DSC5103 Statistics

Session 6. Regularization in Linear Models

Review of last session

- Model selection in (generalized) linear models
  - The model selection workflow: forward, backward, best subset
  - The traditional vs. modern performance measures

- Validation methods: a tool for numerically estimating out-of-sample error
  - Validation set, LOOCV, K-fold CV
  - Auto vs. Manual CV

Plan for today

- Model Selection
  - Best Subset and Stepwise Selection using Cross-Validation

- Shrinkage Methods (Regularization) for linear models
  - Ridge Regression
  - The Lasso
  - Elastic Net

- Regularization in general

Ridge Regression

- Ordinary Least Squares (OLS) minimizes

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

- Ridge Regression imposes a slightly different objective to minimize

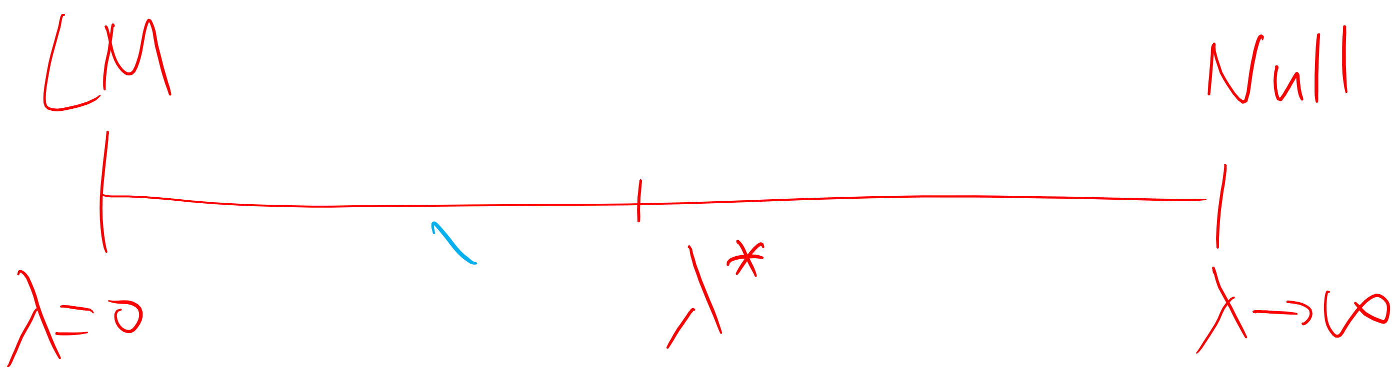
$$\min_{\beta} \qquad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

  - Effectively, Ridge Regression adds a penalty term to linear regression
  - $\lambda \geq 0$ is a tuning parameter

$\sim p$

Ridge Regression

- It still tries to find estimator of β to reduce the RSS
- In addition, it tries to "shrink" large values of β's towards zero $\lambda \sum_{j=1}^{p} \beta_j^2,$
- Parameter λ serves to control the relative weight of the two objectives
  - When λ = 0, it reduces to linear regression (OLS)
  - When λ goes to infinity, it becomes the null model without predictors
  - We shall use cross-validation to find the best λ

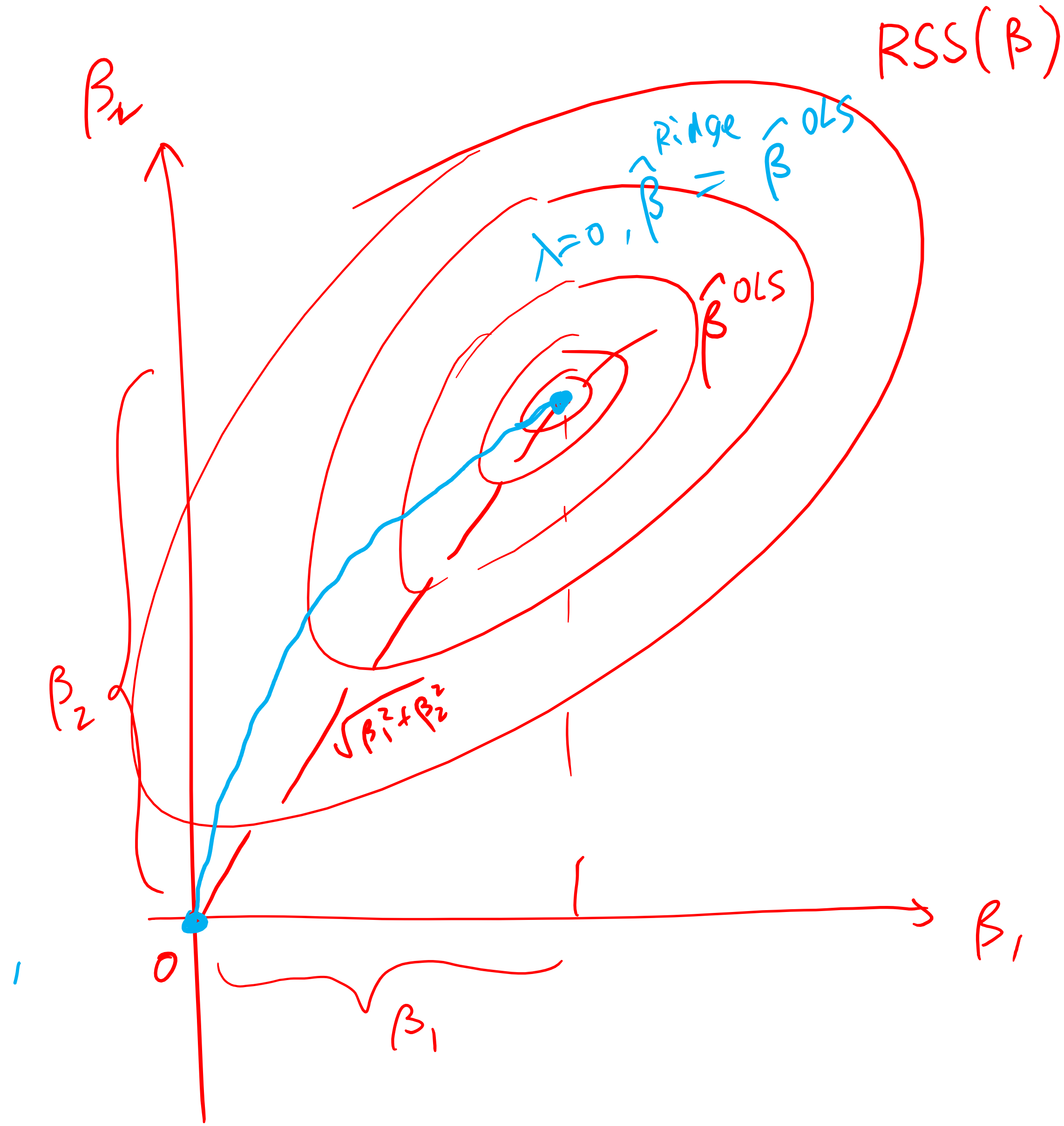


$Y,\ X = (x_1, x_2)$

$\beta = (\beta_0, \beta_1, \beta_2)$

LM: $\min\limits_{\beta} \left\{ \underline{RSS(\beta)} \right.$

Ridge: $+\lambda\left(\underline{\beta_1^2 + \beta_2^2}\right)$



Credit Data: Ridge Regression

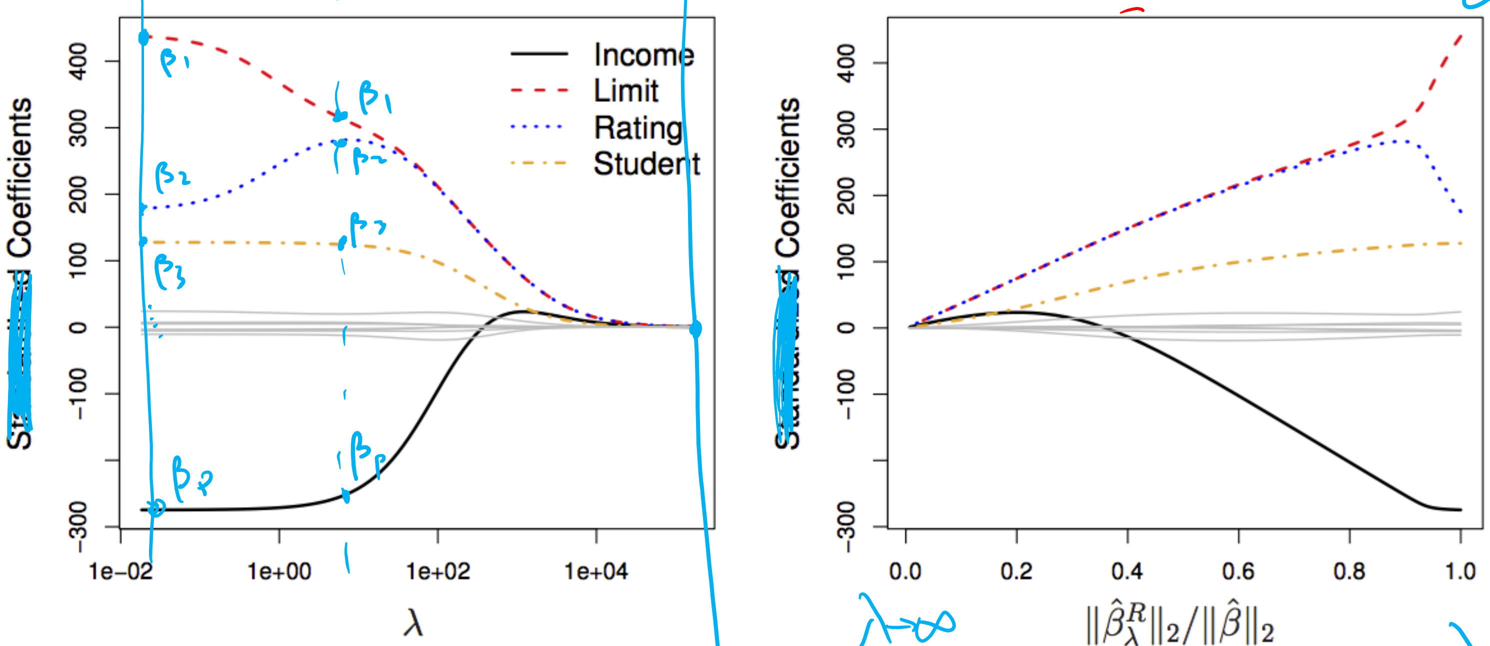
L-2: $\|\beta\|_2 = \sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2}$

Euclidean dist

L-2 dist

- As λ increases, the coefficients shrink towards zero.

LM Null Null LM



$\lambda \to \infty$ $\lambda \to 0$

$\lambda \to 0$ λ $\lambda \to \infty$

Scaling of Predictors

- The standard least squares coefficient estimates are scale equivariant
  - multiplying X_j by a constant **c** simply leads to a scaling of the least squares coefficient estimates by a factor of **c**.
  - regardless of how the **j**-th predictor is scaled, $\beta_j X_j$ will remain the same

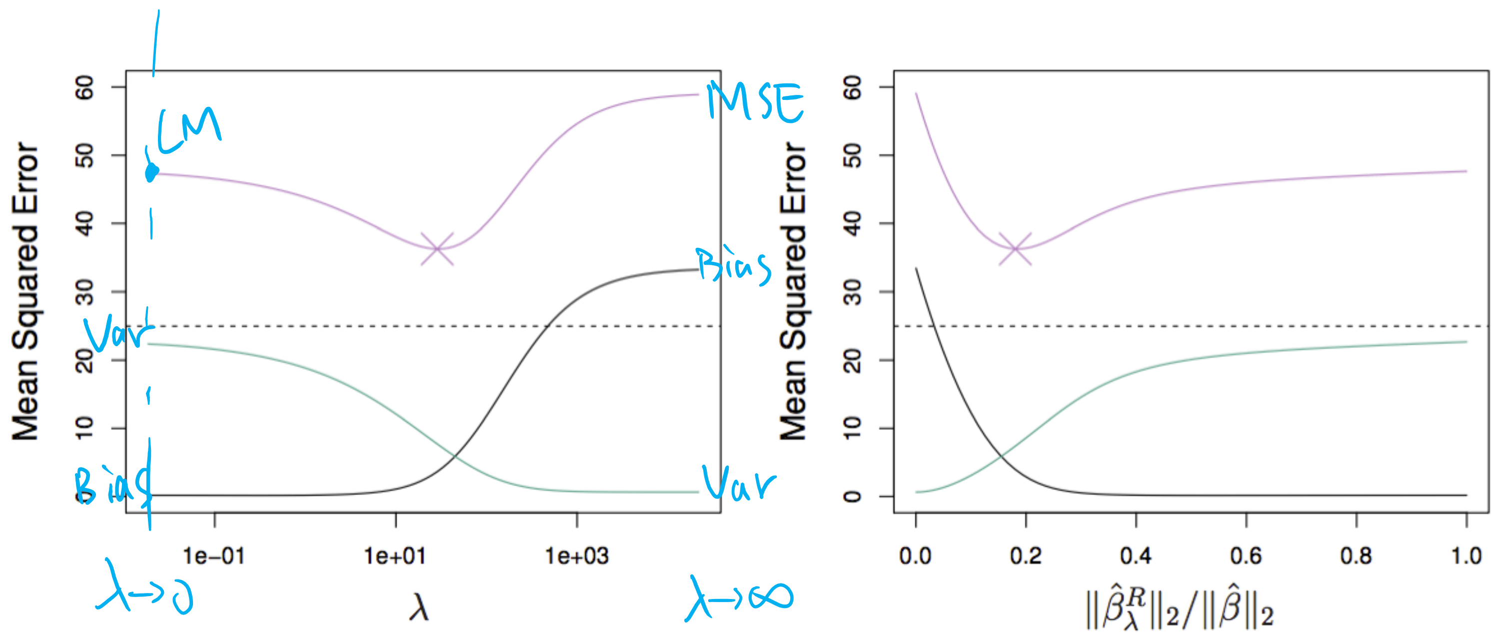
- The ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant due to the penalty term
  - it is best to **standardize** the predictors first by rescaling by their standard deviation

$$y \sim \frac{x_1 - u_1}{\sigma(x_1)} + \cdots + \frac{x_p - u_p}{\sigma(x_p)}$$

Why Shrinkage Works?

- OLS minimizes bias but can be highly variable
  - When there is multicollinearity
  - In particular when n and p are of similar size or when $n < p$
- Ridge regression can substantially reduce **variance** at the cost of **bias**
  - Parameter λ to balance the bias-variance trade-off
  - hence potentially improve the out-of-sample performance

Bias and Variance in Ridge Regression

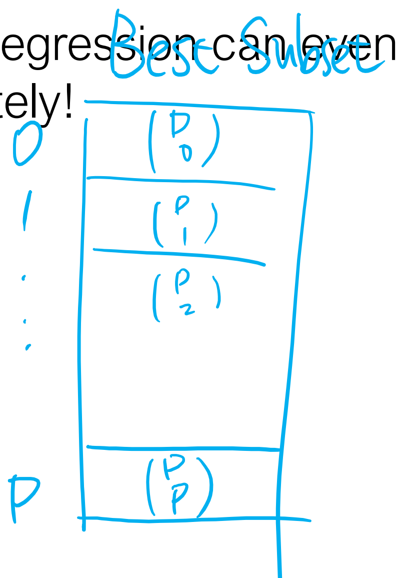


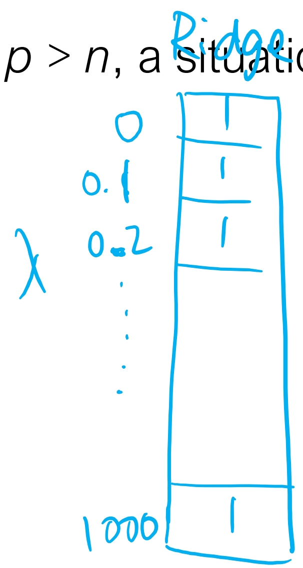
- Black: Bias
- Green: Variance
- Purple: MSE

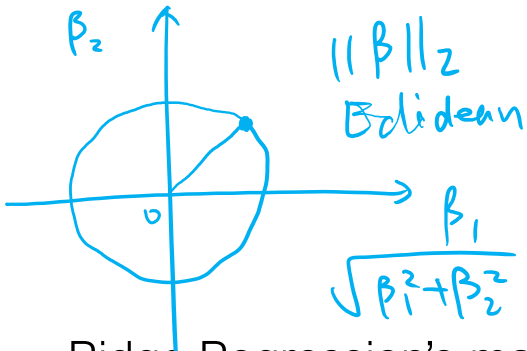
Advantages of Ridge Regression

Computation

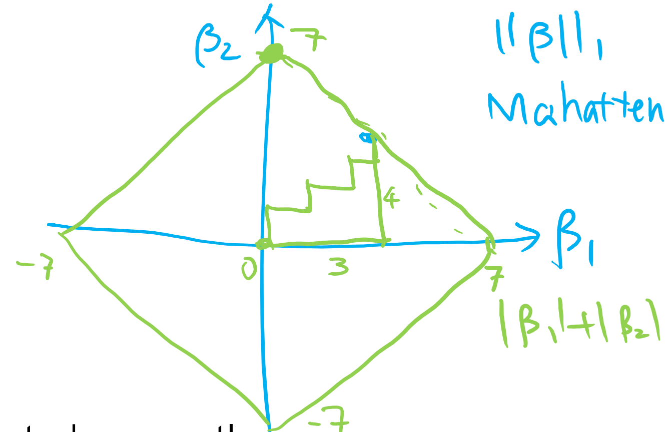
- If p is large, then using the best subset selection approach requires searching through enormous numbers of possible models
- With Ridge Regression, for any given λ, we only need to fit one model and the computations turn out to be very simple
- Ridge Regression can even be used when $p > n$, a situation where OLS fails completely!







The LASSO



- Ridge Regression's major disadvantage
  - the penalty term will never force any of the coefficients to be exactly zero
  - the final model will include all variables, which makes it harder to interpret

- A more modern alternative is the LASSO (Least Absolute Shrinkage and Selection Operator)

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

  - Similar to Ridge Regression, except it uses a different penalty term
  - L-1 versus L-2 norm

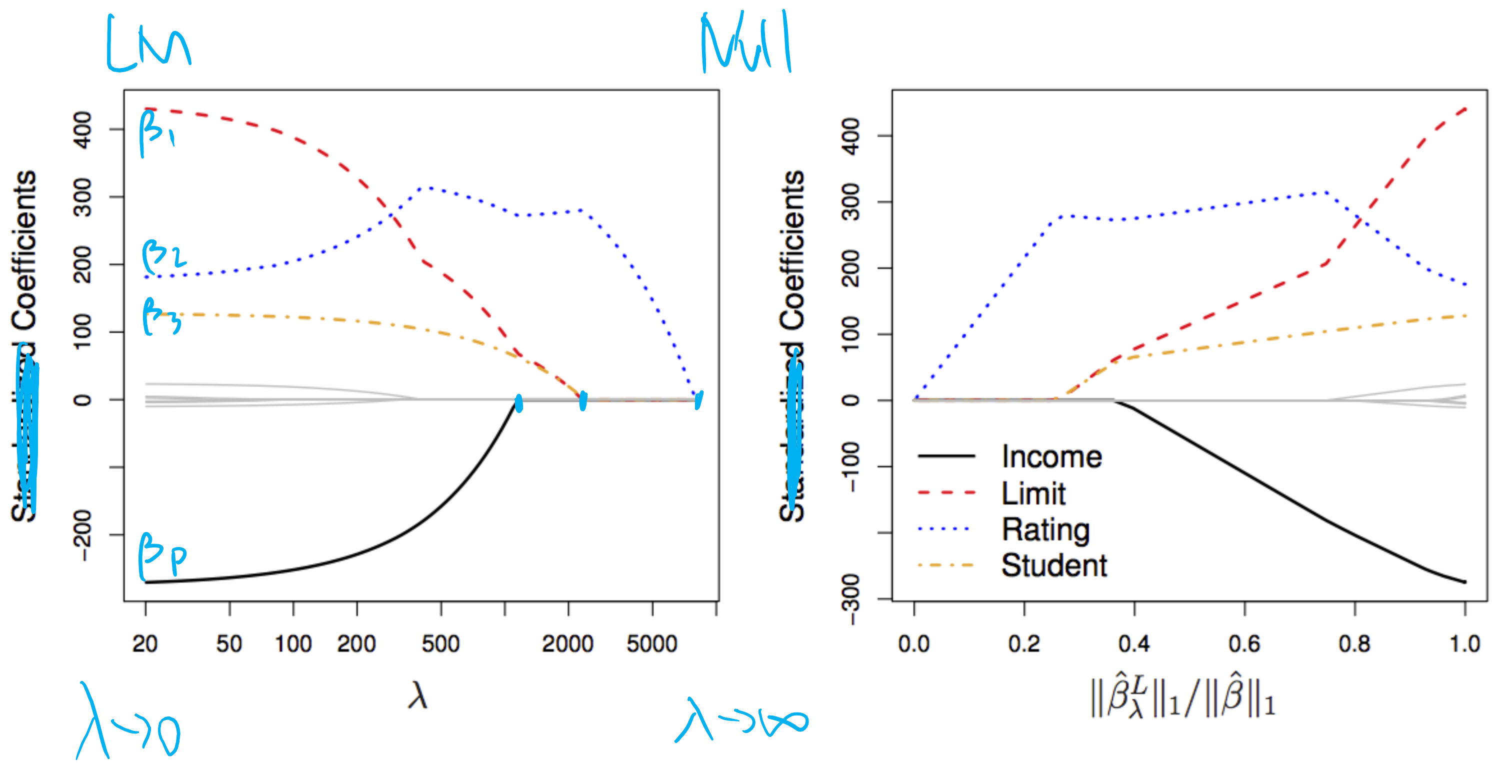
What’s the difference?

- Using this penalty, the LASSO forces coefficient estimates to be exactly zero

- The LASSO effectively does **variable selection** (together with **parameter estimation**) β
  - It yields ***sparse*** models that are easier to interpret

- With LASSO, we can produce a model that has high predictive power and it is simple to interpret

Credit Data: LASSO



Ridge Regression and LASSO

- An optimization perspective
  - View λ as a Lagrangian multiplier

- Ridge Regression

$\min_\beta RSS + \lambda \cdot \|\beta\|_2$ $\quad \|\beta\|_2 < s'$

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

- LASSO

$\min_\beta RSS + \lambda \|\beta\|_1$ $\quad \|\beta\|_1 \leq s$

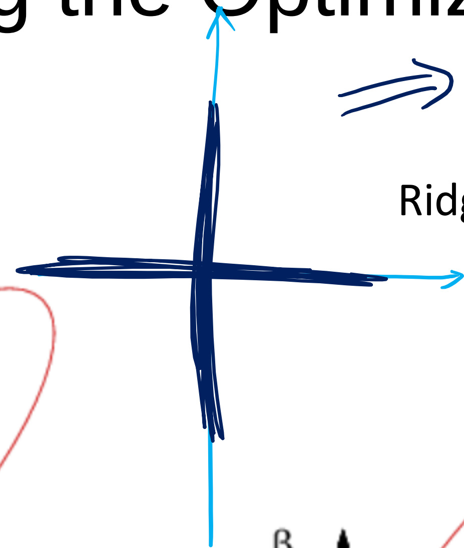
$$\Leftrightarrow \underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

- Best Subset

$\min_\beta RSS \quad \text{subject to} \quad \|\beta\|_0 \leq a$

$\beta = (\beta_1, \beta_2)$

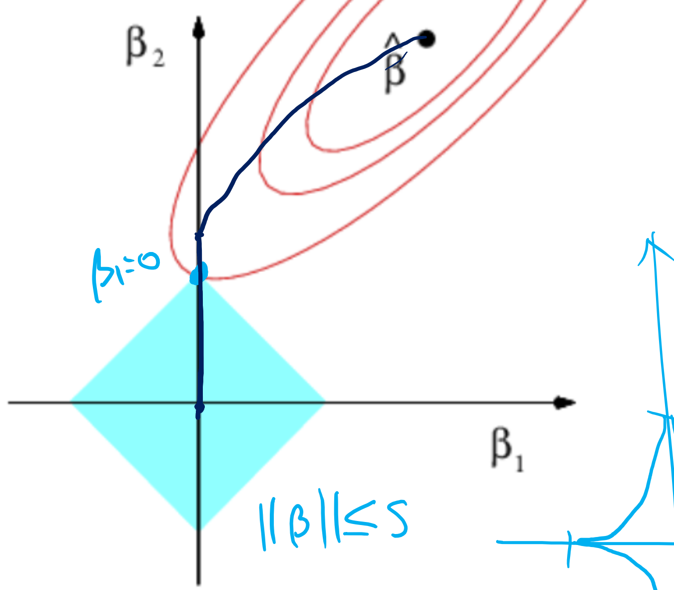
Visualizing the Optimization

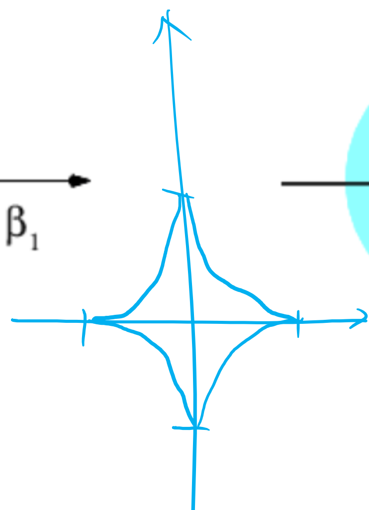


$\Longrightarrow$ Best Subset

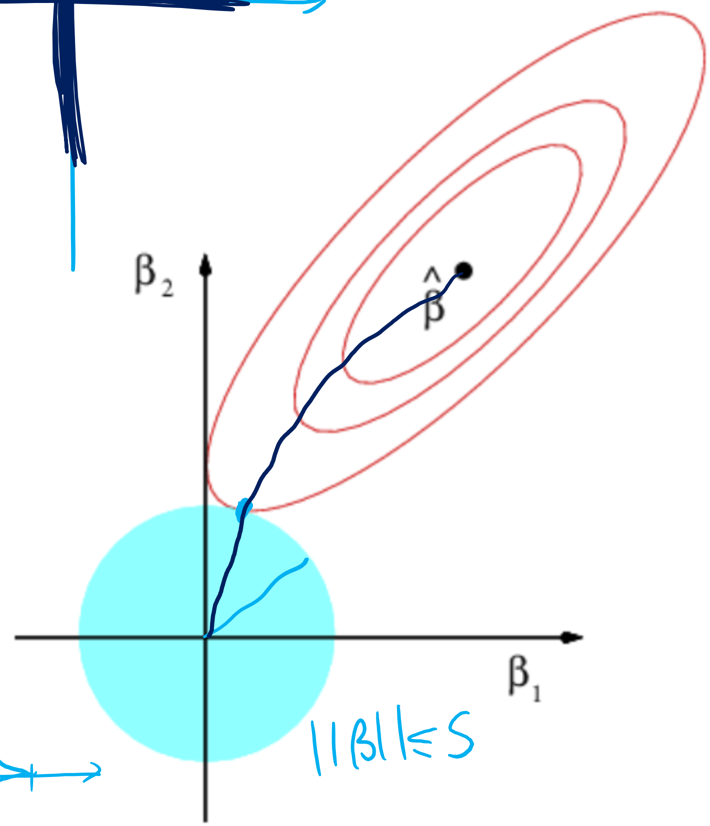
$\|\beta\|_0$: # of nonzero elements in the vector

LASSO



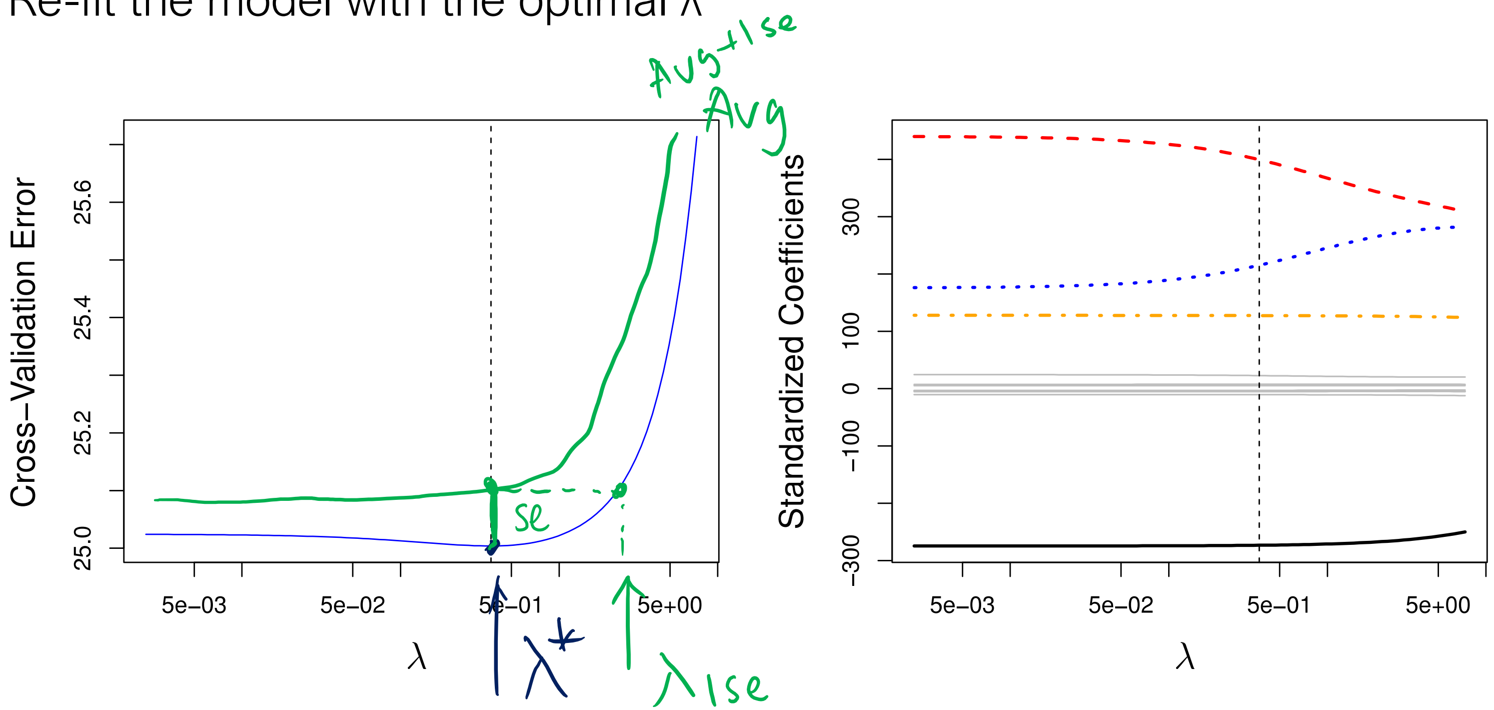


Ridge Regression



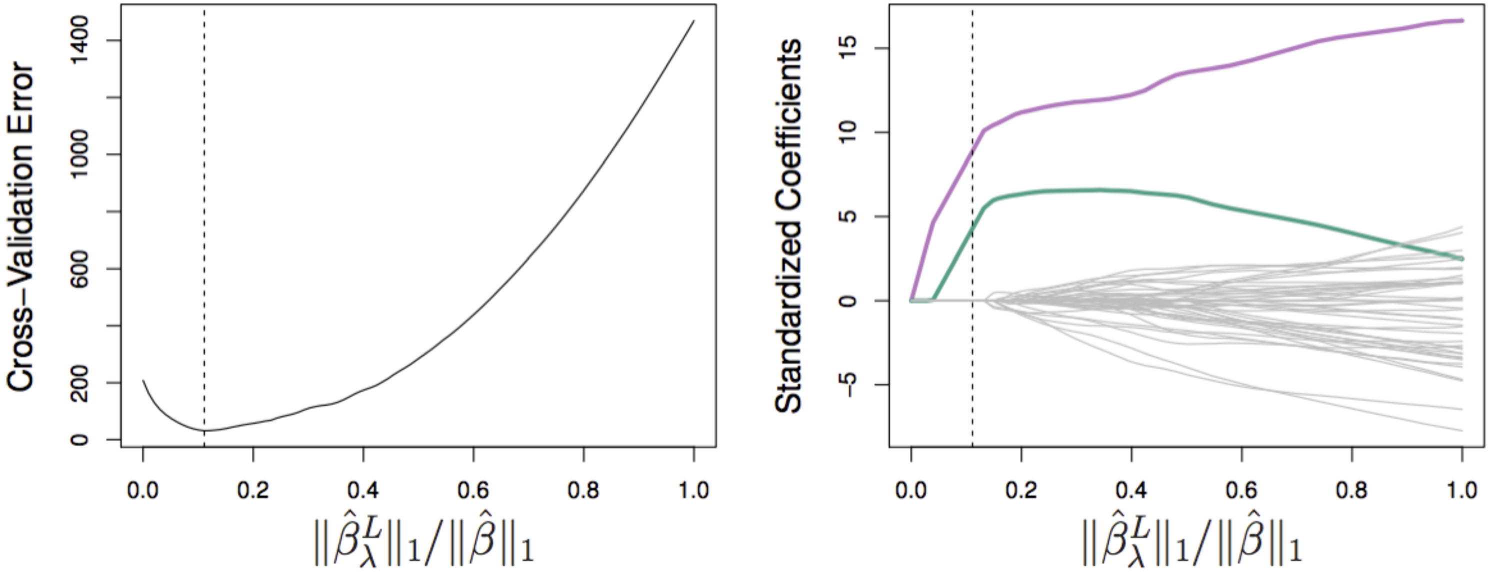
Selecting λ by Cross-Validation

- Select a grid of potential values, compute cross-validation error rate (for each value of λ), and select the one that gives the least error rate
- Re-fit the model with the optimal λ



Selecting λ by Cross-Validation

- LASSO



$y = X$

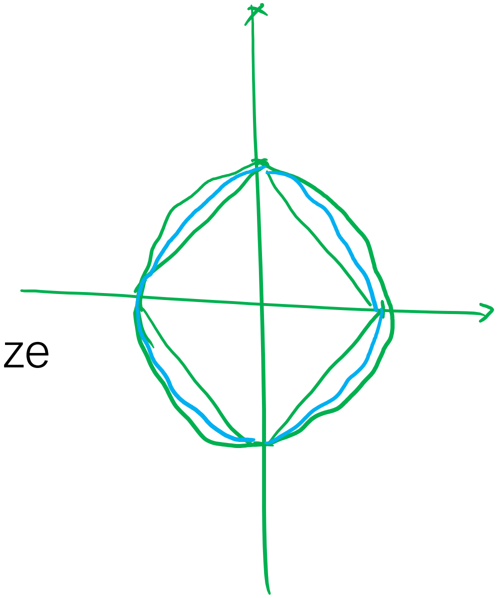
$y \sim X + x'$ (β_1, β_2)

$\lambda(|\beta_1| + |\beta_2|)$

$\lambda(\beta_1^2 + \beta_2^2)$

$$y = \begin{cases} 1 \cdot X + 0X' \\ 2X - X' \\ \vdots \\ 0 \cdot X + 1X' \end{cases} \Rightarrow \text{Ridge: } 0.5X + 0.5X'$$

Elastic Net



- A combination of Ridge Regression and LASSO: to minimize

$$\mathrm{RSS} + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

  - Two tuning parameters λ_1 and λ_2

- In R implementation (function *glmnet()*) of elastic net

$$\mathrm{RSS} + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^{p} \beta_j^2 \right)$$

  - Two tuning parameters λ and α ($0 \le \alpha \le 1$)
  - Special cases: $\alpha = 0$ is ridge regression; $\alpha = 1$ is lasso

$\alpha = 0$: Ridge

$0 < \alpha < 1$: EN

$\alpha = 1$: LASSO

Regularization in General

- Simultaneous parameter estimation and variable selection

- The general idea of regularization applies to a much wider class of tools
  - Generalized Linear models
  - Tree pruning
  - SVM
  - Neural Network and Deep Learning
  - …

- Allow for much more complicated models without overfitting
- Appropriate for p >> n problems

Group Project

- Requirement/assessment
  - Problem definition (5): research questions, data
  - Analysis execution (5): choice of tools, model generation and comparison
  - Report (5) and presentation (5)
- Report (Technical Appendix)
  - As concise as possible (penalty term for number of pages)
  - Rmarkdown is good enough
- Presentation
  - Around 15-minute self-recorded video
  - To cover the high-level messages not technical details
  - Better to involve all the team members
- Submission deadline: Nov 23