widehat{h}(x)).
\end{aligned}$$

Recall that we defined the in-sample risk to be

$$R_{in}(\widehat{h}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\Big(\widehat{h}(x_i) - f(x_i)\Big)^2],$$

where the expectation is over $Y_i = f(x_i) + \varepsilon_i$.

**Example 1.** (actually everything we will look at) Let $\mu = [f(x_i)]_{i=1}^n = \mathbb{E}[Y]$. Suppose

$$\widehat{y} = Hy = [\widehat{h}(x_i)]_{i=1}^n,$$

for some matrix $H$ that is not necessarily a projection matrix or symmetric. If $\varepsilon \sim (0, \sigma^2 I_n)$, then

$$\begin{aligned}
nR_{in}(\widehat{h}) &= \mathbb{E}\left[\|\widehat{y} - \mu\|_2^2\right] \\
&= \mathbb{E}\left[\|Hy - \mu\|_2^2\right] \\
&= \mathbb{E}\left[\|H\mu - H\varepsilon - \mu\|_2^2\right] \\
&= \mathbb{E}\left[\|H\varepsilon - (I - H)\mu\|_2^2\right] \\
&= \|(I - H)\mu\|_2^2 + \mathbb{E}[\|H\varepsilon\|_2^2] \\
&= \text{Bias}(H)^2 + \sigma^2 \operatorname{tr}(H^T H).
\end{aligned}$$

Likewise for the residual sum of squares (RSS) we have

$$\begin{aligned}
\mathbb{E}[RSS] &= \mathbb{E}[\|\widehat{y} - y\|_2^2] \\
&= \mathbb{E}[\|y - Hy\|_2^2] \\
&= \mathbb{E}[\|(I - H)\mu + (I - H)\varepsilon\|_2^2] \\
&= \|(I - H)\mu\|_2^2 + n\sigma^2 - 2\sigma^2 \operatorname{tr}(H) + \sigma^2 \operatorname{tr}(H^T H) \\
&= nR_{in}(\widehat{h}) + n\sigma^2 - 2\sigma^2 \operatorname{tr}(H).
\end{aligned}$$

## 3 Ridge regression

### 3.1 Idea

If $X$ is "ill-conditioned", then some errors in $y$ can induce changes in $\widehat{\beta}$ that are too large if we are using OLS.

We can formalize this with the SVD of $X$. Suppose $X = U\Gamma V^T$ is the SVD of $X$ and $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_d)$. Then

$$\begin{aligned}\widehat{\beta} &= (X^T X)^{-1} X^T y \\ &= V\Gamma^{-1} U^T y.\end{aligned}$$

Suppose $U = [u_1, \ldots, u_d] \in \mathbb{R}^{n \times d}$. Consider two alternative responses

$$\begin{aligned}y' &= y + tu_1, \quad t \in \mathbb{R}, \\ y'' &= y + tu_d, \quad t \in \mathbb{R}.\end{aligned}$$

We thus have

$$\begin{aligned}\widehat{\beta}(y') &= \widehat{\beta}(y) + \frac{tv_1}{\gamma_1} \\ \widehat{\beta}(y'') &= \widehat{\beta}(y) + \frac{tv_d}{\gamma_d}.\end{aligned}$$

If $\gamma_d$ is very small (say $\gamma_d = \frac{1}{1000}$), then $\widehat{\beta}(y'')$ will be very far from $\widehat{\beta}(y)$ even if $t$ is small. This issue arises when $\frac{\gamma_1}{\gamma_d}$ is very large. The quantity $\text{cond}(X) = \frac{\gamma_1}{\gamma_d}$ is called the *condition number* of $X$. In numerical analysis, $\text{cond}(X) \geq 10^6$ leads to problems. In statistics things are more sensitive and we run into trouble if $\text{cond}(X) \gtrsim 100$.

## 3.2 Ridge estimator

One solution to the sensitivity of $\widehat{\beta}$ is to regularize $\widehat{\beta}$ by forcing $\|\widehat{\beta}\|_2^2$ to be small.

**Definition 1.** For $\lambda \geq 0$, define the *ridge estimator* to be

$$\widehat{\beta}_\lambda := \underset{b}{\text{argmin}} \left\{ \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}.$$

Since $\partial_b \left( \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right) = X^T Xb - X^T y + \lambda b$, we have that

$$\widehat{\beta}_\lambda = (X^T X + \lambda I)^{-1} X^T y.$$

This method of estimating $\beta$ is called ridge regression or Tikhonov regularization.

**Fact 1.** *If $X = U\Gamma V^T$, then*

$$\widehat{y}_\lambda = X\widehat{\beta}_\lambda = \sum_{j=1}^{d} \frac{\gamma_j^2}{\gamma_j^2 + \lambda} u_j u_j^T y = H_\lambda y,$$

*where*

$$H_\lambda = U \, \text{diag}\left( \frac{\gamma_j^2}{\gamma_j^2 + \lambda} \right) U^T.$$

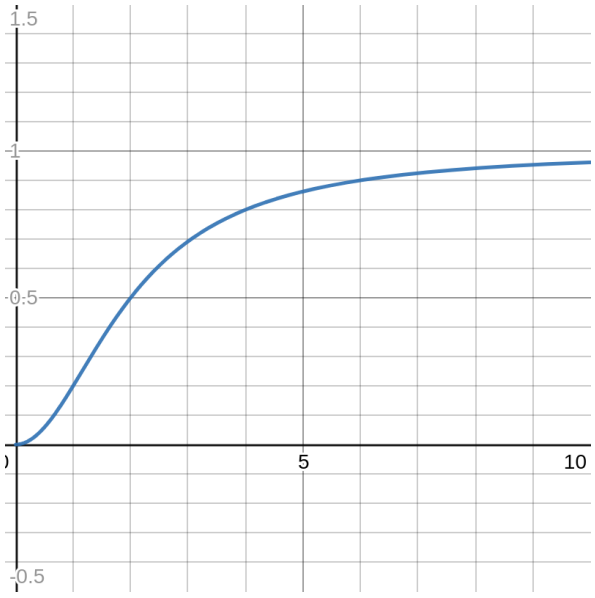
*Also*

$$\frac{\partial}{\partial \lambda} R_{in}(\widehat{\beta}_\lambda)|_{\lambda=0} = -\frac{2\sigma^2 \, \text{tr}(\Gamma^{-2})}{n} < 0.$$

Some thoughts on the above facts before we prove them. The normal OLS estimator satisfies

$$\begin{aligned}\widehat{y} &= Hy \\ &= UU^T y \\ &= \sum_{j=1}^{d} u_j u_j^T y \\ &= \Pi_{\text{range}(X)} y.\end{aligned}$$

On the other hand $\widehat{y}_\lambda$ is the projection of $y$ onto $\text{range}(X) = \text{span}\{u_i\}$ but then shrunk by $\frac{\gamma_j^2}{\gamma_j^2+\lambda}$. The function $\gamma_j \mapsto \frac{\gamma_j^2}{\gamma_j^2+\lambda}$ looks like this ($\lambda = 4$ was used in this plot):




When $\gamma_j = 0$, $\frac{\gamma_j^2}{\gamma_j^2+\lambda} = 0$ and as $\gamma_j \to \infty$, $\frac{\gamma_j^2}{\gamma_j^2+\lambda} \to 1$. When $\gamma_j = \sqrt{\lambda}$, $\frac{\gamma_j^2}{\gamma_j^2+\lambda} = \frac{1}{2}$. The idea is that for small values of $\gamma_j$, there is too much variance and we shrink away. When $\gamma_j$ is large we have lot of "juice" in that direction and we do not shrink much.

**Remark 1.** In the above we are shrinking towards 0 but typically we either replace $y$ with $y - \bar{y}\mathbf{1}$ or use

$$\widehat{\beta}_\lambda = \underset{(b_0,b)}{\text{argmin}} \left\{ \|Xb + b_0\mathbf{1} - y\|_2^2 + \lambda \|b\|_2^2 \right\}.$$

Either of these changes mean that we shrink toward $\bar{y}$ instead of 0 and we do not regularize the intercept term. Note that if we replace $y$ with $y - \bar{y}\mathbf{1}$, our estimator becomes

$$\widehat{y}_\lambda = \bar{y}\mathbf{1} + U \,\text{diag}\left(\frac{\gamma_j^2}{\gamma_j^2+\lambda}\right) U^T(y - \bar{y}\mathbf{1}),$$

thus we can see that we shrink towards $\bar{y}$ if there is little info in $X$.

The statement $\frac{\partial}{\partial\lambda} R_{in}(\widehat{\beta}_\lambda)|_{\lambda=0}$ says that there is always a ridge estimator that beats OLS estimation. Thus introducing a bit of bias always improves estimation (even if $y = X\beta + \varepsilon$ is true).

*Proof of Fact.* We know $\widehat{y}_\lambda = X\widehat{\beta}_\lambda$ and that $\widehat{\beta}_\lambda = (X^TX + \lambda I)^{-1}X^Ty$, then $\widehat{y}_\lambda = H_\lambda y$ where

$$\begin{aligned} H_\lambda &= X(X^TX + \lambda I)^{-1}X^T \\ &= U\Gamma V^T(V\Gamma^2V^T + \lambda I)^{-1}V\Gamma U^T \\ &= U\Gamma V^T(V(\Gamma^2 + \lambda I)V^T)^{-1}V\Gamma U^T \\ &= U\Gamma(\Gamma^2 + \lambda I)^{-1}\Gamma U^T \\ &= U\,\mathrm{diag}\left(\frac{\gamma_j^2}{\gamma_j^2 + \lambda}\right)U^T. \end{aligned}$$

Thus we have proved the first part of the fact. For the second part define $\dot{H}_\lambda = \frac{\partial}{\partial\lambda}H_\lambda$, then

$$\dot{H}_\lambda = U\,\mathrm{diag}\left(\frac{d}{d\lambda}\left(\frac{\gamma_j^2}{\gamma_j^2 + \lambda}\right)\right)U^T = -U\,\mathrm{diag}\left(\frac{\gamma_j^2}{(\gamma_j^2 + \lambda)^2}\right)U^T.$$

Recall that

$$\begin{aligned} nR_{in}(\widehat{\beta}_\lambda) &= \|(I - H_\lambda)\mu\|_2^2 + \sigma^2\,\mathrm{tr}(H_\lambda^2) \\ &= \mu^T(I - 2H_\lambda + H_\lambda^2)\mu + \sigma^2\,\mathrm{tr}(H_\lambda^2) \end{aligned}$$

Thus

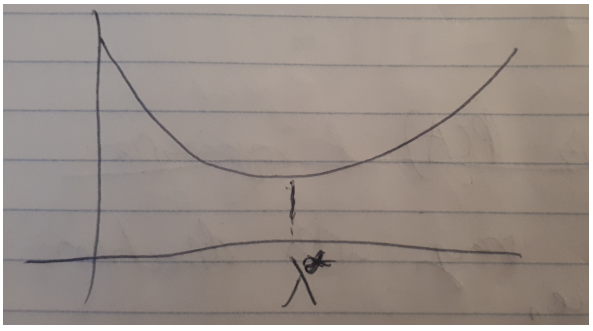
$$\frac{\partial}{\partial\lambda}nR_{in}(\widehat{\beta}_\lambda) = -2\mu^T\dot{H}_\lambda\mu + 2\mu^TH_\lambda\dot{H}_\lambda\mu + 2\sigma^2\,\mathrm{tr}(H_\lambda\dot{H}_\lambda).$$

When $\lambda = 0$, $H_\lambda = UU^T$ and $H_\lambda\dot{H}_\lambda = \dot{H}_\lambda = \dot{H}_0$. Thus

$$\begin{aligned} \frac{\partial}{\partial\lambda}\left(\frac{n}{2}R_{in}(\widehat{\beta}_\lambda)\right)\Big|_{\lambda=0} &= -\mu^T\dot{H}_0\mu + \mu^T\dot{H}_0\mu + \sigma^2\,\mathrm{tr}(\dot{H}_0) \\ &= \sigma^2\,\mathrm{tr}(\dot{H}_0) \\ &= -\sigma^2\,\mathrm{tr}\left(U\Gamma^{-2}U^T\right) \\ &= -\sigma^2\,\mathrm{tr}(\Gamma^{-2}). \end{aligned}$$

□

John said that the matrix derivatives are not important for this course but it's good for us to see the calculation and understand the implication that the OLS estimator is worse than an ridge estimator. Too much regression can be a bad thing. If $\lambda \to \infty$, then $\widehat{y}_\lambda = \bar{y}\mathbf{1}$ which is not a very good predictor. The risk as a function of $\lambda$ tends to look something like this:



A natural question is can we find $\lambda^*$? The number $\lambda^*$ is the value of $\lambda$ with the smallest risk. We know that

$$\mathbb{E}[RSS_\lambda] = nR_{in}(\widehat{\beta}_\lambda) + n\sigma^2 - 2\sigma^2\,\mathrm{tr}(H_\lambda).$$

Thus if we can estimate $\sigma^2$ with $\widehat{\sigma}^2$, then we can choose $\lambda$ to minimize

$$\mathbb{E}[RSS_\lambda] + 2\widehat{\sigma}^2\,\mathrm{tr}(H_\lambda).$$

On the homework we will look at different estimators of $\widehat{\sigma}^2$.

### 3.3 Regularization path

We often want to compute $\widehat{\beta}_\lambda$ for all $\lambda \geq 0$. This gives us a path of solutions parametrized by $\lambda$ which is called a *regularization path*. This can be done very efficiently if we have the SVD of $X$. Suppose $X = U\Gamma V^T$, then

$$\begin{aligned}\widehat{\beta}_\lambda &= (X^T X + \lambda I)^{-1} X^T y \\ &= V\left(\Gamma^2 + \lambda I\right)^{-1} \Gamma U^T y \\ &= \sum_{i=1}^{d} v_j \frac{\gamma_j}{\gamma_j^2 + \lambda} u_j^T y.\end{aligned}$$

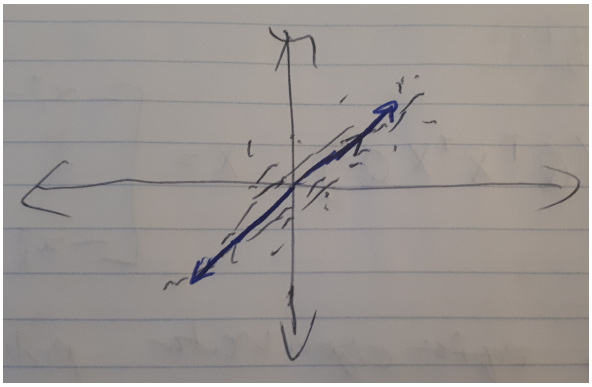
If we store $\widetilde{y} = U^T y$, then we can compute

$$\widehat{\beta}_\lambda = V \operatorname{diag}\left(\frac{\gamma_j}{\gamma_j + \lambda}\right) \widetilde{y},$$

which can be done in $O(d^2)$ time.

## 4 Principal component analysis (PCA) and regression

Idea: Often $y$ should vary with directions in $x$ that have high variance. This is the idea of *principal component analysis (PCA)*: find directions in $X$-space that vary the most (equivalently find direction in $X$-space that most accurately reconstruct our data $\{x_i\}$ ). For example if our data looks like the below picture, we wish to find the blue line.



Suppose we have a subspace $\mathcal{S} = \operatorname{span}\{q_1, \ldots, q_k\}$ where $q_i \in \mathbb{R}^d$ are orthonormal, that is

$$q_i^T q_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Then the projection onto $\mathcal{S}$ is $\Pi_{\mathcal{S}} = QQ^T$ where $Q = [q_1, \ldots, q_k] \in \mathbb{R}^{d \times k}$. Then for all $x \in \mathbb{R}^d$,

$$\begin{aligned}\operatorname{dist}^2(x, \mathcal{S}) &= \|x - \Pi_{\mathcal{S}} x\|_2^2 \\ &= \left\|(I - QQ^T)x\right\|_2^2 \\ &= \|x\|_2^2 - \left\|Q^T x\right\|_2^2.\end{aligned}$$

This gives us a procedure for PCA. We wish to find a $k$-dimensional subspace which is closest to all our data $x_1, \ldots, x_n$. Thus we wish to solve

$$\begin{aligned}
\min_{Q^TQ=I_k} \sum_{i=1}^n \text{dist}^2(x_i, \text{span}(Q)) &= \min_{Q^TQ=I_k} \sum_{i=1}^n \left\|(I - QQ^T)x_i\right\|_2^2 \\
&= \min_{Q^TQ=I_k} \sum_{i=1}^n \|x\| - \left\|Q^Tx_i\right\|_2^2 \\
&= \max_{Q^TQ=I_k} \sum_{i=1}^n \left\|Q^Tx_i\right\|_2^2 \\
&= \max_{Q^TQ=I_k} \text{tr}(Q^TX^TXQ),
\end{aligned}$$

where

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

**Remark 2.** This is an eigenvalue/eigenvector problem. If $X = U\Gamma V^T$, then our problem becomes

$$\max_{Q^TQ=I_k} \text{tr}(Q^TV\Gamma^2V^TQ).$$

Our idea: put all "juice" for the first $k$ directions in $V$ into $Q$ since

$$\Gamma^2 = \text{diag}(\gamma_1^2, \ldots, \gamma_d^2),$$

where $\gamma_1 \geq \ldots \geq \gamma_d$. So $Q = [v_1, \ldots, v_k]$. For reference this concept is called:

- Variational characterization of eigenvectors.
- Rayleigh-Ritz eigenvector characterization.