Variable Selection and Regularized Methods [1]

Shaobo Li

University of Kansas

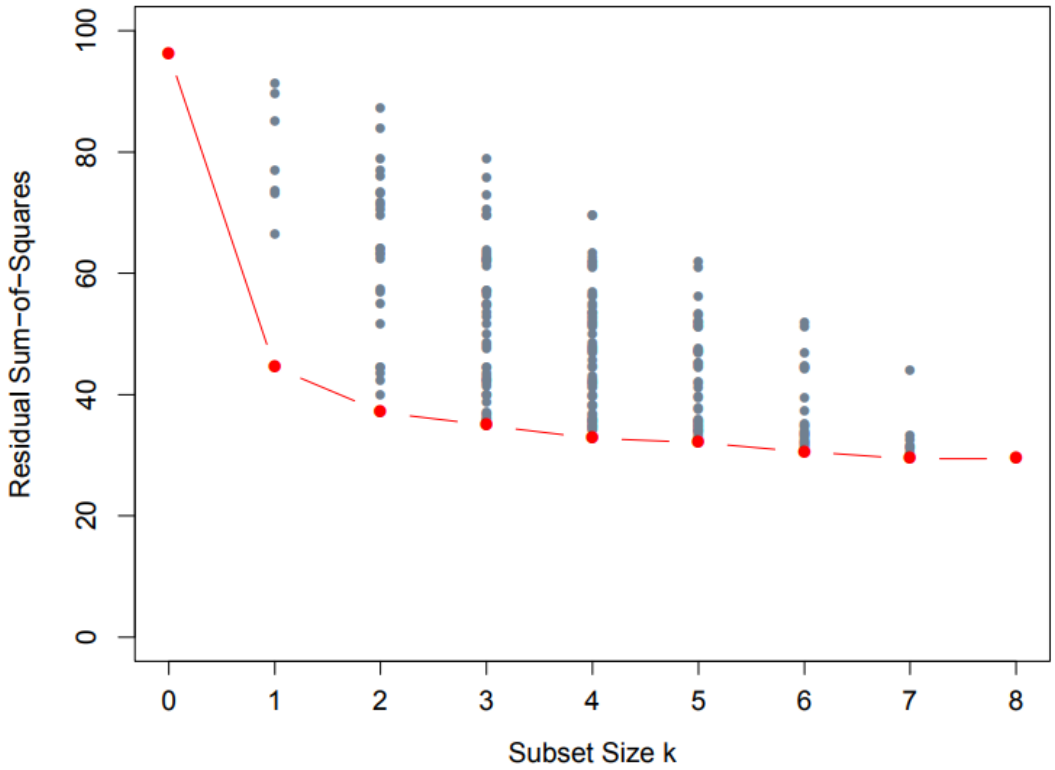
[1] Partially based on Hastie, et al. (2009) ESL, and James, et al. (2013) ISLR

Introduction

- Variable selection – excluding unnecessary variables
  - Interpretation and simplicity
  - Prediction stability and accuracy
  - Bias-variance tradeoff
- Common approaches
  - Stepwise selection (for parametric models)
  - Shrinkage method (also called *regularization*)
  - Dimension reduction (project p predictors to an m-dimensional subspace)
- It can be subjective that some times domain knowledge may force certain variables to be included in the model.

Best Subset Selection

- Search through all subset predictors



- Computationally expensive even infeasible (not commonly used)
  - How many models do we need to evaluate for 10 candidate X's?

Selection Criteria – AIC, BIC

- Akaike information criterion (AIC), the smaller the better

$$AIC = -2\log(\hat{L}) + 2p$$

- Bayesian information criteria (BIC), the smaller the better

$$BIC = -2\log(\hat{L}) + \log(n)p$$

  where $\hat{L}$ is estimated likelihood function
- For linear regression, $-2\log(\hat{L})$ is equivalent to RSS
- What is the difference between AIC and BIC?
  - BIC weighs more on p comparing to AIC. What does it mean?

Forward, Backward, and Stepwise Selection

- Computationally less expensive than best subset
  - practically useful when the number of X is not too large
- Iteratively adding or dropping one variable at a time
- Forward/backward is **greedy** procedure. That is, they won't adjust any added/dropped variables in previous step
- Stepwise: start with forward, and then iteratively add and drop variables
- Commonly used selection criteria: AIC, BIC
- An illustration: click here

Shrinkage Methods for High-Dimensional Regression

- Number of predictor is very large (even larger than sample size)
- Traditional OLS or MLE fails when $p > n$. Why?
- Needs a regularization term (penalty) in the loss.
- This penalty is an additional constraint to the classical OLS or MLE
  - It shrinks the regression coefficients toward 0.
  - Certain penalties can shrink weak coefficients to exact 0.
- Computationally efficient.
- Estimates are usually *biased*.

LASSO

- Least absolute shrinkage and selection operator (LASSO)
- Introduced by Tibshirani (1996)
- One of the most popular variable selection methods
- It estimates the coefficients and selects variables simultaneously.
- A tuning parameter λ controls the "power" of selection.
- Need to standardize all predictors. Why?

LASSO

- LASSO solves the *(L_1) penalized least square*

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \left\{ \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{Least squares}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{L_1 \text{ penalty}} \right\}$$

LASSO

- LASSO solves the *(L_1) penalized least square*

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \Bigg\{ \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{Least squares}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{L_1 \text{ penalty}} \Bigg\}$$

- It is a *convex optimization* problem
- λ is a tuning parameter, called shrinkage parameter

LASSO

- LASSO solves the *(L_1) penalized least square*

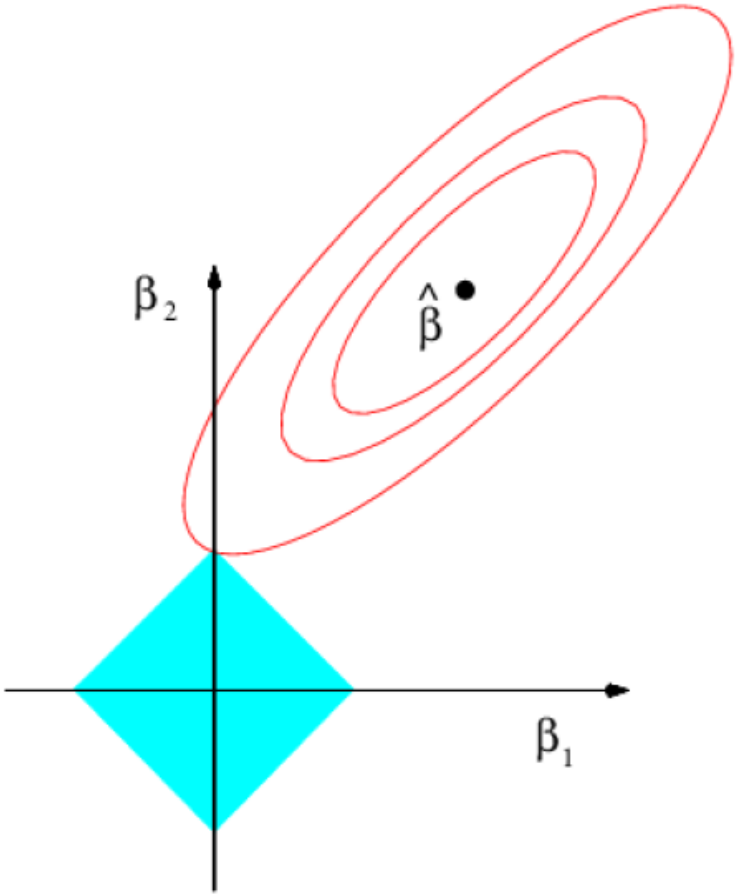
$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \Bigg\{ \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{Least squares}} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{L_1 \text{ penalty}} \Bigg\}$$

- It is a *convex optimization* problem
- λ is a tuning parameter, called shrinkage parameter
- It is equivalent to solve a constrained optimization problem

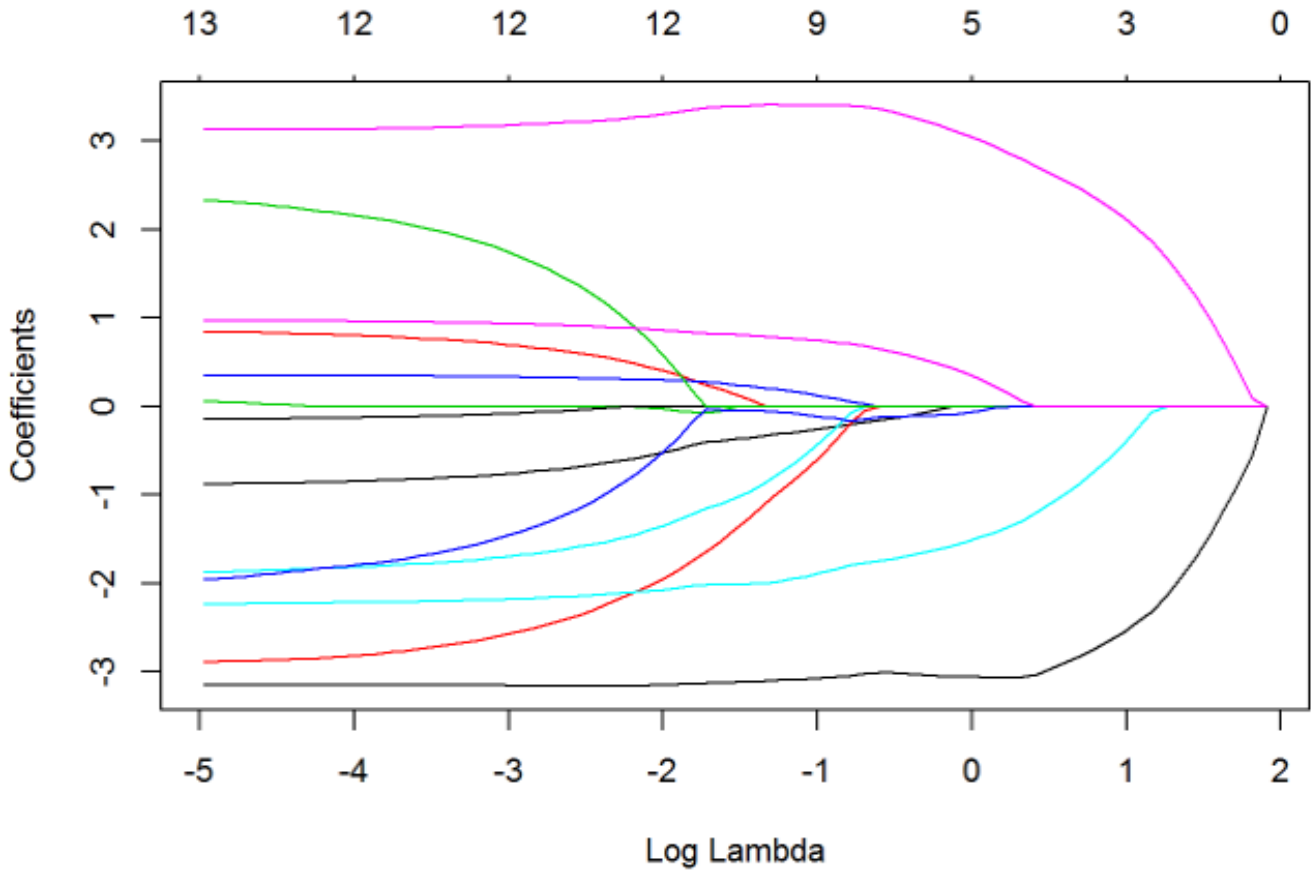
$$\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\text{s.t.} \sum_{j=1}^{p} |\beta_j| = a$$

An Illustration

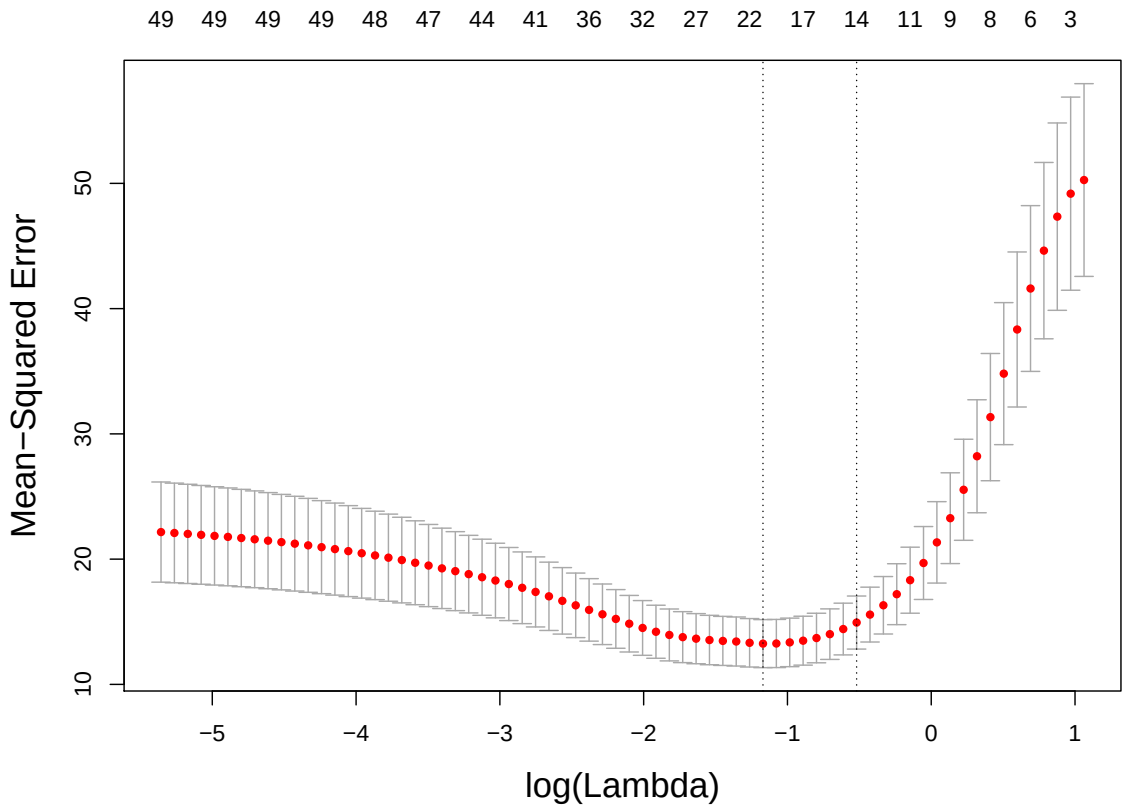


LASSO Regression Solution Path – Boston Housing Data



Tuning Parameter λ Selection via Cross-validation

- λ controls the shrinkage level (different λ associates with different estimated model)



Ridge Regression

- Ridge regression solves a *(L_2) penalized least square*

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \Bigg\{ \underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{\text{Least squares}} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{L_2 \text{ penalty}} \Bigg\}$$

- Writing in matrix form, we can get the analytical solution

$$\hat{\beta}_{Ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{(Challenge yourself: Show it!)}$$

Ridge Regression

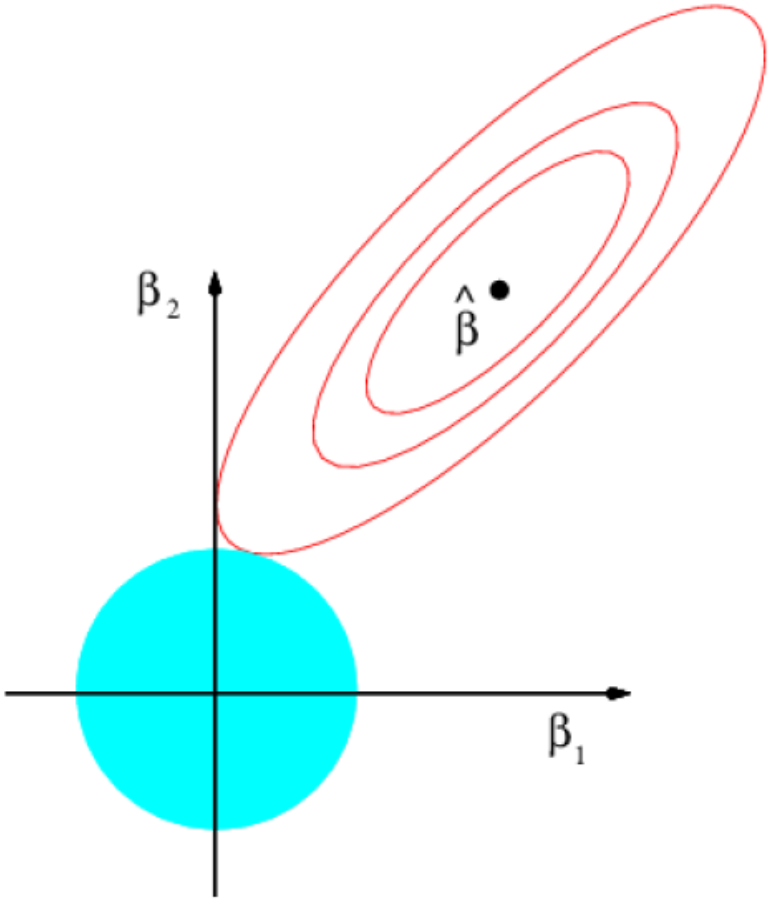
- It is equivalent to solve a constrained optimization problem

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

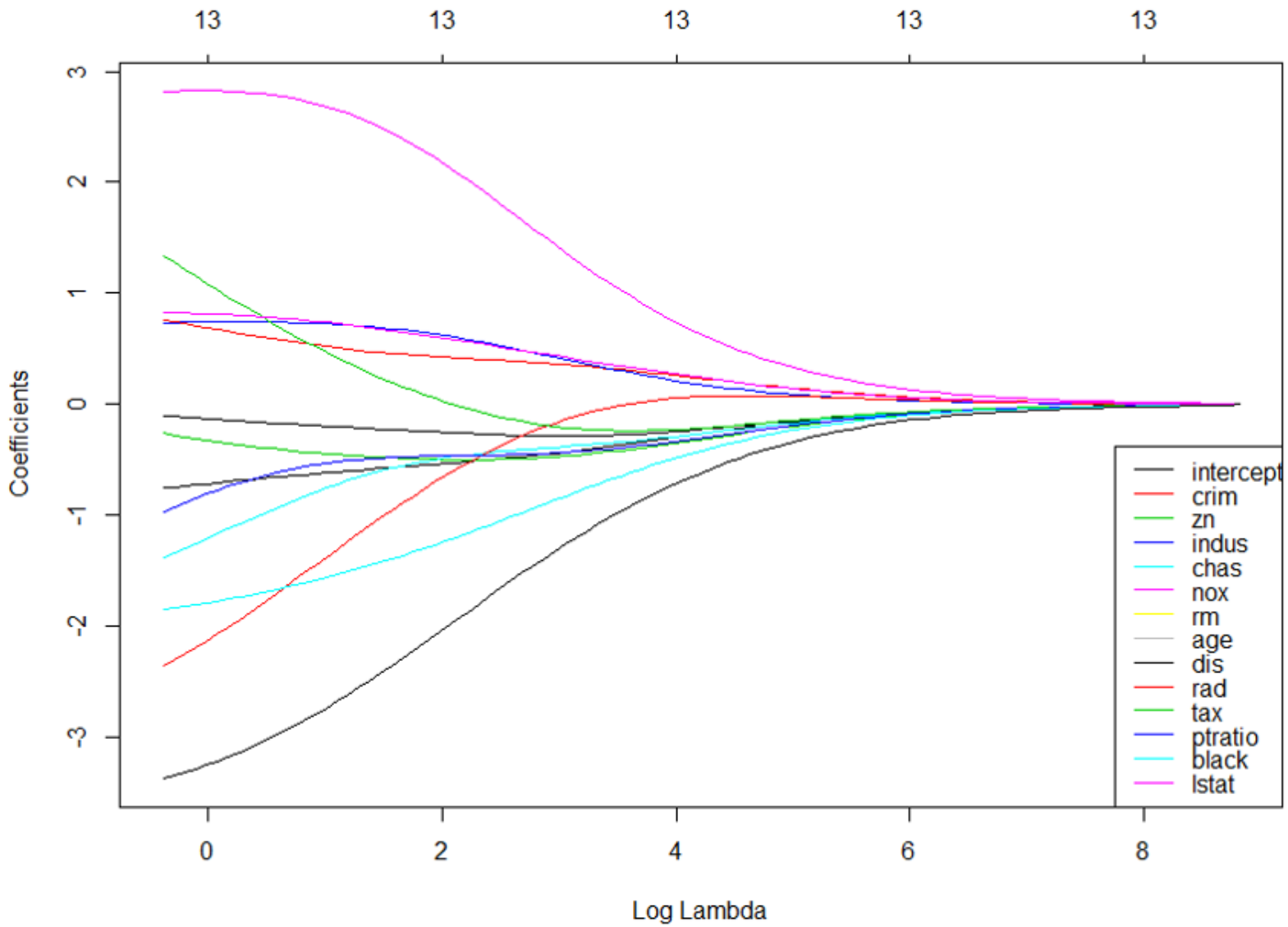
$$\text{s.t.} \sum_{j=1}^{p} \beta_j^2 = a$$

- a corresponds to the tuning parameter λ

An Illustration



Ridge Regression Solution Path – Boston Housing Data



Elastic Net Regression

- Introduced by Zou and Hastie (2005)
- Combination of Ridge and LASSO

$$\hat{\beta}_{EN} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \beta \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

- Convex optimization
- Ridge and LASSO are special cases of Elastic Net
- It incorporates the advantages of both Ridge and LASSO
  - Ridge regression: lower variance; multicollinearity
  - LASSO: variable selection (selects at most n variables if $p > n$)

Some Variants of LASSO

- L_1 penalty:
  - LASSO (Tibshirani, 1996)
  - Adaptive-LASSO (Zou, 2006)
  - SCAD (Fan and Li, 2001)
  - MCP (Zhang, 2010)
- L_2-norm penalty (when X has group structure):
  Group Lasso (Yuan and Lin, 2006)
- $L_1 + L_2$ penalty:
  - Elastic Net (Zou and Hastie, 2005)

Estimation Uncertainty

- Under standard assumptions, OLS provides standard errors and confidence intervals for regression coefficients:

$$\widehat{\mathrm{SD}}(\hat{\beta}_j) = \hat{\sigma}\sqrt{[(X^\top X)^{-1}]_{jj}}, \qquad \hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-p-1}.$$

- Does it apply to penalized regression?
  - Generally, No!
- For penalized regression, what kind of uncertainty do we want to quantify?
  - Standard deviation of $\hat{\beta}$
  - Selection uncertainty (LASSO) – how likely a variable is in?
- What can we do?

Assessing Uncertainty via Bootstrap

- Resampling method
- A powerful tool to quantify uncertainty
  - standard error
  - confidence interval
- Random sampling with replacement
- The general procedure:
  - fit a model **B** times based on **B** bootstrap samples
  - store all the parameter estimates
  - calculate standard error and confidence interval
  - (for selection uncertainty) find the frequency that a variable receives a nonzero coefficient.