INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo
Dept. of Bio-industrial Mechatronics Engineering
National Taiwan University

Today:

- Constraint least squares regression methods
- Bootstrapping

Outline

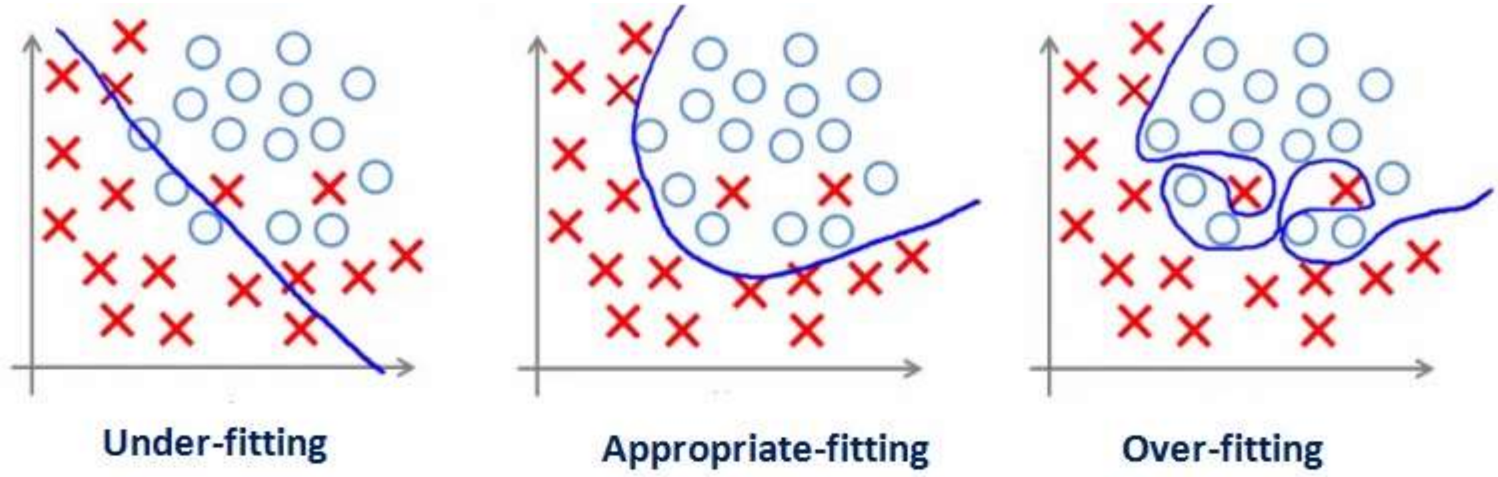
- Goal of the lecture
- Constraint least-squares
- Bootstrapping

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 <u>excessively complex</u>
- A overfitted model describes random error or noise instead of the underlying relationship



https://affineanalytics.wordpress.com/

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: $\hat{y} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}$
- Given a response vector $\boldsymbol{y}$ and explanatory matrix $\boldsymbol{X}$ from N observations, the model coefficients $\boldsymbol{\beta}$ can be obtained in least-squares sense, i.e.,

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

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

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:

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

- This is equivalent to

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

where $0 < \lambda \in \mathfrak{R}$ 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:

$$\nabla_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}\left(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||^2\right)$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}}\left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right)$$

$$= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\right)$$

$$= -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} = 2\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)\boldsymbol{\beta} - 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

$$\nabla_{\boldsymbol{\beta}} = 0 \quad \Longrightarrow \quad \boldsymbol{\beta}_{Ridge} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Recall: SVD of A Matrix

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

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

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

$$\underset{M \times N}{\begin{bmatrix} \\ \boldsymbol{X} \\ \\ \end{bmatrix}} = \underset{M \times M}{\begin{bmatrix} | & & | \\ \boldsymbol{e}_1 & \cdots & \boldsymbol{e}_M \\ | & & | \end{bmatrix}} \underset{M \times N}{\begin{bmatrix} \sigma_1 & & \varnothing \\ & \ddots & \\ \varnothing & & \sigma_M \\ \hdashline \varnothing & & \varnothing \end{bmatrix}} \underset{N \times N}{\begin{bmatrix} — & \boldsymbol{f}_1 & — \\ — & \boldsymbol{f}_N & — \end{bmatrix}}$$

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

Algebraic Interpretation of Ridge Regression

- Consider the fitted response:

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

- Introduce the SVD of $\boldsymbol{X}$

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

$$\hat{\boldsymbol{y}} = \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\left(\left(\boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\right) + \lambda\boldsymbol{I}\right)^{-1}\left(\boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\right)^{\mathrm{T}}\boldsymbol{y}$$

$$= \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\left(\boldsymbol{F}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{F}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{y}$$

$$= \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\left(\boldsymbol{F}\boldsymbol{\Sigma}^{2}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{F}(\lambda\boldsymbol{I})\boldsymbol{F}^{\mathrm{T}}\right)^{-1}\boldsymbol{F}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{y}$$

Algebraic Interpretation of Ridge Regression (Cont'd)

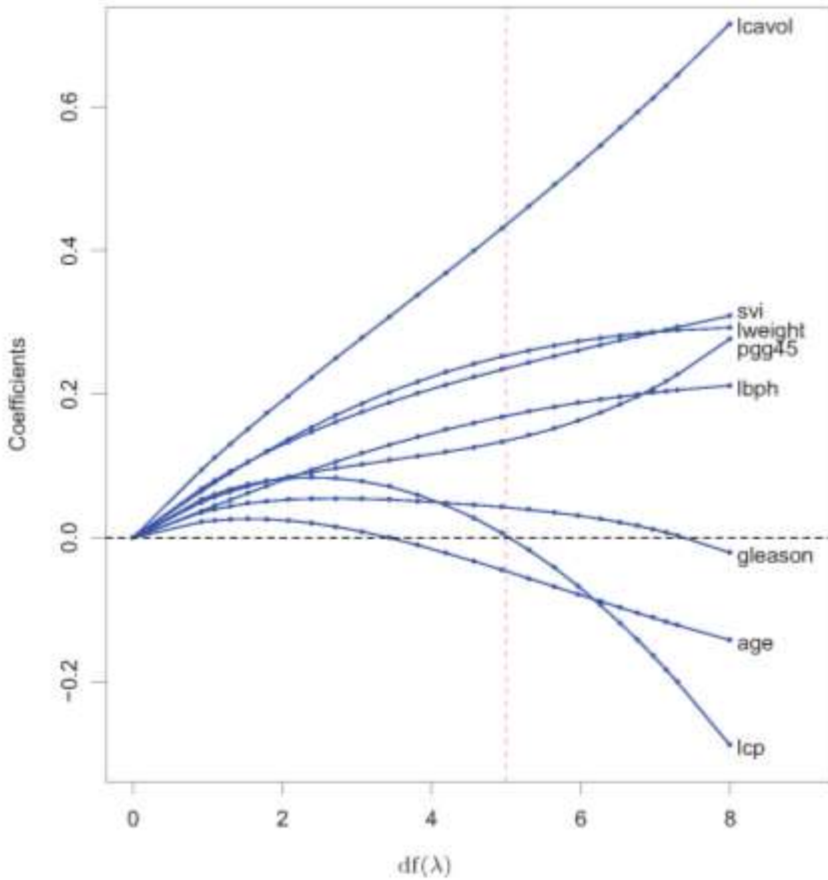
$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\left(\boldsymbol{F}(\boldsymbol{\Sigma}^2 + \lambda\boldsymbol{I})\boldsymbol{F}^{\mathrm{T}}\right)^{-1}\boldsymbol{F}\boldsymbol{\Sigma}^{\mathrm{T}}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{y} \\
&=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

- 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 $\boldsymbol{\beta}$ is measured in L_1 sense, the optimization problem can be written as:

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

- This is equivalent to

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

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

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

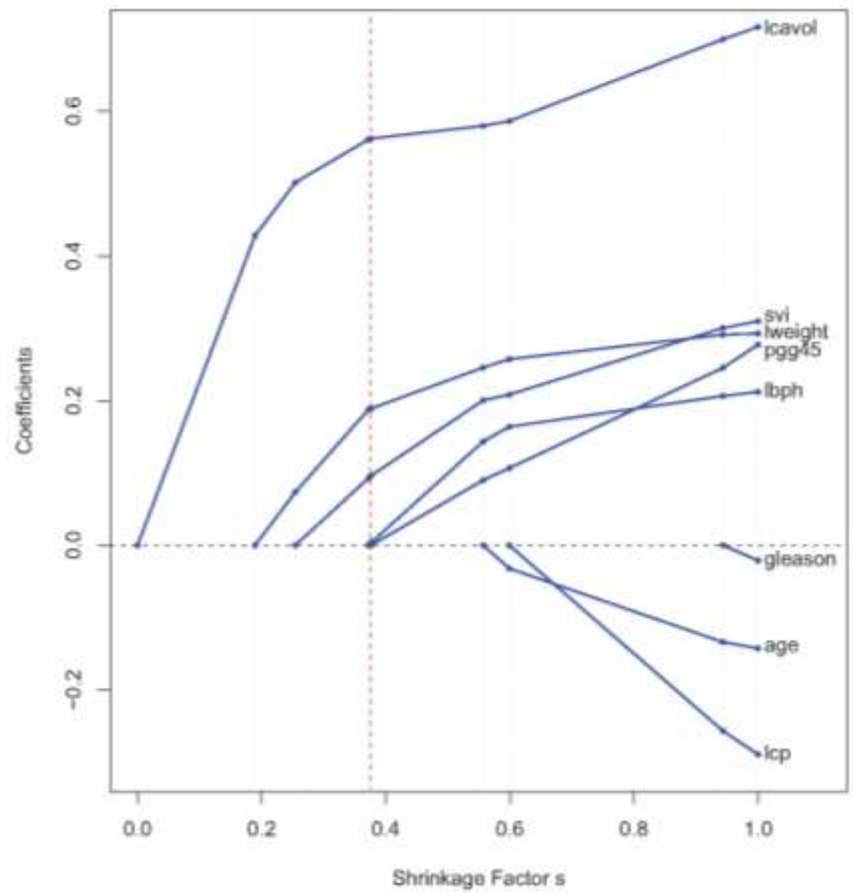
LASSO Coefficient against Factor s

- The shrinkage factor s is defined as

$$s = \frac{c}{\sum_{i=1}^{N} |\boldsymbol{\beta}_{LS}|} \in \Re$$

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

LASSO

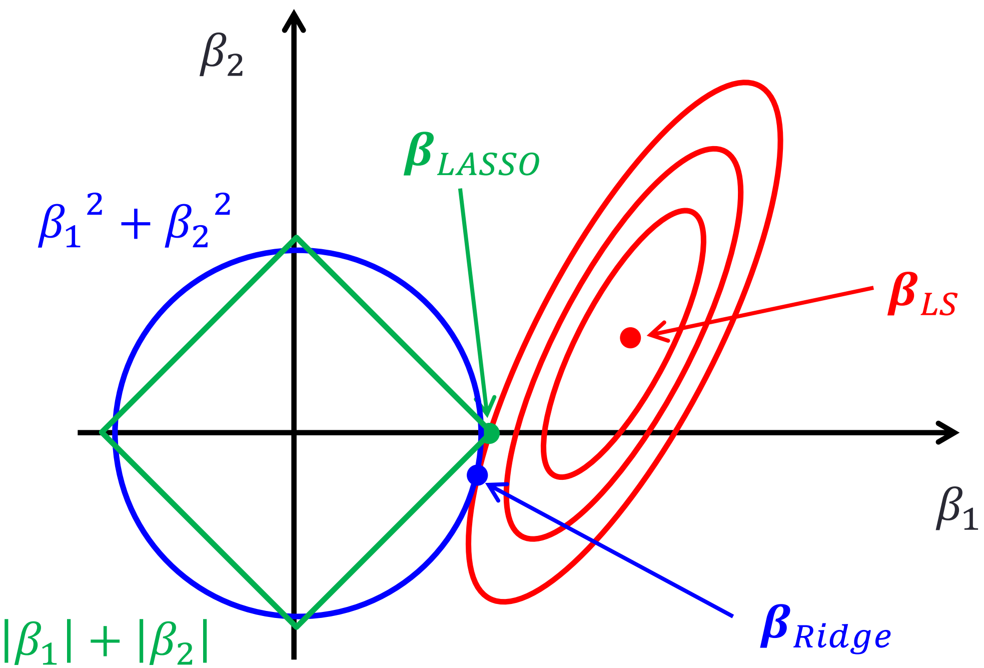


Hastie, Tibshirani, and Friedman.
The Elements of Statistical Learning

$$s = \frac{c}{\sum_{i=1}^{N} |\boldsymbol{\beta}_{LS}|}$$

Geometry Interpretation of Ridge & LASSO

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



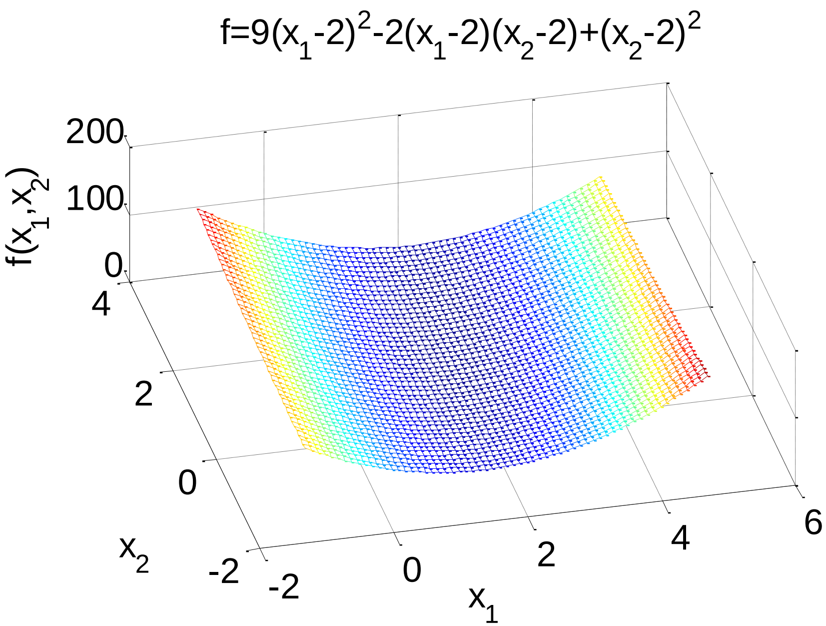
2D interpretation of the Ridge and LASSO constraints

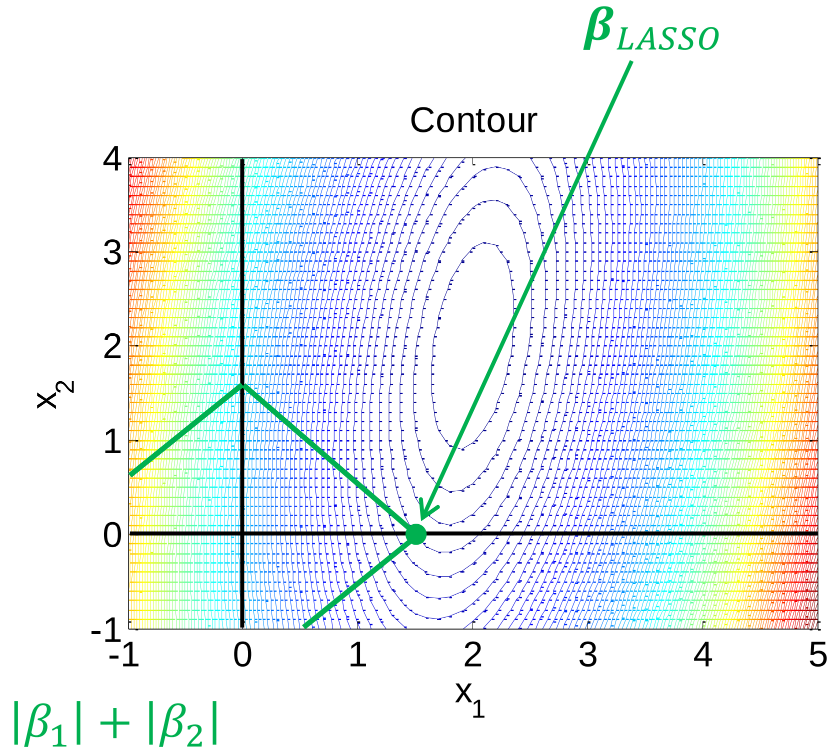
Example

- Cost function:

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

- Model coefficients $\boldsymbol{\beta}$ of less significant explanatory variables goes to zero first





Example MATLAB Code

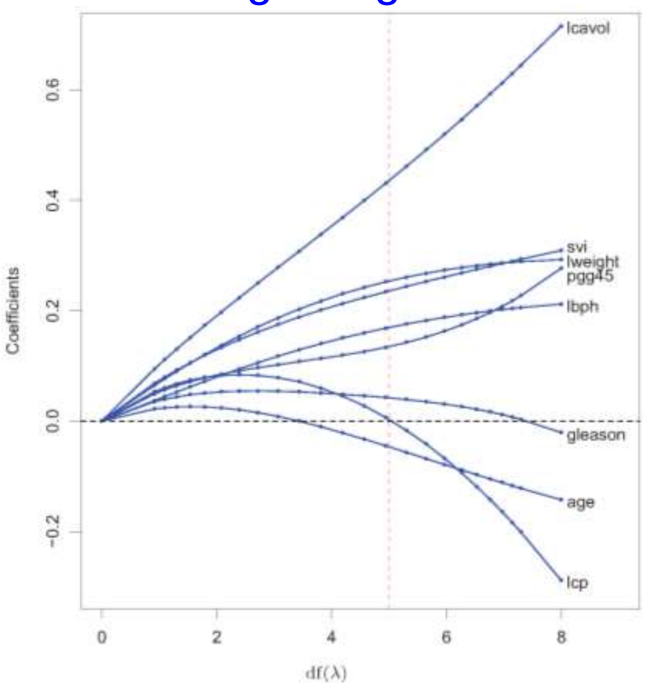
```
clear; close all;
% generate data
[x1,x2]=meshgrid(-1:.1:5,-1:.1:4);
y=9*(x1-2).^2-2*(x1-2).*(x2-2)+(x2-2).^2;
mesh(x1,x2,y); zlabel('f(x_1,x_2)', 'FontSize', 16);
xlabel('x_1', 'FontSize', 16); set(gca,'FontSize', 16);
ylabel('x_2', 'FontSize', 16); set( gcf, 'Color', 'w');
title('f=9(x_1-2)^2-2(x_1-2)(x_2-2)+(x_2-2)^2');

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');
```

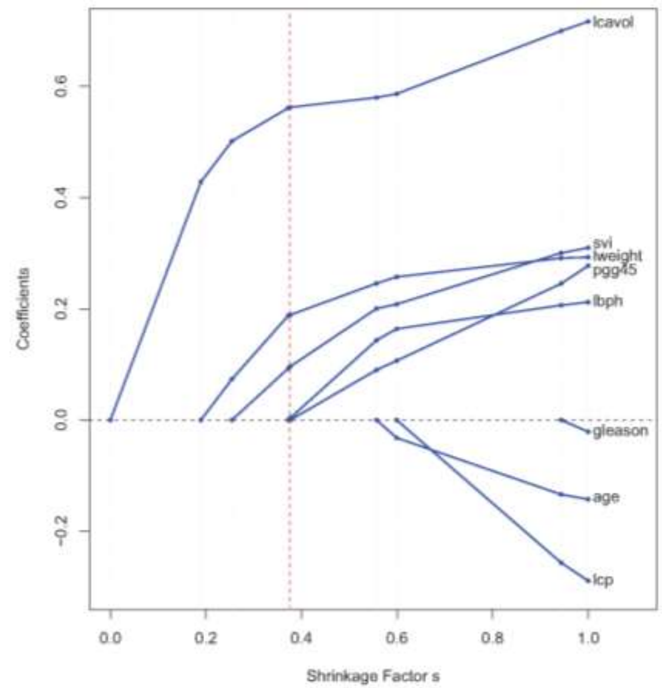
Why LASSO?

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

Ridge Regression



LASSO

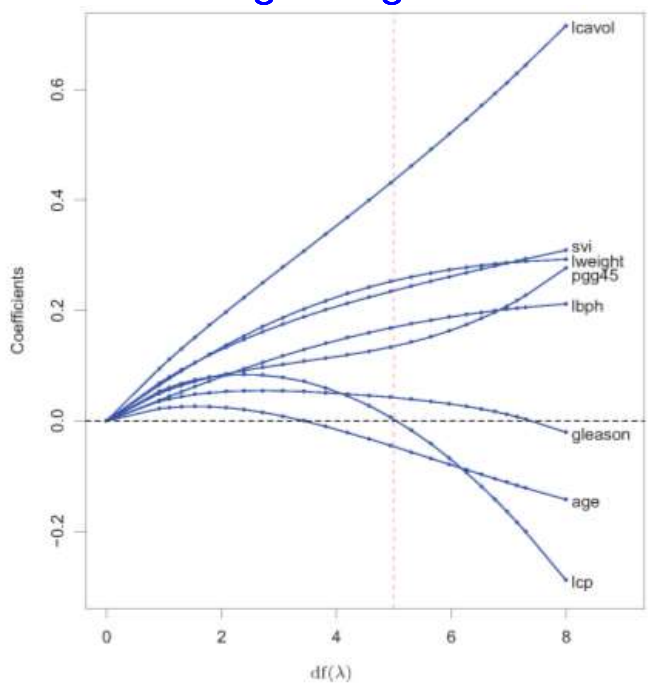


Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning*

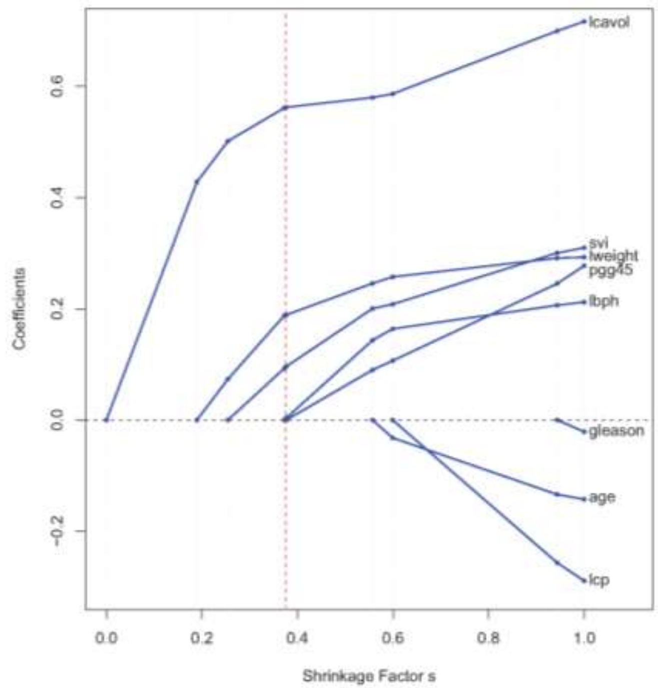
How to Choose the Number of the Coefficients?

- 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

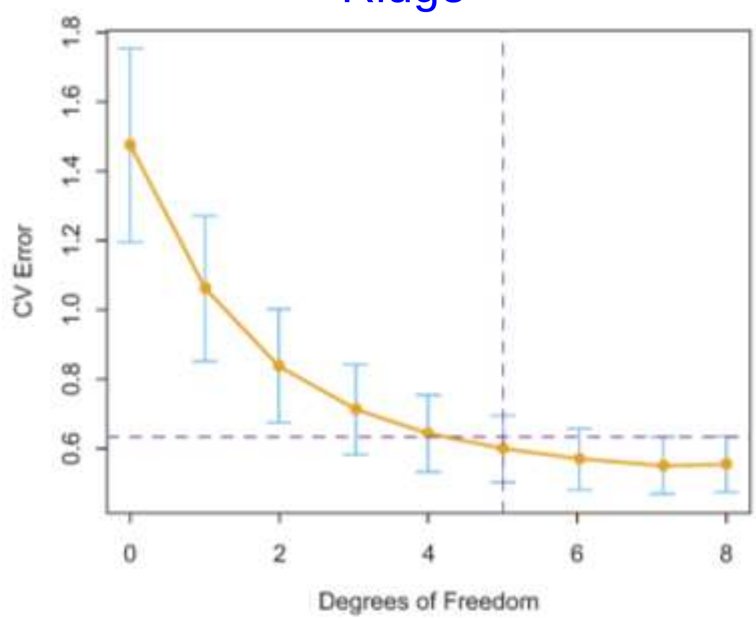


Hastie, Tibshirani, and Friedman. *The Elements of Statistical Learning*

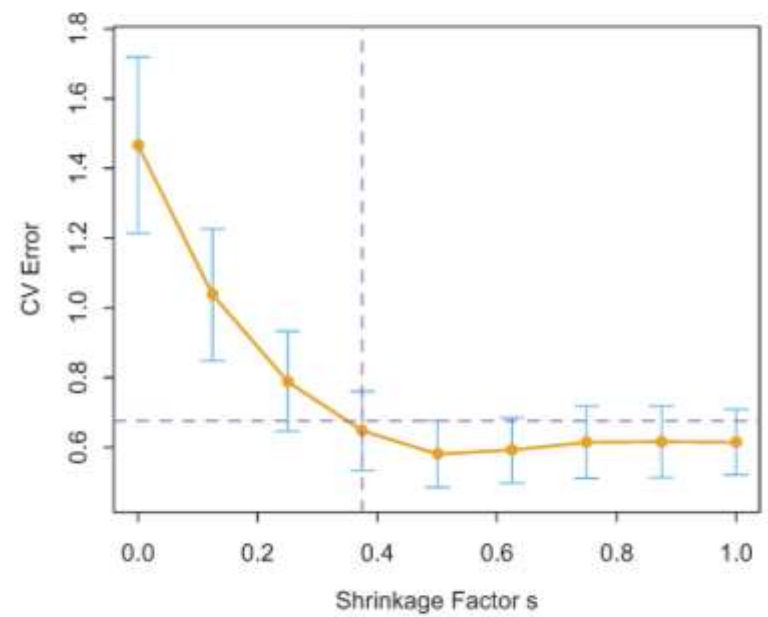
How to Choose the Number of the Coefficients?

- 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



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

Why Sparsity Important?

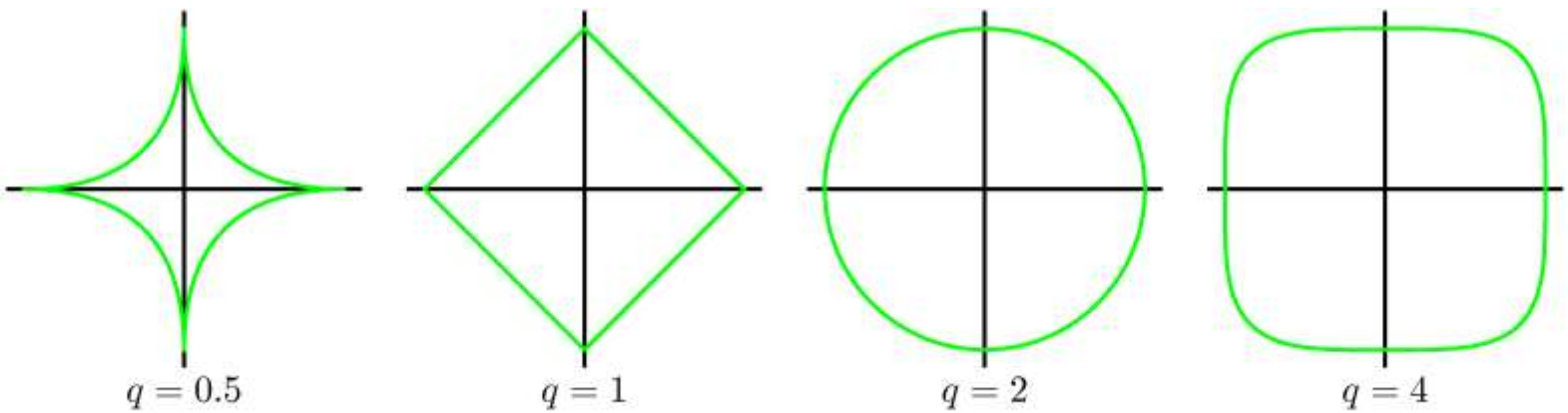
- In many cases, the response of a model is determined by just a small subset of the explanatory variables
- For example, identifying if the person in the picture wears glasses or not

A Unifying View

- Constraint least-squares:

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

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



LASSO Ridge

Bishop. *Pattern Recognition and Machine Learning*

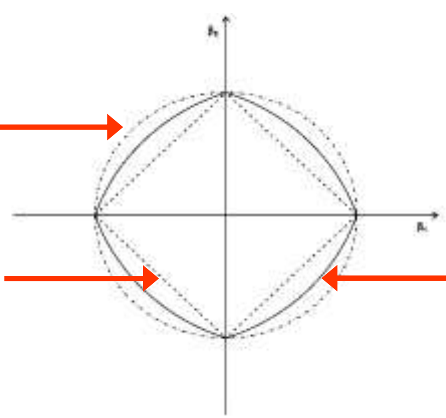
Elastic Net

- 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:

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

Ridge penalty λ_2

LASSO penalty λ_1



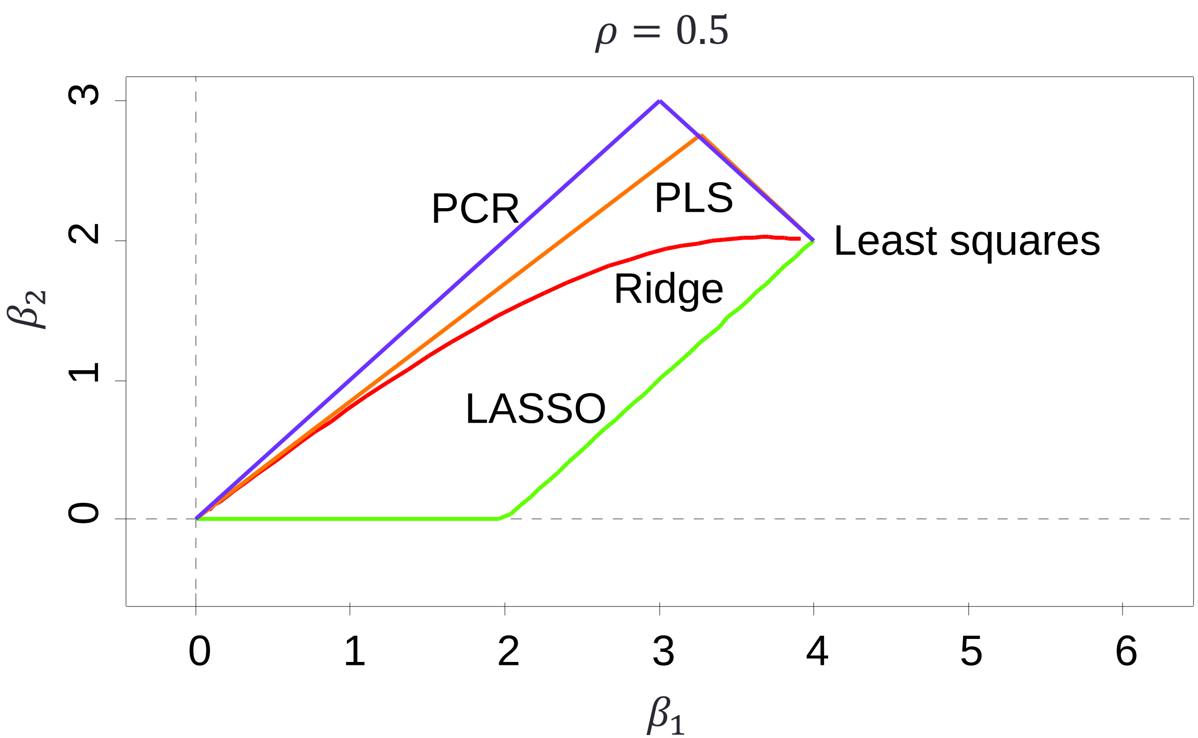
Elastic net penalty

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

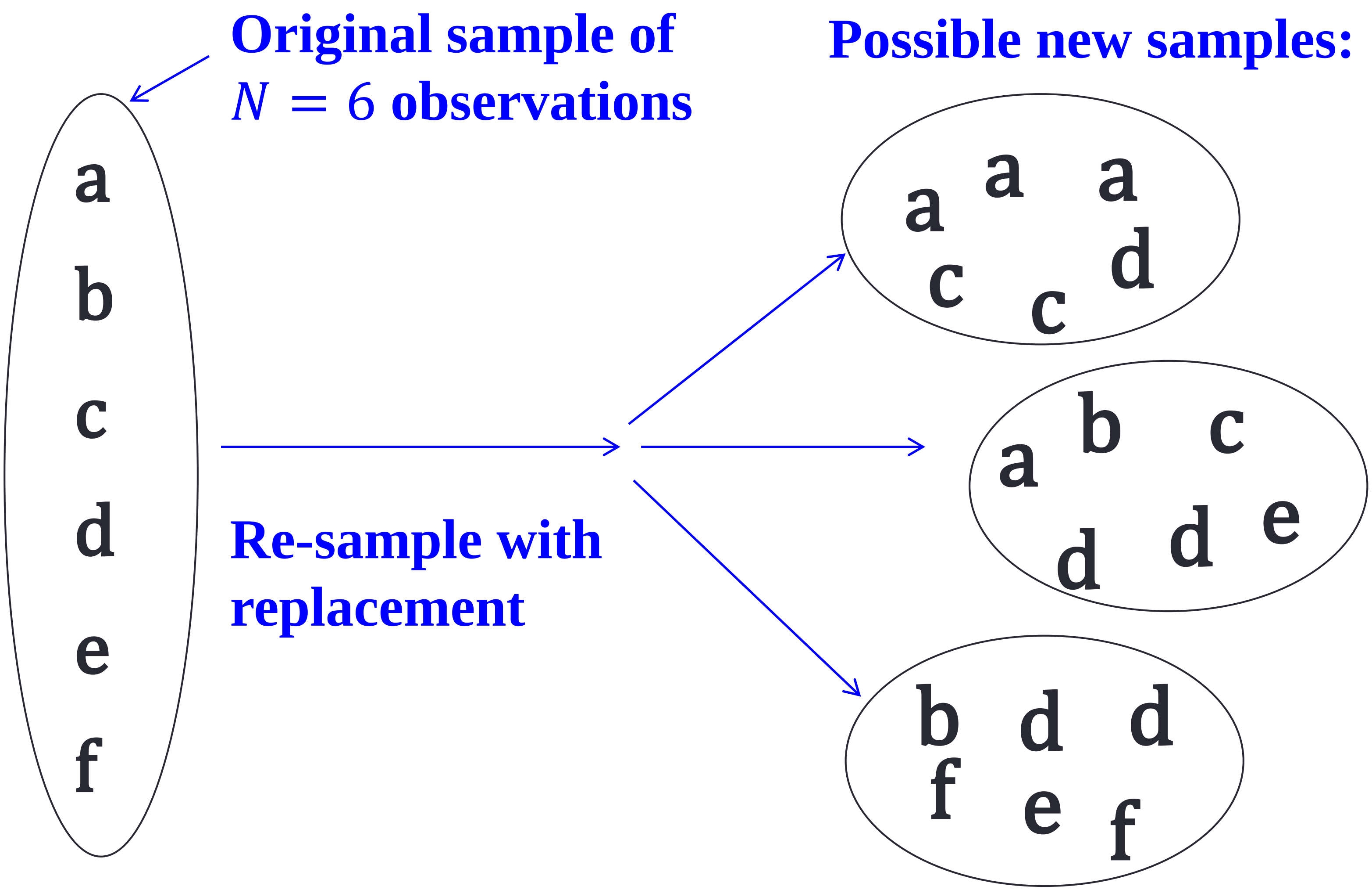
- All models are wrong; some, though are better than others and we can search for the better alternatives (Mc Cullagh & Nelder 1983)
- Increasing dimensionality of features increases the data requirements exponentially
- Dimensionality reduction techniques are tools to reduce the risk of overfitting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

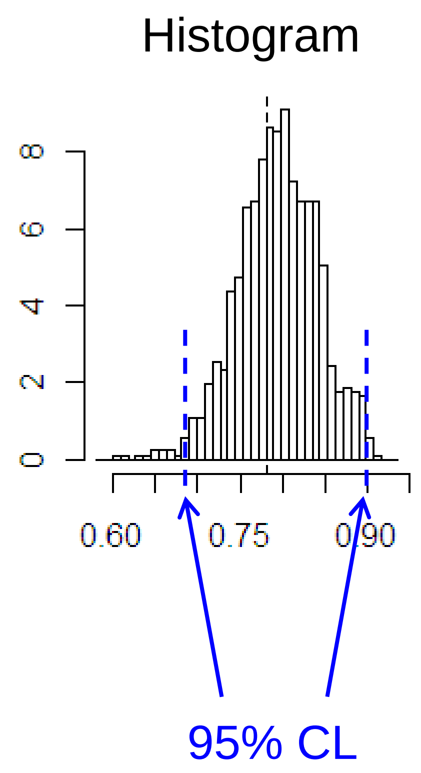
- Why bootstrap – it makes no assumptions about the underlying distribution in the population

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