

INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo

Dept. of Bio-industrial Mechatronics Engineering

National Taiwan University

Today:

- Constraint least squares regression methods
- Bootstrapping

Outline

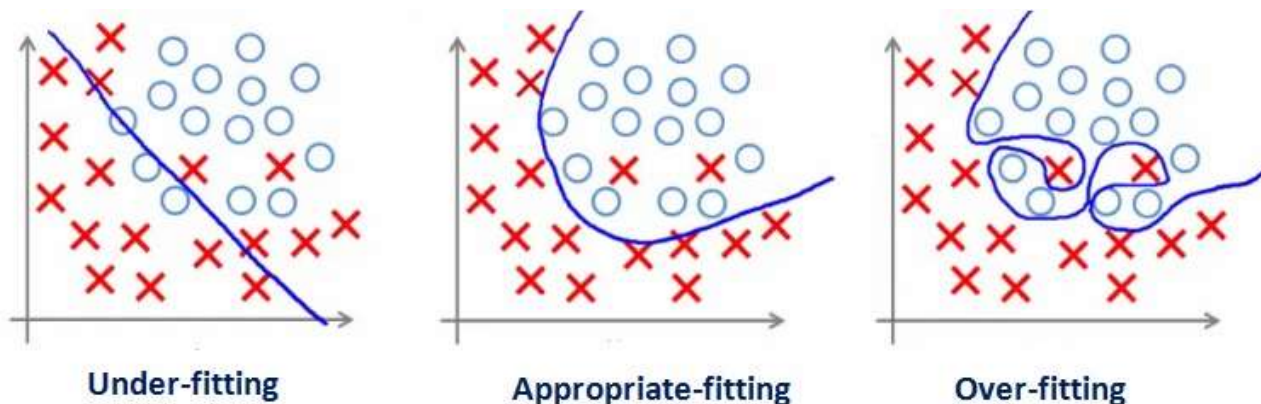
- Goal of the lecture
- Constraint least-squares
- Bootstrapping

Goals

- After this, you should be able to:
 - Understand the risk of overfitting
 - Perform constraint least squares regression methods

Review of Overfitting

- A phenomena that a model is excessively complex
- A overfitted model describes random error or noise instead of the underlying relationship



Tackle Overfitting – Shrinkage Methods

- Shrinkage is a regression method that involves the use of penalties or constraints to shrink the coefficients of model parameters
- The constraint makes the coefficients of the variables smaller, hence “shrinkage”
- Also called regularization
- Typical shrinkage methods:
 - Ridge regression
 - LASSO

Review – Linear Regression

- A linear multiple regression model that predicts a response variable from an explanatory vector has the form: $\hat{y} = \mathbf{x}^T \boldsymbol{\beta}$
- Given a response vector \mathbf{y} and explanatory matrix \mathbf{X} from N observations, the model coefficients $\boldsymbol{\beta}$ can be obtained in least-squares sense, i.e.,

$$\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 ,$$

$$\text{where } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NM} \end{bmatrix}$$

Shrinkage Method – Ridge Regression

- Constraint least-squares methods impose penalty on the “size” of $\boldsymbol{\beta}$
- The “size” is measured in a few different ways
- If the “size” is measured in L_2 sense, the optimization problem can be written as:

$$\boldsymbol{\beta}_{Ridge} = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|^2 < c \in \Re$$

- This is equivalent to

$$\boldsymbol{\beta}_{Ridge} = \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2), \quad (\text{why?})$$

where $0 < \lambda \in \Re$ is the regularization parameter

- This is called **ridge regression** (1970) (why?)

Solving the Ridge Regression Problem

- Assume the data is normalized, the gradient of $\boldsymbol{\beta}_{Ridge}$ is:

$$\begin{aligned}\nabla_{\boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} \left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \right) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}) \\ &= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} + 2\lambda \boldsymbol{\beta} = 2(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\boldsymbol{\beta} - 2\mathbf{X}^T \mathbf{y}\end{aligned}$$

$$\nabla_{\boldsymbol{\beta}} = 0 \quad \Rightarrow \quad \boldsymbol{\beta}_{Ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Recall: SVD of A Matrix

- It is always possible to decompose any matrix \mathbf{X} into

$$\mathbf{X} = \mathbf{E}\mathbf{\Sigma}\mathbf{F}^T = \sum_{i=1}^M \sigma_i \mathbf{e}_i \mathbf{f}_i^T$$

Note that the variance of the PC is represented by σ_i^2

$$\begin{array}{ccccccc}
 \left[\begin{array}{c} \mathbf{X} \\ \hline \end{array} \right] & = & \left[\begin{array}{c|c|c} & & \\ \mathbf{e}_1 & \cdots & \mathbf{e}_M \\ & & \end{array} \right] & \left[\begin{array}{cc} \sigma_1 & \emptyset \\ & \ddots \\ \emptyset & \sigma_M \\ \hline \emptyset & \emptyset \end{array} \right] & \left[\begin{array}{c} \text{---} \mathbf{f}_1 \text{---} \\ \text{---} \mathbf{f}_N \text{---} \end{array} \right] \\
 M \times N & & M \times M & M \times N & N \times N
 \end{array}$$

where \mathbf{E} and \mathbf{F} are orthogonal matrices, i.e., $\mathbf{E}^T \mathbf{E} = \mathbf{I}$ and $\mathbf{F}^T \mathbf{F} = \mathbf{I}$

Algebraic Interpretation of Ridge Regression

- Consider the fitted response:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}_{Ridge} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- Introduce the SVD of \mathbf{X}

$$\mathbf{X} = \mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T$$

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T \left((\mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T)^T (\mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T) + \lambda\mathbf{I} \right)^{-1} (\mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T)^T \mathbf{y} \\ &= \mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T (\mathbf{F}\boldsymbol{\Sigma}^T\mathbf{E}^T\mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T + \lambda\mathbf{I})^{-1} \mathbf{F}\boldsymbol{\Sigma}^T\mathbf{E}^T \mathbf{y} \\ &= \mathbf{E}\boldsymbol{\Sigma}\mathbf{F}^T (\mathbf{F}\boldsymbol{\Sigma}^2\mathbf{F}^T + \mathbf{F}(\lambda\mathbf{I})\mathbf{F}^T)^{-1} \mathbf{F}\boldsymbol{\Sigma}^T\mathbf{E}^T \mathbf{y}\end{aligned}$$

Algebraic Interpretation of Ridge Regression (Cont'd)

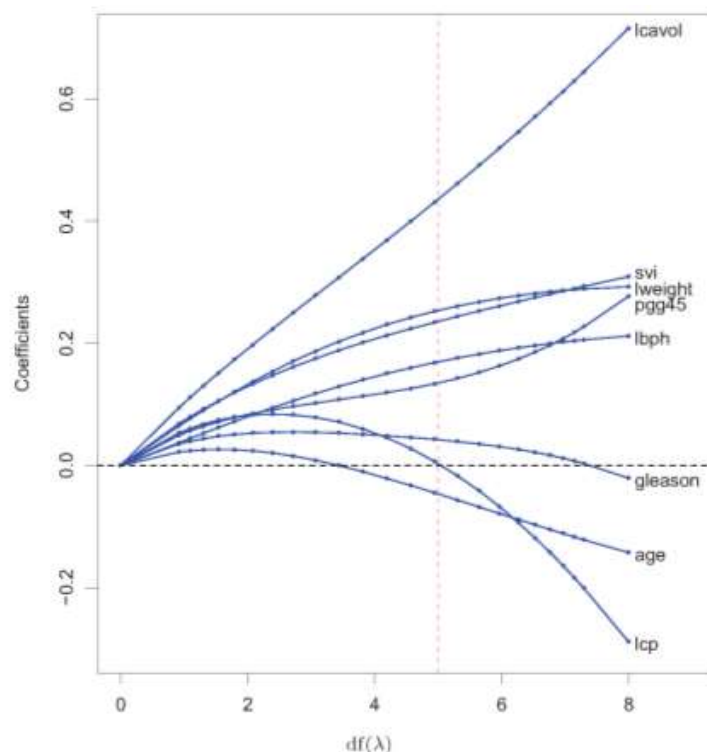
$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{E}\mathbf{\Sigma}\mathbf{F}^T(\mathbf{F}(\mathbf{\Sigma}^2 + \lambda\mathbf{I})\mathbf{F}^T)^{-1}\mathbf{F}\mathbf{\Sigma}^T\mathbf{E}^T\mathbf{y} \\ &= \mathbf{E}\mathbf{\Sigma}\mathbf{F}^T(\mathbf{F}^T)^{-1}(\mathbf{\Sigma}^2 + \lambda\mathbf{I})^{-1}(\mathbf{F})^{-1}\mathbf{F}\mathbf{\Sigma}^T\mathbf{E}^T\mathbf{y} \\ &= \mathbf{E}\mathbf{\Sigma}(\mathbf{\Sigma}^2 + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}^T\mathbf{E}^T\mathbf{y} \\ &= \sum_{i=1}^M \mathbf{e}_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \mathbf{e}_i^T \mathbf{y}, \quad \text{where} \quad \sum_{i=1}^M \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \triangleq df(\lambda) \in \mathfrak{R}\end{aligned}$$

- Ridge regression shrinks the coefficients with respect to the orthonormal basis formed by the principal components

Ridge Regression Coefficient against λ

- Coefficients with respect to the principal components with smaller variances are shrunk more
- The coefficients are NOT shrunk to zero until the $df(\lambda)$ is zero, i.e., $\lambda = \infty$

Ridge Regression



Hastie, Tibshirani, and Friedman.
The Elements of Statistical Learning

$$\sum_{i=1}^M \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \triangleq df(\lambda)$$

Shrinkage Method – LASSO

- If the size of β is measured in L_1 sense, the optimization problem can be written as:

$$\beta = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 \quad \text{subject to} \quad |\beta| < c \in \Re$$

- This is equivalent to

$$\beta = \min_{\beta} (\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda |\beta|)$$

where $0 < \lambda \in \Re$ is the regularization parameter

- This is called **LASSO** (least absolute shrinkage and selection operator) (1996) (**why?**)
- NO explicit solution!



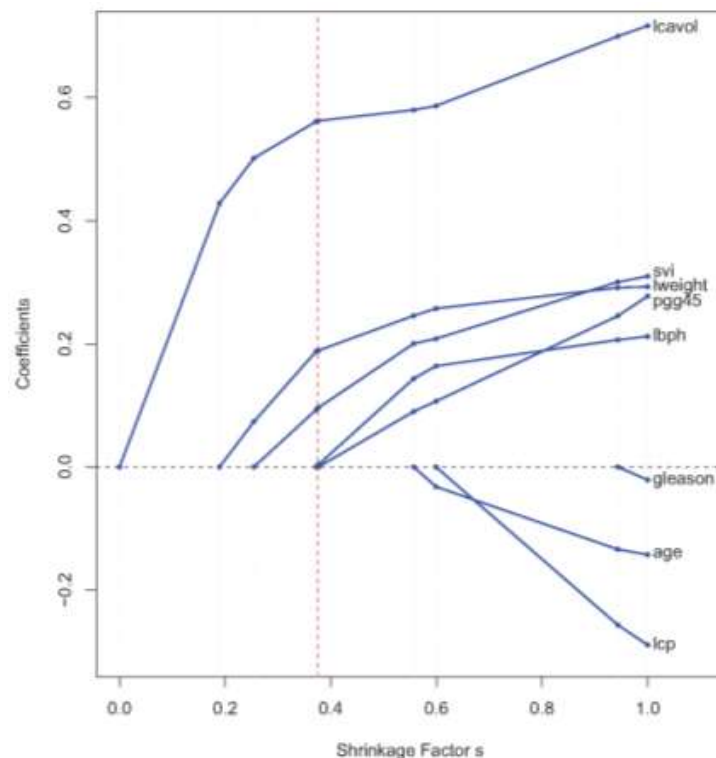
LASSO Coefficient against Factor s

- The shrinkage factor s is defined as

$$s = \frac{c}{\sum_{i=1}^N |\beta_{LS}|} \in \mathbb{R}$$

- If $c = \sum_{i=1}^N \beta_{LS}$, there is no shrinkage at all
- If s is small enough, some coefficients are 0 and the LASSO acts as subset selection method

LASSO

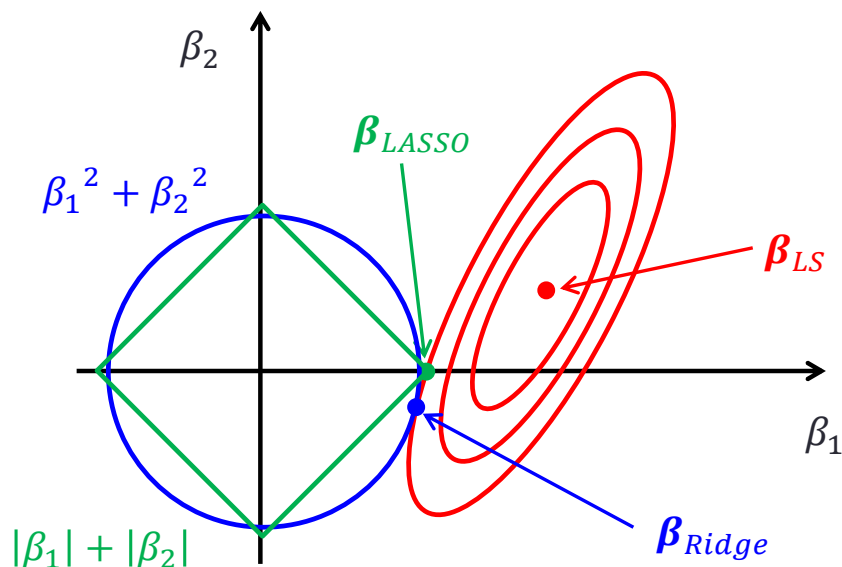


Hastie, Tibshirani, and Friedman.
The Elements of Statistical Learning

$$s = \frac{c}{\sum_{i=1}^N |\beta_{LS}|}$$

Geometry Interpretation of Ridge & LASSO

- **Cost function:** $\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$
- **Ridge constraint:** $\|\boldsymbol{\beta}\|^2 < c \Rightarrow \beta_1^2 + \beta_2^2 < c$
- **LASSO constraint:** $|\boldsymbol{\beta}| < c \Rightarrow |\beta_1| + |\beta_2| < c$



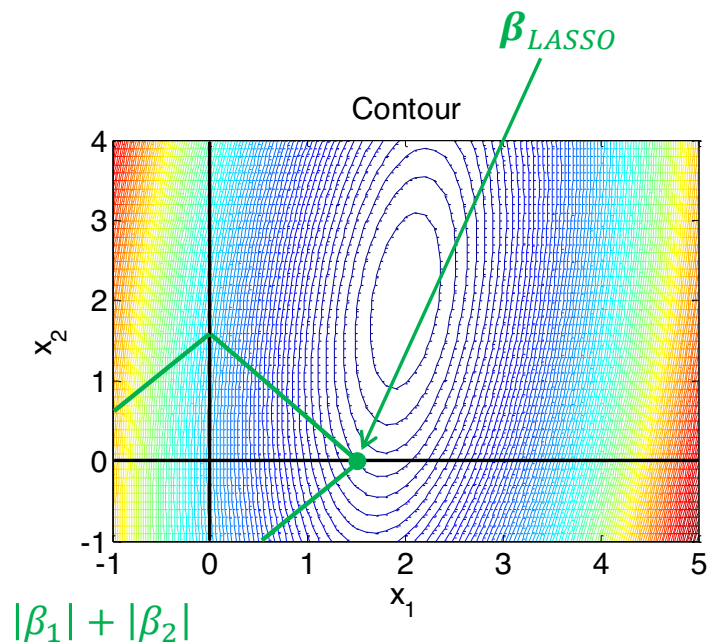
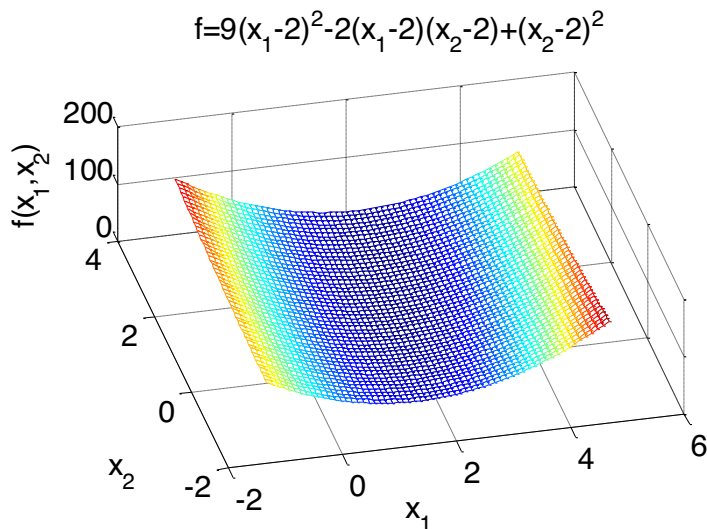
2D interpretation of the Ridge and LASSO constraints

Example

- Cost function:

$$f = 9(x_1 - 2)^2 - 2(x_1 - 2)(x_2 - 2) + (x_2 - 2)^2$$

- Model coefficients β of less significant explanatory variables goes to zero first



Example MATLAB Code

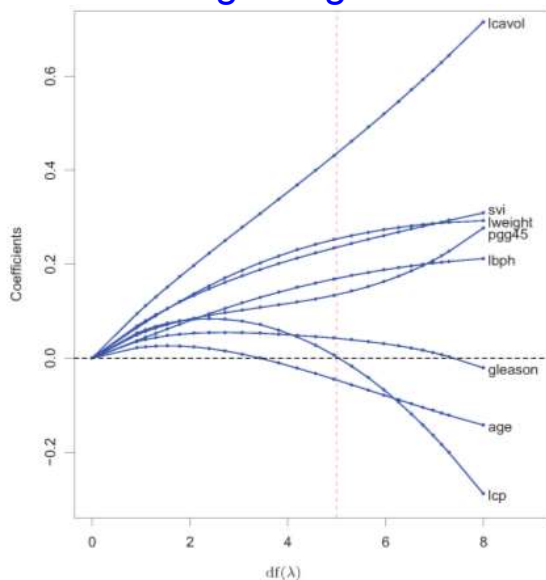
```
clear; close all;
% generate data
[x1,x2]=meshgrid(-1:.1:5,-1:.1:4);
y=9*(x1-2).^2-2*(x1-2).*(x2-2)+(x2-2).^2;
mesh(x1,x2,y); zlabel('f(x_1,x_2)', 'FontSize', 16);
xlabel('x_1', 'FontSize', 16); set(gca,'FontSize', 16);
ylabel('x_2', 'FontSize', 16); set(gcf, 'Color', 'w');
title('f=9(x_1-2)^2-2(x_1-2)(x_2-2)+(x_2-2)^2');

figure; contour(x1,x2,y,100); set(gca,'FontSize', 16);
xlabel('x_1', 'FontSize', 16); line([0 0],[-1 4]);
ylabel('x_2', 'FontSize', 16); line([-1 5],[0 0]);
set(gcf, 'Color', 'w'); title('Contour');
```

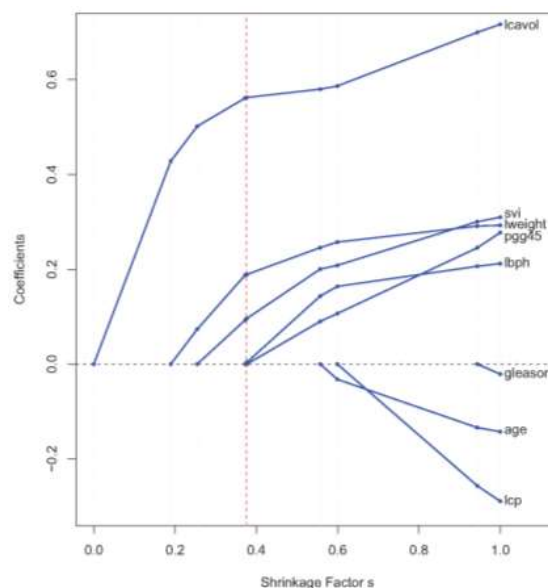
Why LASSO?

- LASSO is parsimonious
- LASSO results in sparse models which lend themselves more easily for interpretation

Ridge Regression



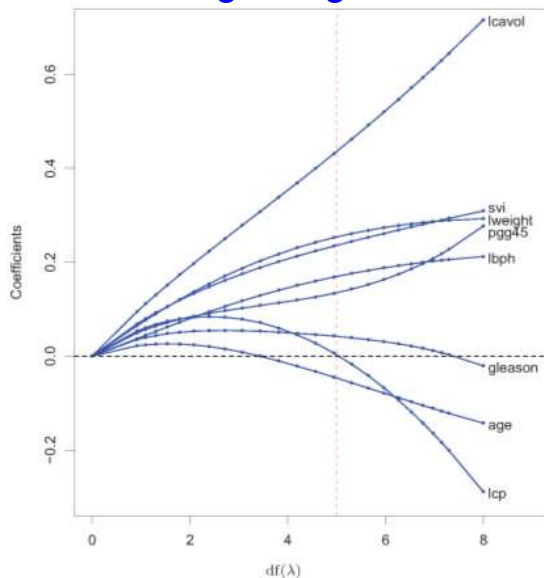
LASSO



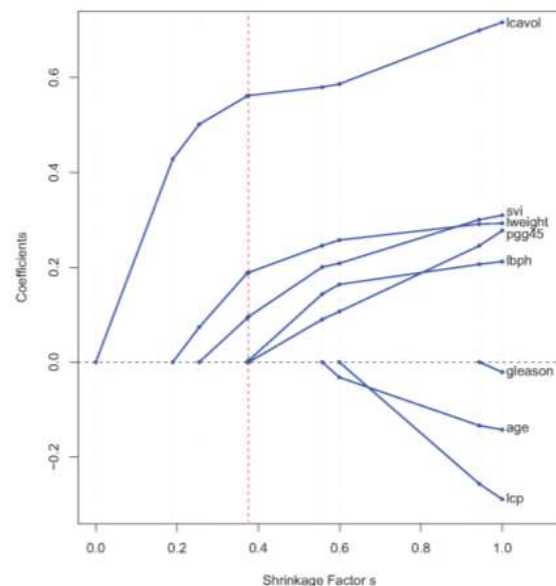
How to Choose the Number of the Coefficients?

- One can obtain many Ridge and LASSO models following the procedure
- Which one is the “best”, i.e., the degree of shrinkage?

Ridge Regression



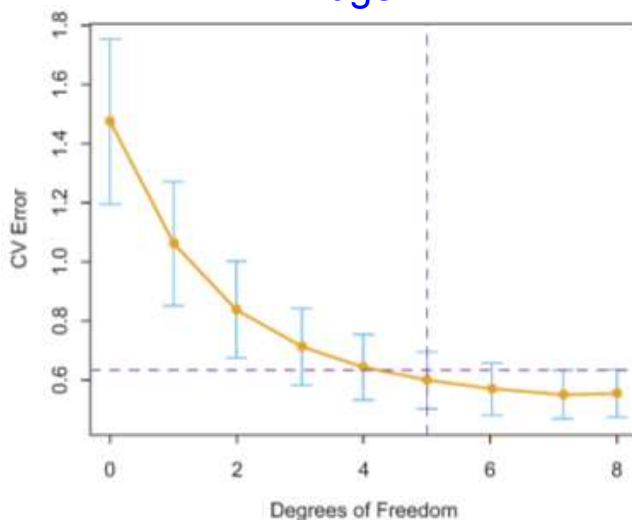
LASSO



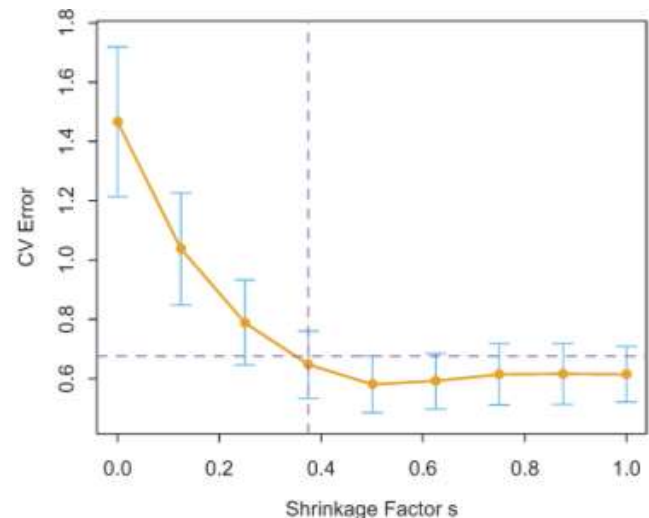
How to Choose the Number of the Coefficients?

- Typically based on cross-validation
- One standard error rule – choose the most parsimonious model whose error is no more than one standard error above the error of the best model

Ridge



LASSO



Why Sparsity Important?

- In many cases, the response of a model is determined by just a small subset of the explanatory variables
- For example, identifying if the person in the picture wears glasses or not

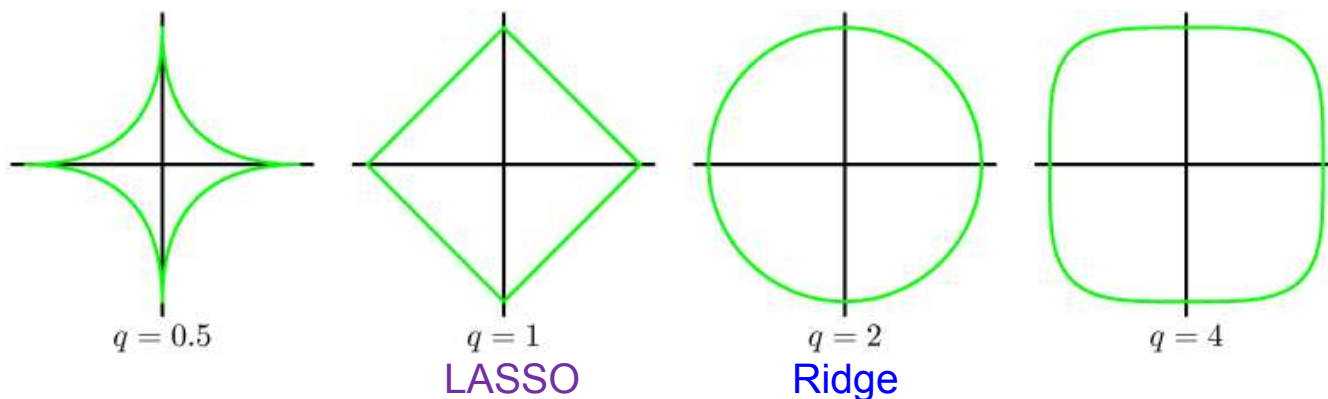


A Unifying View

- Constraint least-squares:

$$\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda |\boldsymbol{\beta}|^q)$$

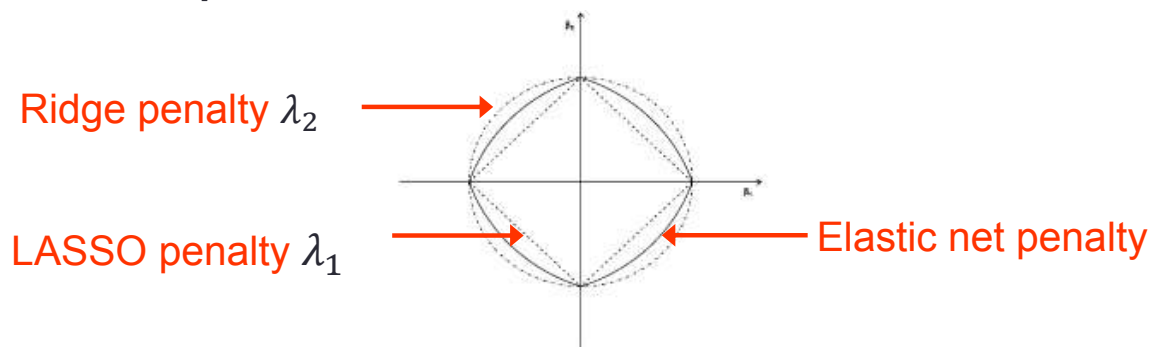
- $\lambda = 0$: least squares
- $\lambda > 0, q = 0$: subset selection
- $\lambda > 0, q = 1$: LASSO
- $\lambda > 0, q = 2$: Ridge regression



Elastic Net

- A method that overcome the limitation of LASSO
- Especially works well with data with a high degree of multicollinearity
- It compromises between L_1 and L_2 penalty
- Lost function:

$$\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|^2)$$

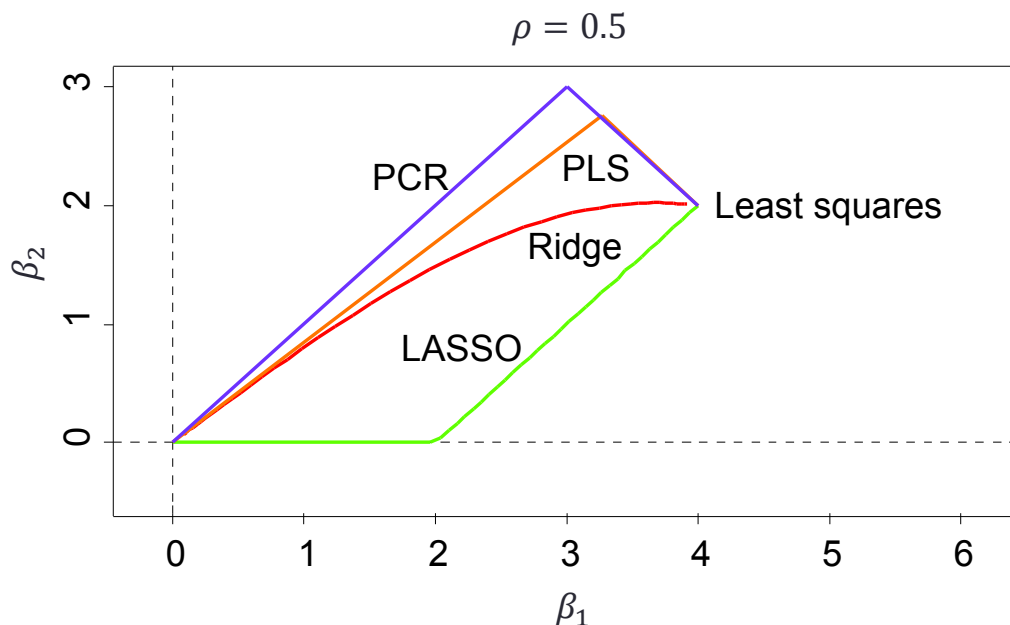


Comparison of PCR, PLSR, Ridge, and LASSO

- Biased PCR discards lower variance components
- PLSR shrinks the low-variance direction, while inflates some high variance direction
- PCR and PLSR are discrete shrinkage methods, while Ridge regression and LASSO are continuous methods
- Ridge shrinks all directions but shrinks the low-variance directions most
- Biased PCR, PLSR, and Ridge regression can't zero out coefficients; thus, you end up including all the coefficients in the model
- LASSO does both parameter shrinkage and variable selection automatically

Compare Selection and Shrinkage

- Consider an example with two correlated inputs x_1 and x_2 , with correlation $\rho = 0.5$
- Assume that the least squares coefficients are $\beta_1 = 4$ and $\beta_2 = 2$

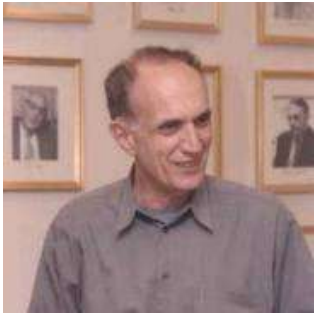


Hastie, Tibshirani,
and Friedman. *The
Elements of
Statistical Learning*

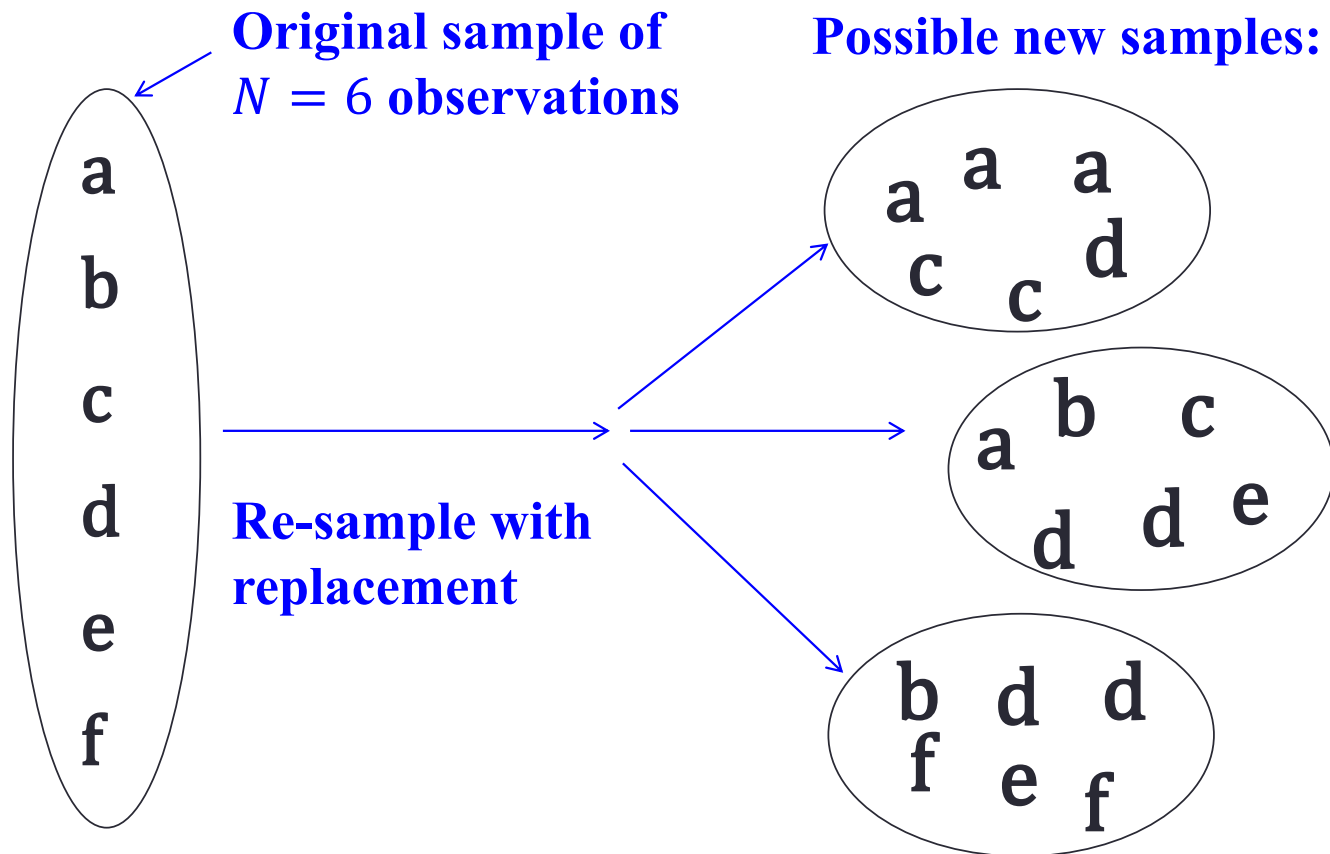
Summary

- All models are wrong; some, though are better than others and we can search for the better alternatives (Mc Cullagh & Nelder 1983)
- Increasing dimensionality of features increases the data requirements exponentially
- Dimensionality reduction techniques are tools to reduce the risk of overfitting
- Remember: standardize the inputs since most of the algorithms are sensitive to scaling of the parameters

Bootstrapping

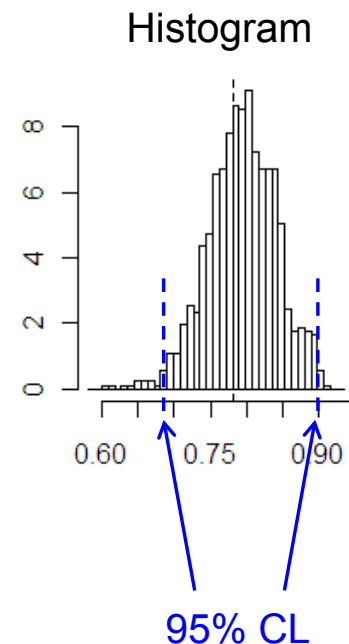
- A technique that allows estimation of the sampling distribution of any statistic using only very limited number of samples
 - This is achieved through resampling – sampling with replacement from an original dataset
 - For use in obtaining statistical estimates, e.g., standard error of model coefficients
 - Described by Bradley Efron in 1979
- 
- A portrait of Bradley Efron, a man with thinning hair, wearing a light-colored button-down shirt, smiling slightly. He is standing in front of a wall with several framed pictures.
- Why bootstrap – it makes no assumptions about the underlying distribution in the population

Sampling with Replacement



Bootstrapping Procedure

1. Number your observations 1, 2, ..., N
2. Draw a random sample of size N with replacement
3. Calculate the statistic, e.g., mean, coefficients, etc., with these data
4. Repeat steps 1-3 many times, e.g., 1000 times
5. Find the confidence intervals directly from the sample of 1000 statistics



References

- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Chapter 3 and 7
- C. M. Bishop, Pattern Recognition and Machine Learning