PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY *of* WASHINGTON

Introduction to Machine Learning

MLEARN 510A – Lesson 7

Recap of Lesson 6

- Resampling Methods
- Validation Set Approach
- Leave-One-Out Cross Validation (LOOCV)
- LOOCV vs. k-fold Cross Validation
- Bias-Variance Tradeoff for Cross Validation
- The Bootstrap Method

Course Outline

1. Introduction to Statistical Learning
2. Linear Regression
3. Classification
4. Model Building, Part 1
5. Model Building, Part 2
6. Resampling Methods
7. **Linear Model Selection and Regularization**
8. Time Series Analysis
9. Frequent Itemset Mining
10. Dimensionality Reduction

Outline of Lesson 7

- Improving Linear Models – Prediction Accuracy and Model Interpretability
- Shrinkage Methods
- Ridge Regression
- Lasso Regression
- Comparison of Shrinkage Methods

Recall Least Squares Model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 \quad = \quad \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Motivation for the Alternatives:

- ➢ Improving prediction accuracy
- ➢ Increasing model interpretability

Prediction Accuracy

- Least squares estimates have low bias and low variance
  - when the relationship between label and features is linear
  - the number of observations n is much bigger than the number of predictors p ($n \gg p$)

- When $n \approx p$, then the least squares fit can have high variance and may result in over fitting and poor performance on test data

- When $n < p$, then the variability of the least squares fit increases dramatically, and the variance of these estimates in infinite

- **Solution**: *Shrink* or *control* the coefficient estimates to reduce the variance at the cost of some increase in bias

Model Interpretability

- Not all features in the model are associated with the label

- Leaving these variables leads to unnecessary complexity in the resulting model

- The model would be easier to interpret by removing the irrelevant variables

- Need an automated way to ‘zero out’ coefficients of these features

Alternative to Least Squares

- **Subset selection (aka Wrapping Methods).** Identify a subset of the *p* predictors that is believed to be related to the response

- **Shrinkage (aka Embedded Methods).** Fit a model involving all predictors, but some of the coefficients are shrunk towards zero with little loss in performance metrics

- **Dimension reduction.** Project the p predictors into a M-dimensional subspace (M < p). The new predictors are used for typical least squares fitting

Adjusting Error Estimates

- Training MSE underestimates the test MSE
- We can decrease training MSE (or increase R^2) by including more variables in the model
- Training set RSS and training set R^2 cannot be used to select from among a set of models with different numbers of variables
- *Need to adjust* the training error for the model size being used

Metrics to Adjust Error Estimate

- Colin Mallow’s selection Criterion for a model with *“d”* predictors

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

- Akaike’s Information Criterion (AIC)

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

- Bayesian Information Criterion (BIC)

$$\text{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right)$$

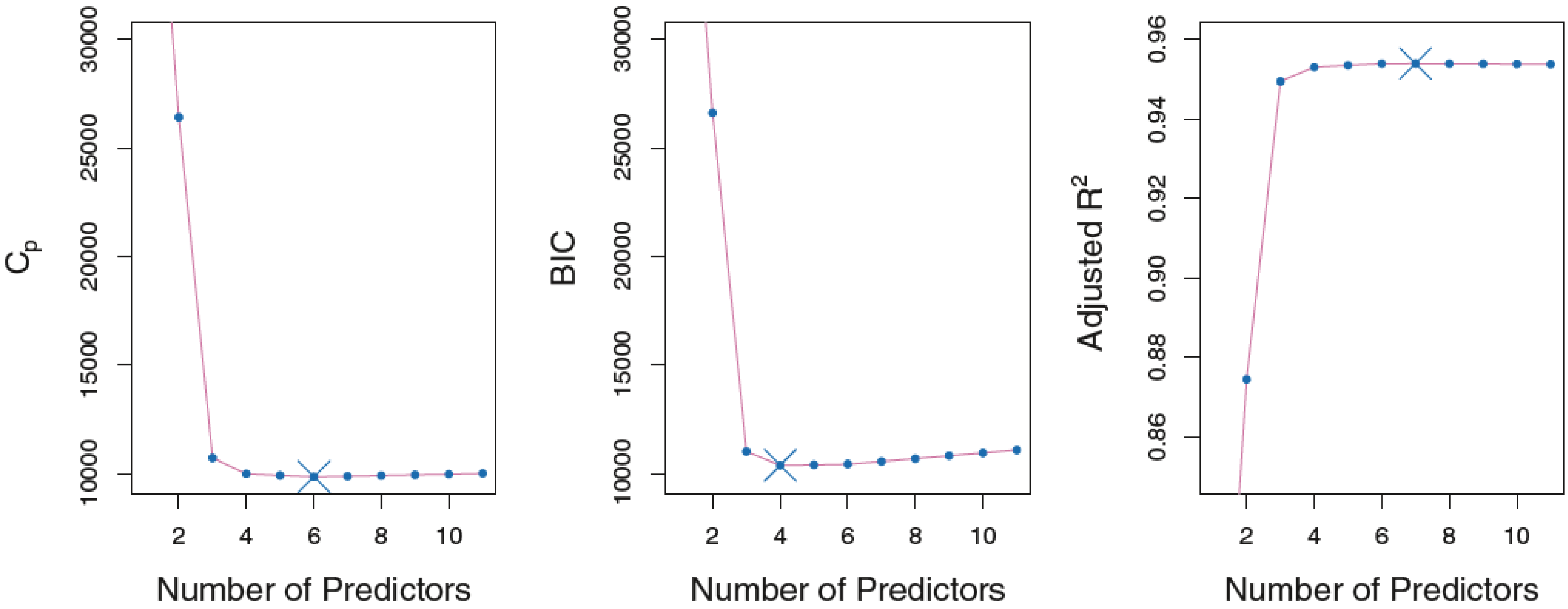
- Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

Comparing Adjusted Metrics

- Metrics tend to have a small value for models with a low test error
- For least squares models, *Cp* and AIC are proportional
- BIC replaces the $2d\hat{\sigma}^2$ used by *Cp* with a $\log(n)d\hat{\sigma}^2$ term, where *n* is the number of observations
- BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than *Cp*
- Ultimately, all metrics have strong theoretical justifications

Comparing Adjusted Metrics



Estimating Test Error

- Indirectly

➢ Estimate test error by making an adjustment to the training error to account for the bias due to overfitting

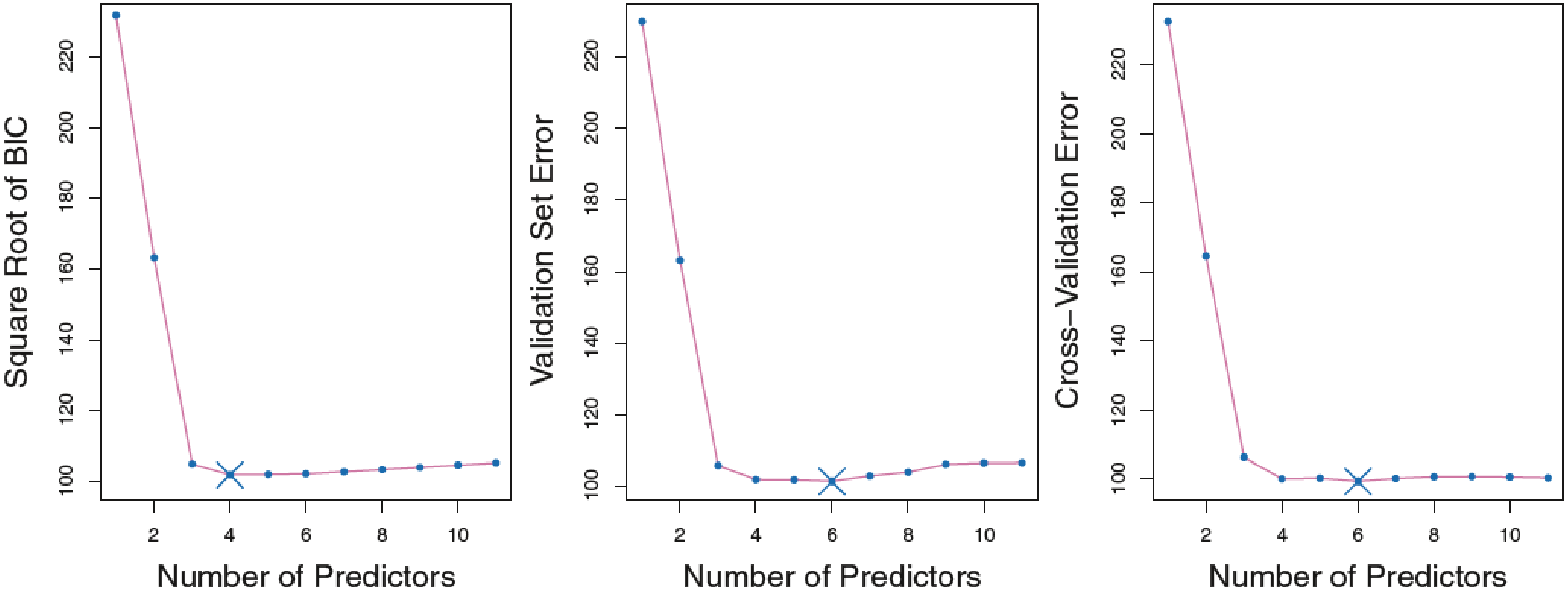
Adjustments are never perfect!

- Directly

➢ Using either a validation set approach or a cross-validation approach, as discussed in previous lectures

Use this one, when possible

Choosing Optimal Model



Shrinkage Methods

- Fit a model containing all *p* predictors using a technique that *constrains* or *regularizes* or *shrinks* the coefficient estimates

- The two best-known techniques for shrinking the regression coefficients

- Ridge Regression

- Lasso Regression

Regularization

- Shrinkage methods come within the realm of Regularization

- ***Regularization*** *is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error*

- How does regularization help?
  - Encourages a more parsimonious description of the model
  - Prevents the weights/learned parameters from becoming too large
  - Smaller weights generate a simpler model and help avoid overfitting

Regularization as Constrained Optimization

- Minimize some loss function while limiting the model complexity

  minimize Loss(Data|Model)

  such that complexity(Model) <= t

- The regularized objective function is written as

  minimize Loss(Data|Model) + λcomplexity(Model)

- Our training optimization algorithm is now a function of two terms:
  - **Loss term:** measures how well the model fits the data
  - **Regularization term:** measures model complexity

- λ – Controls strength of regularization

Ridge Regression

- Ordinary Least Squares (OLS) estimates the coefficients $\beta' s$ by minimizing

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- Ridge Regression uses a slightly different equation

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 ,$$

Ridge Regression

- To determine β, we need to minimize L_2

$$Min_{\beta}\, L_2 = (y - x\beta)^2 + \lambda \sum_{i=1}^{p} \beta_i^2$$

$$\begin{aligned} L_2 &= (y - x\beta)^2 + \lambda\beta^2 \\ &= \mathrm{y}^2 - 2\mathrm{xy}\beta + x^2\beta^2 + \lambda\beta^2 \end{aligned}$$

- Calculate the partial derivative and set it to 0:

$$\frac{dL_2}{d\beta} = 0 \text{ at } \beta = \beta^*$$

$$-2\mathrm{xy} + 2x^2\beta^* + 2\lambda\beta^* = 0$$

- Solve for β

$$\beta^* = \frac{xy}{(x^2+\lambda)}$$

Ridge Regression

- The effect of this equation is to add a penalty term

$$\lambda \sum_{j=1}^{p} \beta_j^2,$$

Where the tuning parameter λ is a positive value.

- This has the effect of “shrinking” large values of $\beta' s$ towards zero.

- Such a constraint improves the fit, because shrinking the coefficients can significantly reduce their variance

Regularization Constant

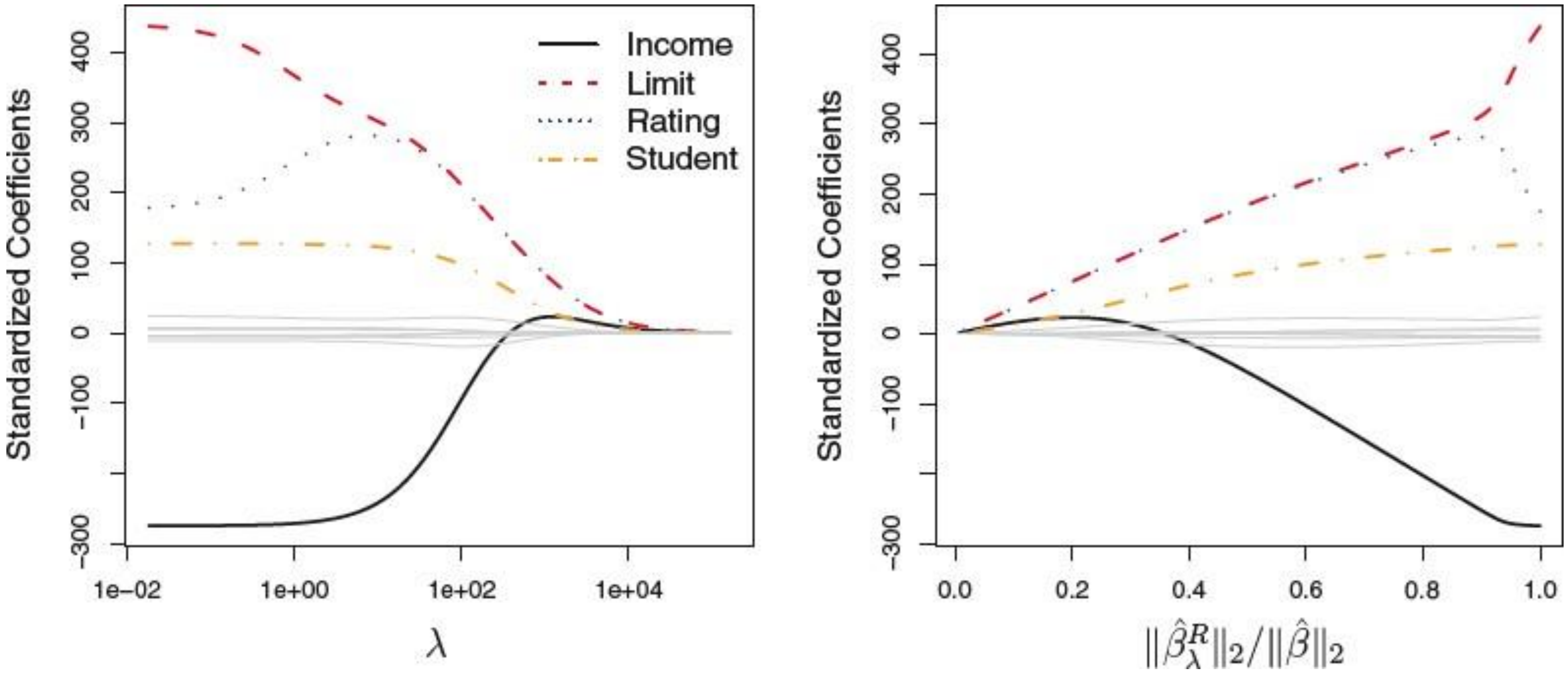
- The tuning parameter λ serves to control the tradeoff between RSS and shrinkage penalty

- When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates

- As $\lambda \rightarrow \infty$, the ridge regression coefficient estimates will approach zero

- Selecting a good value for λ is critical

Credit Data: Ridge Regression

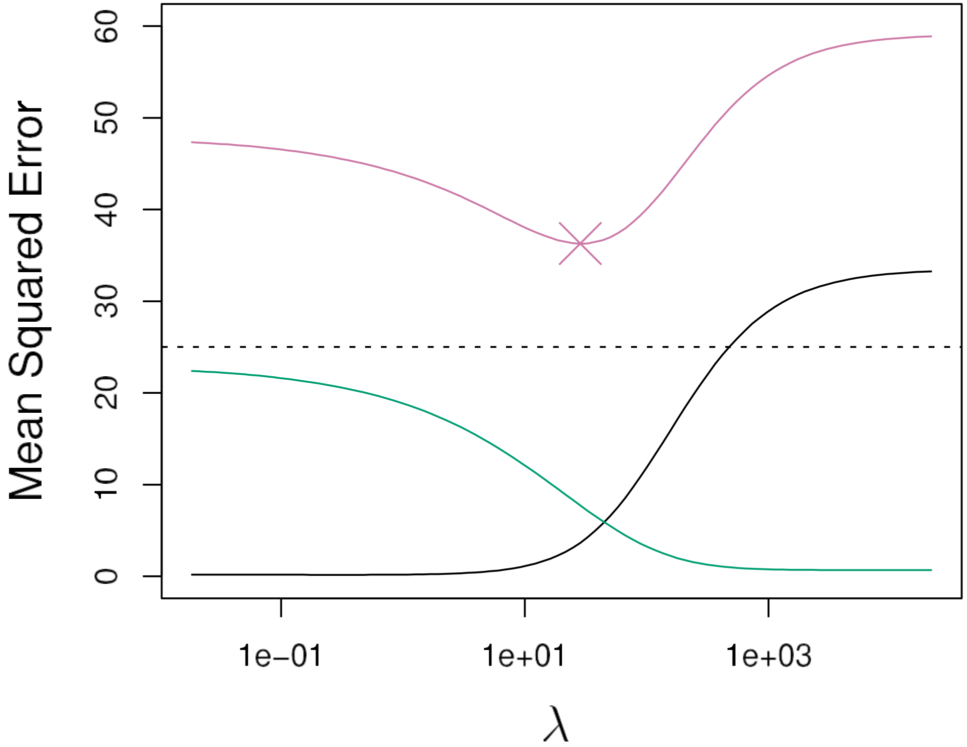


l_2 norm: $\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j{}^2}$

Why Does Ridge Regression Work?

- OLS estimates generally have low bias but can be highly variable
- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance
- Thus, there is a bias/ variance trade-off

Why Does Ridge Regression Work?



Black: Bias
Green: Variance
Purple: MSE

Increase lambda increases bias but decreases variance

Advantages of Ridge Regression

- If p is large, then using the best subset selection approach requires searching through many possible models, which is not the case for ridge regression

- With Ridge Regression, for any given λ, we only need to fit one model

- Ridge Regression can be used even when $p > n$, a situation where OLS do not even have a unique solution

The Lasso

- Ridge Regression has one disadvantage:
  - It will include all p predictors in the final model

- The penalty term will shrink all of the coefficients towards zero but will never force any of them to be exactly zero

- The **Lasso** is an alternative to ridge that overcomes this disadvantage

- The Lasso works in a similar way to Ridge Regression, except it uses a different penalty term

Penalty Term of Lasso

- Ridge Regression minimizes

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\boxed{\lambda\sum_{j=1}^{p}\beta_j^2,}$$

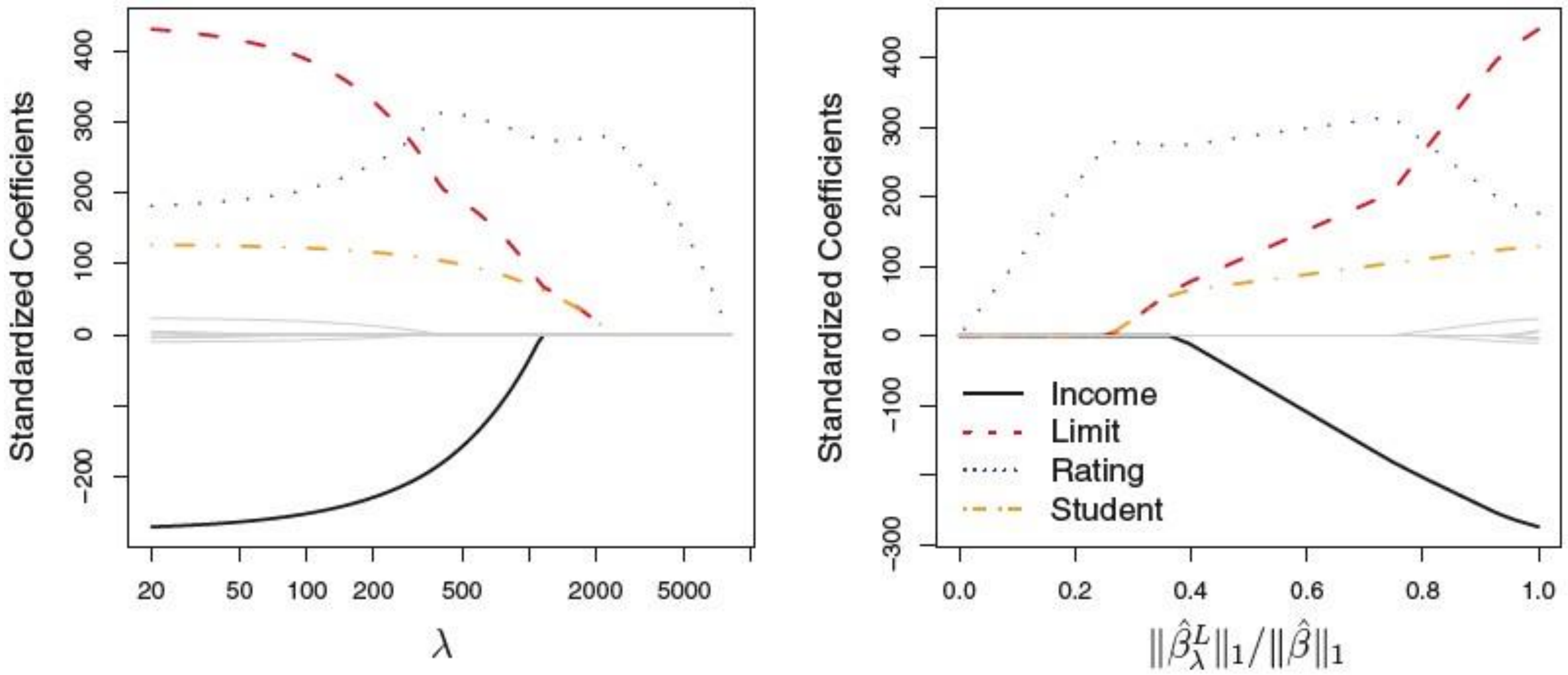
- The LASSO minimizes

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\mathrm{RSS}+\boxed{\lambda\sum_{j=1}^{p}|\beta_j|.}$$

Effect of the Penalty Term

- Lasso shrinks the coefficient estimates towards zero
- The l_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero
- The lasso performs *variable selection*
- Models generated from the lasso are generally much easier to interpret than those produced by ridge regression
- Lasso yields *sparse* models
- Selecting a good value of λ for the lasso is critical

The Lasso


Standardized Coefficients
400
300
200
100
0
−200
20
50
100
200
500
2000
5000
λ
Standardized Coefficients
400
300
200
100
0
−100
−300
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$
Income
Limit
Rating
Student

Another Formulation of Ridge/Lasso

Lasso:

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j|\le s$$

Ridge:

$$\underset{\beta}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \quad \text{subject to} \quad \sum_{j=1}^{p}\beta_j^2\le s$$

Variable Selection Property of Lasso

- Why is it that the lasso, unlike ridge regression, results in coefficient estimates that are exactly equal to zero?

- Difference lies in the nature of the penalty term

Lasso Regression

- To determine β, we need to minimize L_1

$$Min_{\beta}\, L_1 = (y - x\beta)^2 + \lambda \textstyle\sum_{i=1}^{p} |\beta_i|$$

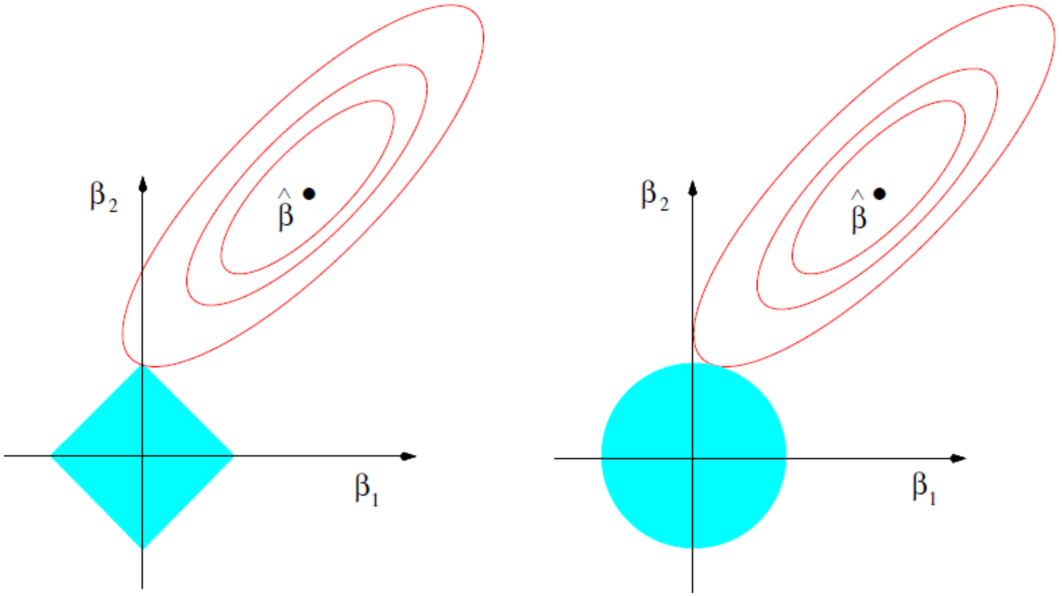
- Let's represent β_i as β

$$\begin{aligned} L_1 &= (y - x\beta)^2 + \lambda|\beta| \\ &= \mathrm{y}^2 - 2\mathrm{xy}\beta + x^2\beta^2 + \lambda|\beta| \end{aligned}$$

- Can you apply partial derivative? The answer No because L_1 is not a continuous function
- Based on optimization theory, the optima occurs at the point of discontinuities which is 0

How Does the Choice of Regularization Affect the Final Solution?

- Plot shows contours of error (red) and constraint (blue) functions for Lasso (left) and Ridge (right) regularization
- Find the first point where the ellipses hit the constraint region
- Unlike the disk, the diamond has corners; if the solution occurs at a corner, then it has one parameter βjequal to zero i.e., Lasso promotes sparsity by setting some weights to zero



Ridge vs. Lasso Regression

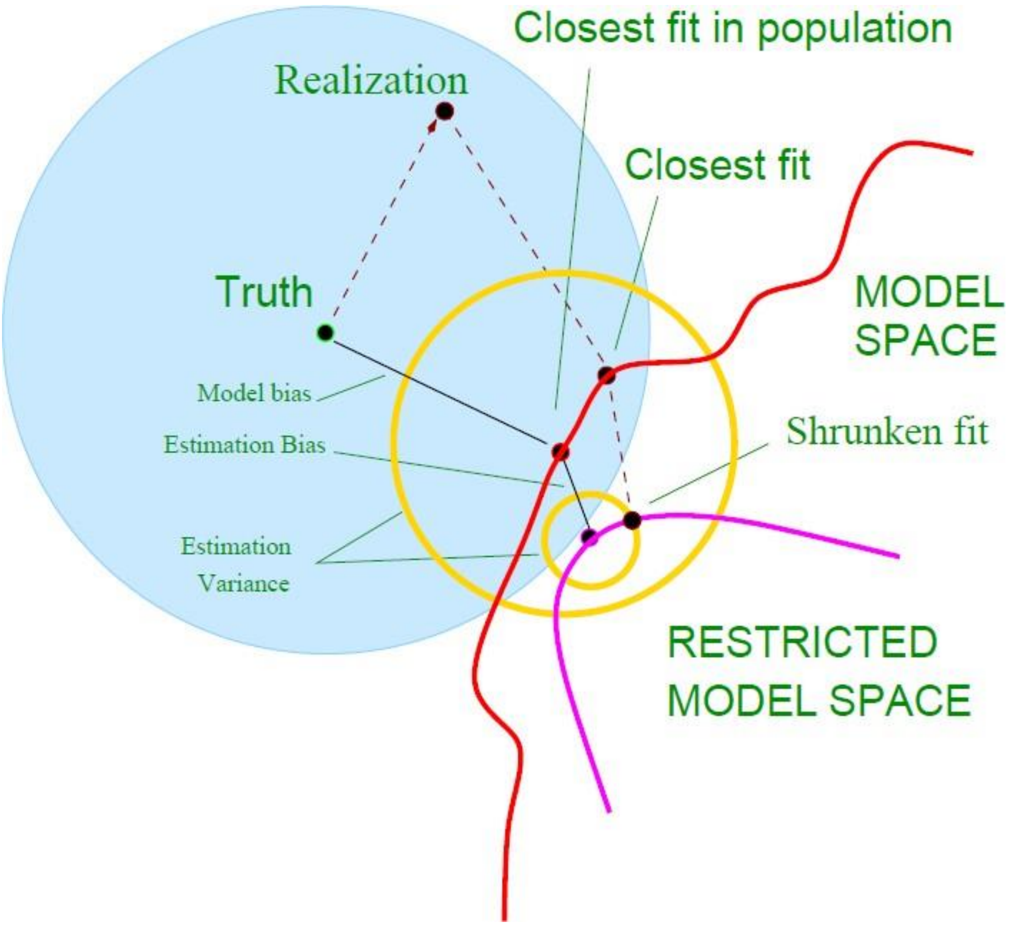
- Neither ridge nor the lasso will universally dominate the other
- In general, Lasso performs better in a setting where a relatively small number of predictors have substantial coefficients
- Ridge regression performs better when the response is a function of many predictors, all with coefficients of roughly equal size
- Cross-validation can be used in order to determine which approach is better on a particular data set

Ridge vs. Lasso Regularization

- The table below captures the differences between the two forms of regularization

Ridge	Lasso
L2 penalizes the sum of squares of weights	L1 penalizes the sum of absolute value of weights
L2 does not have a sparse solution	L1 has a sparse solution
L2 has no feature selection and is not robust to outliers	L1 has in-built feature selection and is robust to outliers
L2 gives better performance when the output is a function of all input variables	L1 models find it hard to learn complex patterns

Putting it All Together

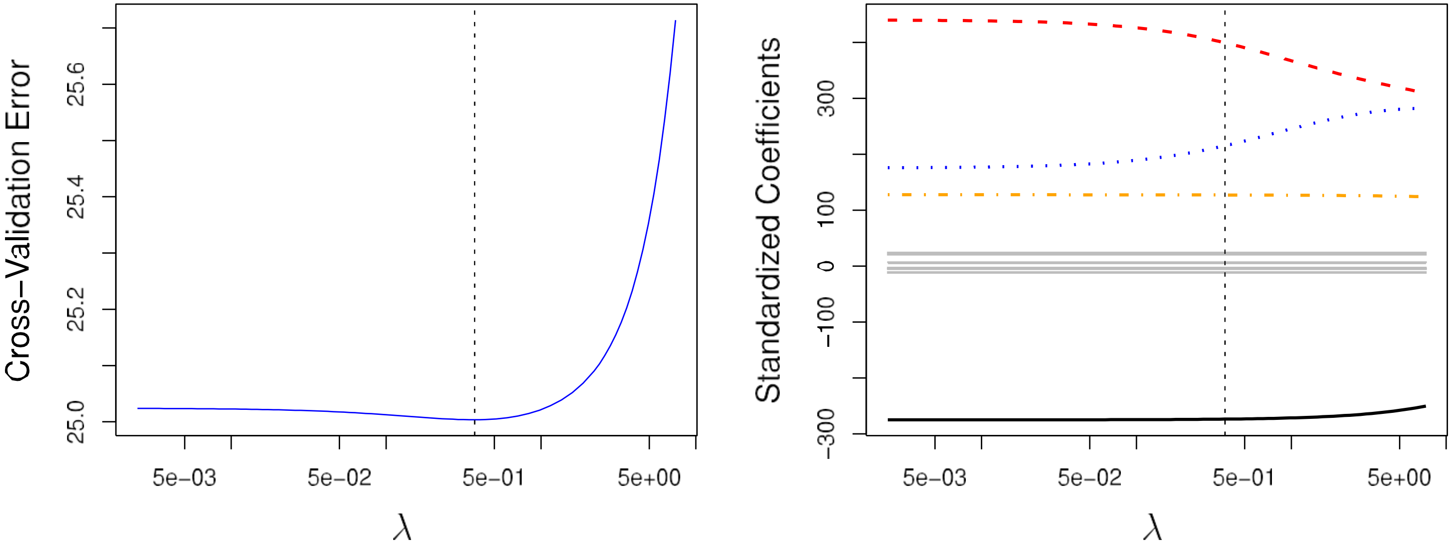


Elements of Statistical Learning – Tibshirani et.al.

Selecting λ

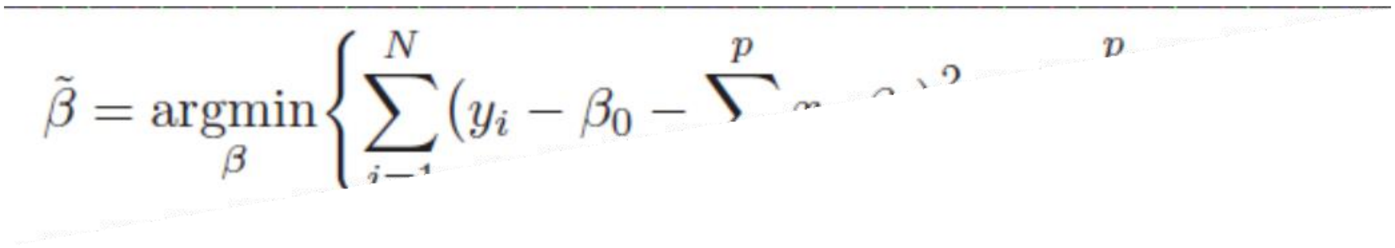
- How to pick a value for λ?

- Select a grid of potential values, use cross validation to estimate the error rate on test data (for each value of λ) and select the value that gives the least error rate

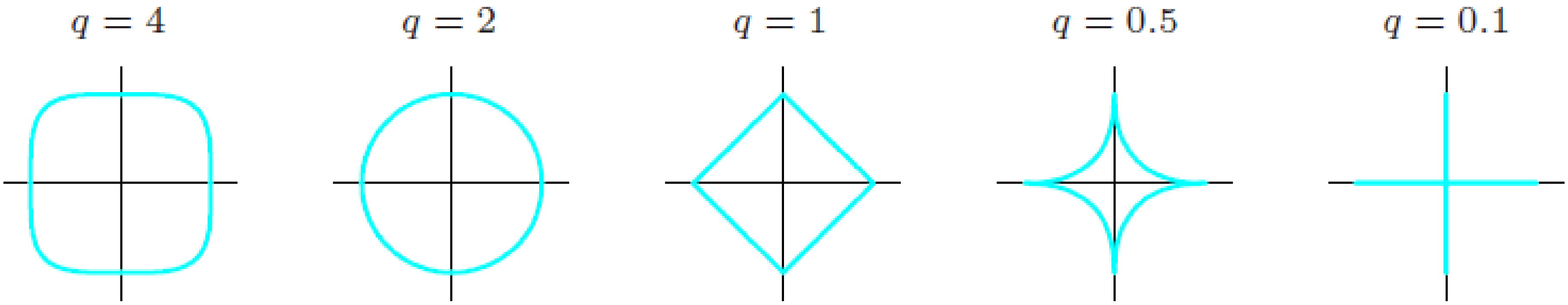


Generalizing Ridge and Lasso Regression

- ➢ Generalize the Ridge and Lasso formulation using L_q penalty



$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum^{p} \right.$$



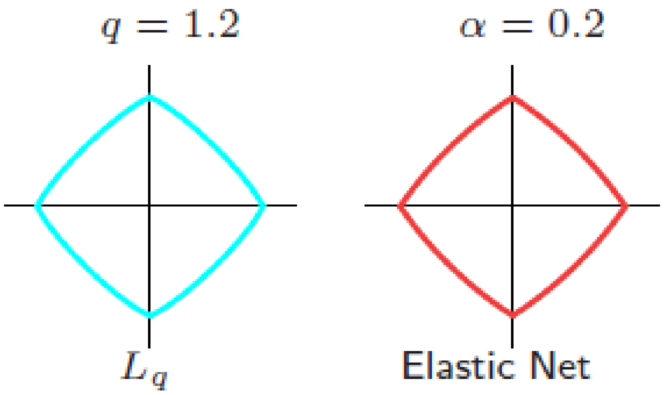
Elastic Net Penalty

- A compromise between Ridge and Lasso

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|),$$

- The elastic-net selects variables like the Lasso, and shrinks together the coefficients of correlated predictors like Ridge

- Considerable computational advantages over the L_q penalties



Resources

- *Chapter 3: Elements of Statistical Learning*

- *Chapter 7: Deep Learning (This is a deeply mathematical treatment of Regularization)*
  http://egrcc.github.io/docs/dl/deeplearningbook-regularization.pdf

Jupyter Notebook

- *Case Study*

ON-BRAND STATEMENT

FOR GENERAL USE

> What defines the students and faculty of the University of Washington? Above all, it’s our belief in possibility and our unshakable optimism. It’s a connection to others, both near and far. It’s a hunger that pushes us to tackle challenges and pursue progress. It’s the conviction that together we can create a world of good. And it’s our determination to Be Boundless. Join the journey at **uw.edu**.