

STAT406- Methods of Statistical Learning Lecture 5

Matias Salibian-Barrera

UBC - Sep / Dec 2018

Comparing models

- For Gaussian errors we have

$$\text{AIC} = n \log \left(\frac{\text{RSS}}{n} \right) + 2p + \text{constant}$$

where

$$\text{RSS} = \sum_{i=1}^n r_i^2,$$

the **constant** depends on n , not on p

Comparing models

- However, many times we find

$$\text{AIC} = \frac{1}{n} \frac{1}{\hat{\sigma}^2} \left(\text{RSS} + 2 p \hat{\sigma}^2 \right) + \text{constant}$$

(e.g. [JWHT13])

Where does this expression come from?

Comparing models

- Regularity assumptions are needed
 - This is an asymptotic approximation, n should be large
 - One of the models should include truth
 - $\theta_1 \neq \theta_2 \Rightarrow f(y, \theta_1) \neq f(y, \theta_2)$
 - Standard large-sample MLE assumptions to obtain asymptotic normality

Comparing models

- AIC suggests a submodel
- Prediction-wise the full model is better
- AIC can be highly variable

“Smoother” model selection

- Ridge regression
- Can be thought as a type of “prediction taming”
- It is a member of a larger class called “shrinkage methods”
- However, its origins are rather different

Without loss of generality...

- If covariates are centered, $\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$

$$\arg \min_{\alpha, \beta} \sum_{i=1}^n (y_i - \alpha - \beta' \mathbf{x}_i)^2$$

satisfies

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n,$$

and

$$\hat{\beta}_{LS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y},$$

Without loss of generality...

- We can always assume that

$$\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$$

and hence

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}_n ,$$

- In what follows, there is no intercept

Shrinkage methods

- When covariates are correlated, LS estimators can be highly variable

$$\hat{\beta}_{LS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

$$\text{var}(\hat{\beta}_n) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- When $\mathbf{X}'\mathbf{X}$ is close to singular...

Ridge Regression

- One way to “avoid” this problem is to add a “ridge” to $\mathbf{X}'\mathbf{X}$...

$$\hat{\beta}_{RR} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y}$$

where $\lambda > 0$ and

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & \dots & \ddots & 0 \\ 0 & \dots & \dots & 1 \end{pmatrix}$$

Ridge Regression

- This is equivalent to solving

$$\min_{\beta} \sum_{i=1}^n (y_i - \beta' \mathbf{x}_i)^2 + \lambda \|\beta\|_2^2$$

Ridge Regression

- And also equivalent to solving

$$\min_{\beta} \sum_{i=1}^n (y_i - \beta' \mathbf{x}_i)^2$$

subject to

$$\sum_{j=1}^p \beta_j^2 \leq C$$

for some $C > 0$

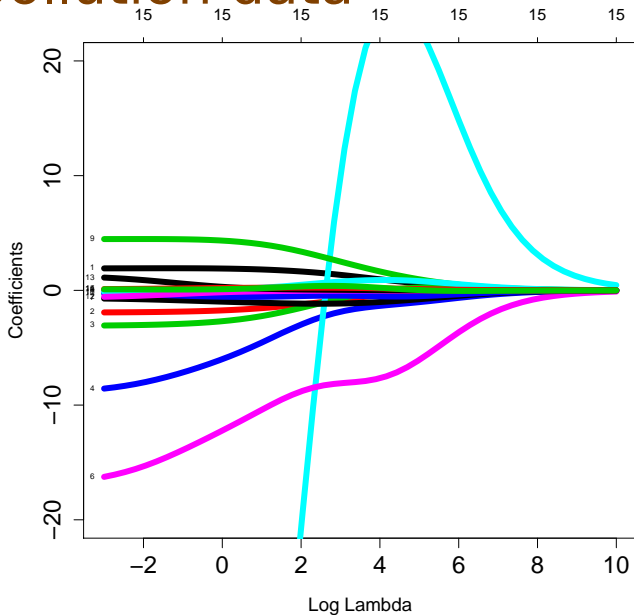
Bias / variance trade-off

- Ridge regression was originally proposed as a “hack” to “push” $\mathbf{X}'\mathbf{X}$ away from singularity
- It can also be thought as a way of reducing the variance of $\hat{\beta}_n$
- This may increase the bias of the estimator, but if the variance is reduced even more, we might gain overall in expected squared error performance...

Ridge regression

- We now have a sequence (“path”) of estimators (one for each $\lambda > 0$)
- $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p$ is always non-singular for $\lambda > 0$ (why?)
- Why are they called “shrinkage methods”?

Air pollution data



Questions

- What does λ measure?
- How do I choose one among these infinitely many “solutions”?

Effective degrees of freedom

- How many “effective” parameters are we using?
- In linear regression, we have p parameters
- A more general definition is as follows. For a fitting method producing $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$,

$$\text{edf} = \frac{1}{\sigma^2} \sum_{i=1}^n \text{cov}(\hat{y}_i, y_i)$$

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association, 81(394):461-470.

Effective degrees of freedom

- It is easy to see that for least squares predictors, we have

$$\hat{\mathbf{y}} = \mathbf{H} \mathbf{y}$$

with

$$\mathbf{H} = \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$$

and

$$\text{edf} = \frac{1}{\sigma^2} \sum_{i=1}^n \text{cov}(\hat{y}_i, y_i) = \text{trace}(\mathbf{H}) = p$$