Chapter 3-4: Linear Models for Regression

Richard Liu

School of Mathematics, XMU

May 3, 2020

Content

Source

- Trevor Hastie, et al. *The Elements of Statistical Learning*
- Jie Hu, *Applied Linear Models Course in XMU*
- Kui Du, *Numerical Linear Algebra Course in XMU*
- Wen Huang, *Numerical Optimization Course in XMU*

Section 1

Introduction

Introduction

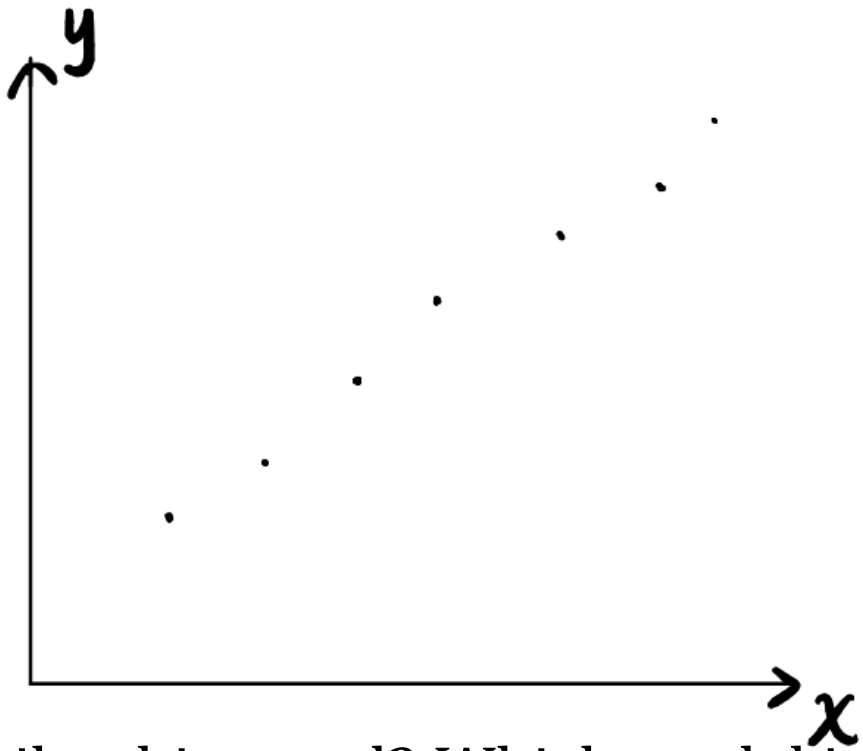
- Consider a classification model with data X and target Y, we know that X are determined objects used for predicting the values of Y.
- And we know that a regression model differs in the property of Y (from categorical to numerical).
- That is why we need Regression Analysis.
  - Note: if X and Y are both random data, then the subject analyzing their relationship is named Correlation Analysis.

Section 2

Univariate OLS

Case Study

Suppose that there are 2 rows of data. Undeniably, x increases, then y increases.



How to describe this trend? Which model is the best?

What are Linear Models?

Definition 1: Linear Models

Assume that $f(x) = \sum_{i=1}^{n} a_i x_i$, then $f(x)$ is called a linear model if $\{a_i = a(y_i)\}$ are all linear for $i = 1, 2, \cdots, n$ w.r.t. y_i.

Definition 2: Independent Variables, Parameters

a_i is called a parameter, while x_i is called an independent variable.

- Example: $f(x) = ax^2 + bx + c$ is a linear model because a, b, c are parameters with power 1. $f(x) = a^2 x + b$ is not a linear model.
- Use variable substitution could help us release the burden analyzing the cases variable x_i is nonlinear.

Why Linear Models?

- Easy!
  - *All models are wrong, but some are useful.*
- Generalizable!
- Theoretical Guaranteed!

We could use linear models to help us describe the trend in the previous case.

Which model is the best?

- Intuitively, when we use a linear function to describe the data, the value the linear function provides should not differ a lot compared with the real data points.
- How to analyze the difference?
  - A given linear function and a given cost function.

Problem Formulation

General Form

$$E(y|x) = y_i = f(x) = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$, inter-independent. (Gauss-Markov)

- Given estimated $\hat{\beta}_0$, $\hat{\beta}_1$, we have $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Y is called dependent variable.

OLS Objective

$$\min_{\beta_0, \beta_1} \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Problem: What are the hats on y_i, β_0, β_1 ?

Final Results

Theorem 1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proof

FOC + Crammer's Law, □

Theoretical Properties of $\hat{\beta}_1$ and $\hat{\beta}_0$

Theorem 2

$E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_0) = \beta_0$

Proof

$$
\begin{aligned}
E(\hat{\beta}_1) &= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} E(y_i) \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} (\beta_0 + \beta_1 x_i) \\
&= \beta_1 \sum_{i=1}^{n} \frac{(x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \beta_1, \square
\end{aligned}
$$

Theoretical Properties of $\hat{\beta}_1$ and $\hat{\beta}_0$

Theorem 3

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \text{Var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \sigma^2$$

Proof(Part 1)

$$\begin{aligned}
\text{Var}(\hat{\beta}_1) &= \text{Var}\left(\sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} y_i\right) \\
&= \sum_{i=1}^n \text{Var}\left(\frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} y_i\right) \\
&= \sigma^2 \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \square
\end{aligned}$$

Theoretical Properties of $\hat{\beta}_1$ and $\hat{\beta}_0$

Proof(Part 2)

$$\begin{aligned}\text{Var}\left(\hat{\beta}_0\right) &= \text{Var}\left(\bar{y} - \hat{\beta}_1\bar{x}\right) = \text{Var}(\bar{y}) + \text{Var}\left(\hat{\beta}_1\bar{x}\right) \\ &= \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]\sigma^2\end{aligned}$$

□

Theorem 4

$$\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -\frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Proof

$$\begin{aligned}\text{Var}(\bar{y}) = \frac{\sigma^2}{n} &= \text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1\bar{x}\right) \\ &= \text{Var}\left(\hat{\beta}_0\right) + \text{Var}\left(\hat{\beta}_1\right)\bar{x}^2 + 2\bar{x}\,\text{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right), \square\end{aligned}$$

Section 3

Multivariate OLS

Introduction

- In the previous case, we assume that 1-d independent variable X is used to explain dependent variable Y.
- However! If there is more than one factor?
  - Multivariate Linear Models!

General Form

$$E(y|x) = y_i = f(x) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Meaning?

Matrix Formulation

Because several factors co-explain the response (dependent variable), we could write as a matrix formulation.

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

$$\text{where } \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Compute $\hat{\beta}$

Given the same OLS objective, how to compute $\hat{\beta}$ more generally?

Theorem 5

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Proof

By FOC, we have $\mathbf{e}^T\mathbf{X} = 0$, where $\mathbf{e} = (e_i)_{i=1}^n$, $e_i = y_i - \hat{y}_i$. Then

$$\mathbf{Y}^T\mathbf{X} = \hat{\beta}^T\mathbf{X}^T\mathbf{X}$$

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\beta}$$

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \square$$

Theoretical Properties of $\hat{\beta}$

Definition 3: Hat Matrix

We denote $H = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ for simplicity.

Theorem 6

$E(\hat{\beta}) = \beta$

Proof

$$E(\hat{\beta}) = E(H\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta, \square$$

Theoretical Properties of $\hat{\beta}$

Theorem 7

$$\mathrm{D}(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Proof

$$\mathrm{D}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{D}(y)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}, \square$$

Explanation

$$\mathrm{D}(A\beta) = A\,\mathrm{D}(\beta)A^T$$

Theoretical Properties of $\hat{\beta}$

Theorem 8

$\text{Cov}(\hat{\beta}, \mathbf{e}) = 0$

Proof

Note that $\mathbf{e} = (I - H)\mathbf{Y}$ and $(I - H)\mathbf{X} = 0$, $\square$.

Section 4

Models with penalty

Subsection 1

Ridge Regression

Introduction

- Why multivariate linear models still insufficient to solve the regression problem?
- Consider the case $\exists i, j$, s. t. β_i, β_j are linear dependent.
- WLOG, assume $\beta_1 = 2\beta_2$, what happens?
  - Multicollinearity!
  - Unable to explain and unable to solve $\hat{\beta}$ for $\mathbf{X}^T\mathbf{X}$ is not full-ranked.
- Models with penalty.

Ridge Regression

Definition 4: Ridge Regression Estimator

We denote $\hat{\beta}(k) = (\mathbf{X}^T\mathbf{X} + kI)^{-1}\mathbf{X}^T\mathbf{Y}$ the Ridge Regression Estimator.

Ridge Regression Objective

$$\hat{\beta} = \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$

Obviously, $\lambda\sum_{j=1}^{p}\beta_j^2$ is the penalty. (Meaning?)

Ridge Regression

Theorem 9

Prove the equivalence in the last page.

Proof

Let $f(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta$, then we have $\frac{\partial f}{\partial \beta} = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) + 2\lambda\beta$, FOC and SOC lead to the result. □

Ridge Regression is one kind of Shrinkage Methods.

Why Shrinkage?

We use several theorems and an important tool to explain this term.

Theorem 10

$\hat{\beta}(k) = A_k \hat{\beta}$, where $A_k = (\mathbf{X}'\mathbf{X} + kI)^{-1} \mathbf{X}'\mathbf{X}$

Proof

$\hat{\beta}(k) = (X'X + kI)^{-1} X'Y = (X'X + kI)^{-1} \left[(X'X) (X'X)^{-1} \right] X'Y, \square$

Prerequisite: Reduced-Rank Singular Value Decomposition

- For each matrix X, we have $X = U_p \Sigma_p V_p^T$, where U_p, V_p are two orthonormal matrix with size $n \times p, p \times p$ and Σ is a squared diagonal matrix.
- For each σ_i in $\Sigma_p = \text{diag}(\sigma_1, \cdots, \sigma_p)$, we call it singular value.
- Tightly correlated with eigenvalue.

Why Shrinkage?

Theorem 11

For all $k > 0$, we have $\|\hat{\beta}(k)\| < \|\hat{\beta}\|$.

Proof

By Theorem 10, we only need to compute $(\mathbf{X}'\mathbf{X} + kI)^{-1}\mathbf{X}'\mathbf{X}$.
By Reduced-Rank SVD and note that
$(\mathbf{X}'\mathbf{X} + kI)^{-1}\mathbf{X}'\mathbf{X} = U(\Sigma^2 + \lambda I)^{-1}\Sigma^2 V^T$, we have

$$\|\hat{\beta}(k)\| = \|(\Sigma^2 + \lambda I)^{-1}\Sigma^2\hat{\beta}\| \leq \|(\Sigma^2 + \lambda I)^{-1}\Sigma^2\|\|\hat{\beta}\|$$

Σ is diagonal leads to the result.□

Why Shrinkage?

Compare two estimators

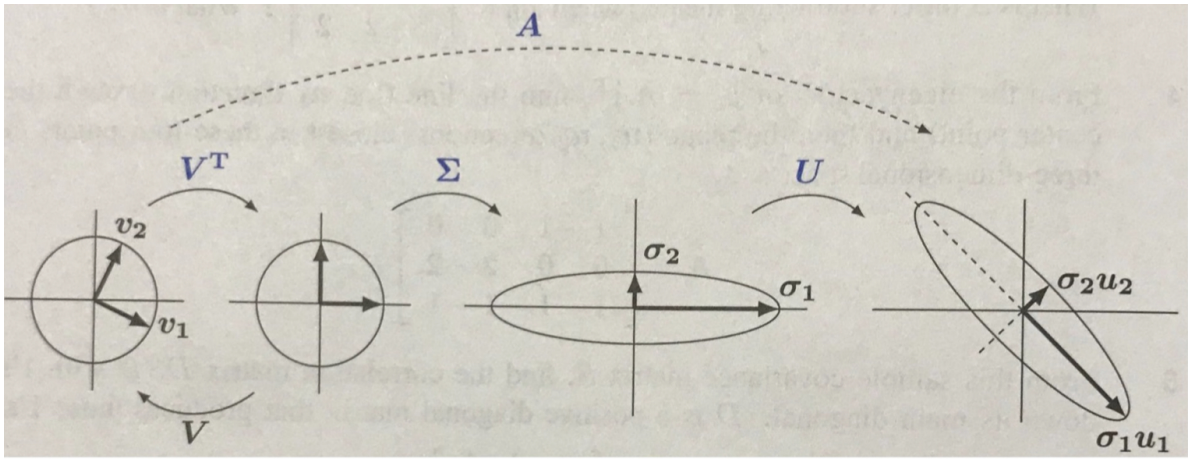
$$\hat{y}_1 = X\hat{\beta}(k) = U\Sigma(\Sigma^2 + \lambda I)^{-1}\Sigma U^T\mathbf{Y} = \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2 + k} u_j^T\mathbf{Y}$$

$$\hat{y}_2 = X\hat{\beta} = UU^T\mathbf{Y} = \sum_{j=1}^{p} u_j u_j^T\mathbf{Y}$$

So we could find that, for smaller σ_j, the shrinkage will be greater (why?).

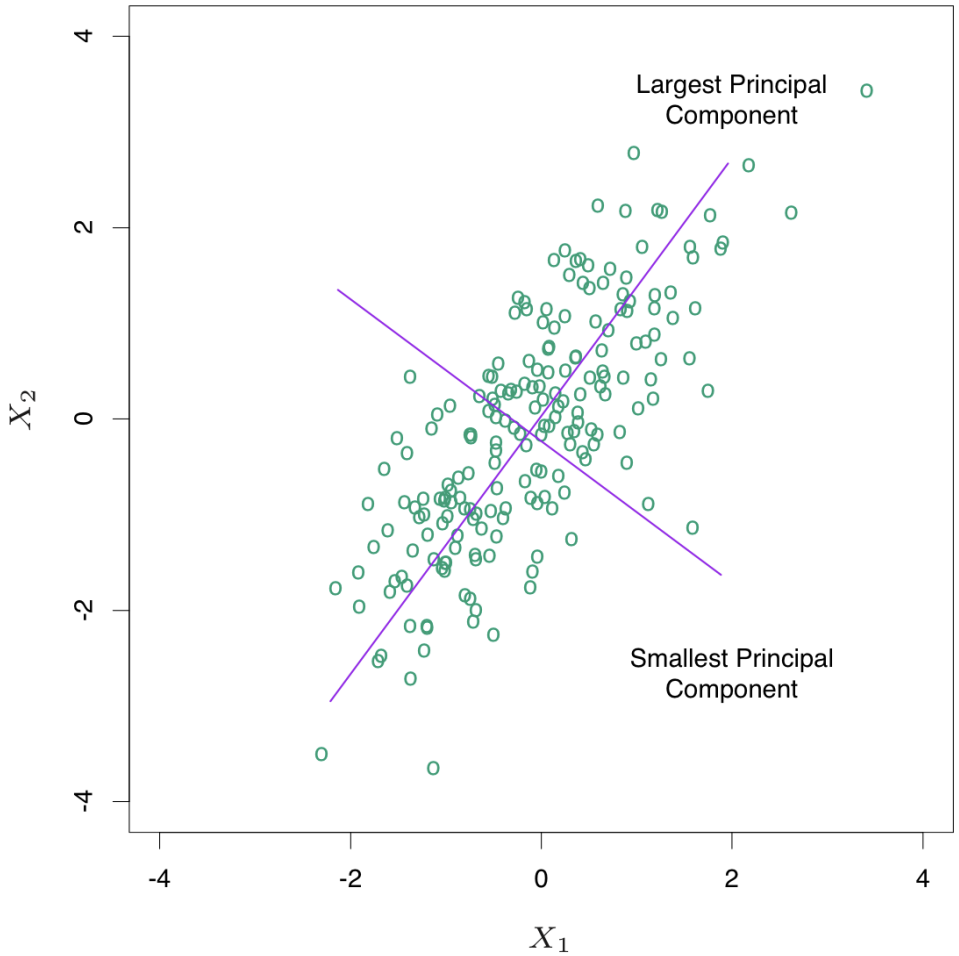
Behavior

Consider the geometric properties of SVD. (Or PCA)



The two values of an SVD are the maximum and the minimum length of two diameters.

Behavior


Largest Principal Component
Smallest Principal Component
X_1
X_2
-4
-2
0
2
4

Behavior

- Because $\mathbf{X}^T\mathbf{X} = V\Sigma^2 V^T$ is the eigen decomposition of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{X}v_1$ has the largest sample variance, which leads to the longest diameter of the ellipse.
- More information about PCA will be mentioned later.

FYI

In fact, there are many other properties about Ridge Regression in Statistics. More information could be seen in https://zhuanlan.zhihu.com/p/51431045.

Subsection 2

LASSO

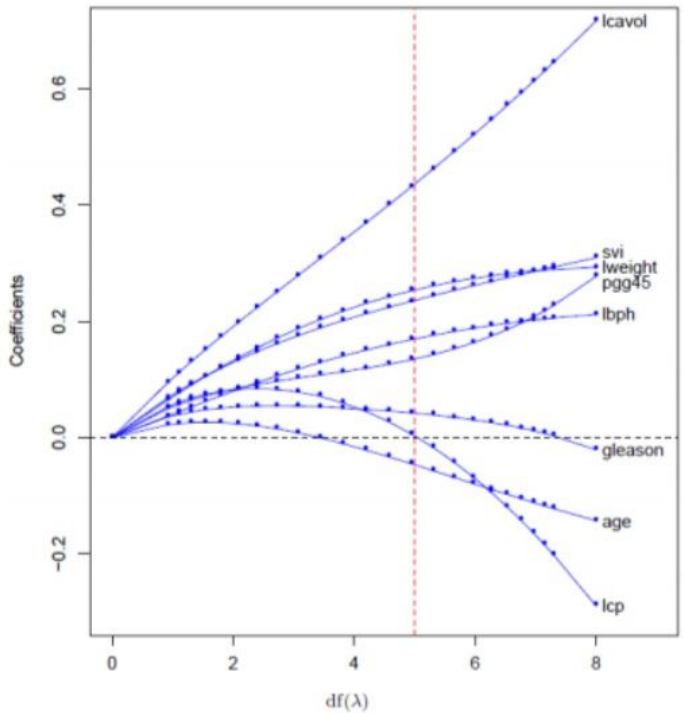
Introduction

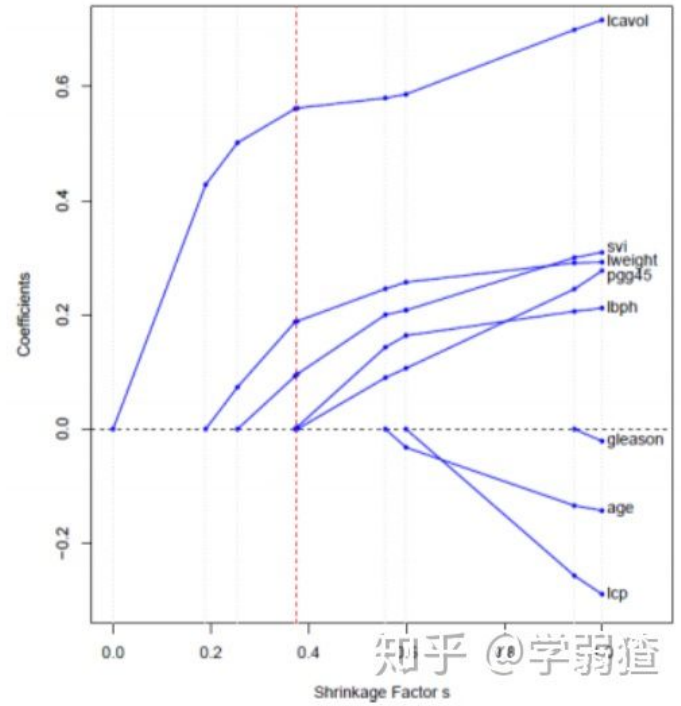
- Could we change the penalty? Obviously could!
- From 2-norm to 1-norm
  - Least Absolute Selection and Shrinkage Operator (LASSO)
  - Have some unexpected properties.

LASSO Objective

$$\hat{\beta} = \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

Comparison with Ridge Regression



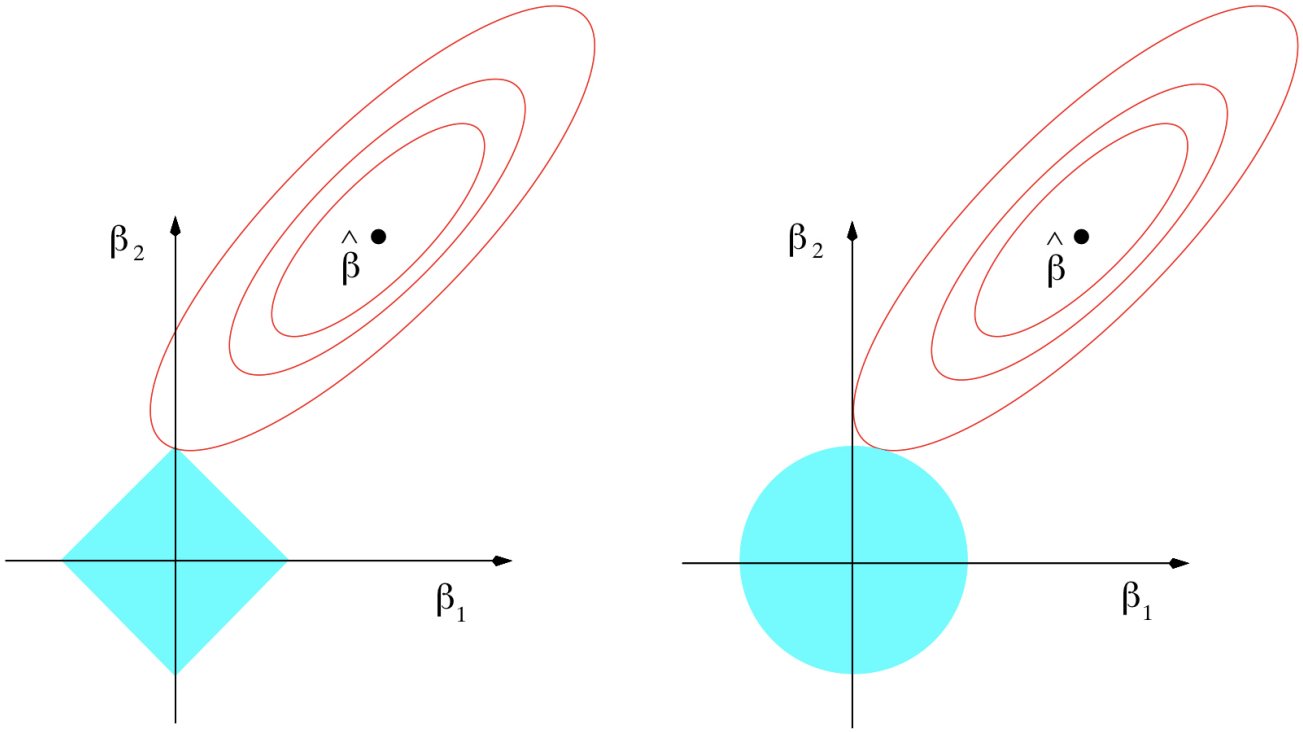


Here $\mathrm{df}(\lambda) = \mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T]$ be the effective degrees of freedom

Discussion

- What is the difference?
  - Shrinkage and Subset Selection.
- Why?

Discussion



- Other penalties: Elastic Net

FYI

In fact, there are many other properties about LASSO in Statistics. More information could be seen in https://zhuanlan.zhihu.com/p/53764089.

Section 5

Supplementary

Subsection 1

Matrix Derivative Computation Examples

Tricks

- Matrix Derivative is more than important in many areas related to data science, such as SVM and RNN.
- Therefore, some tricks of computation could aid a lot in computing gradient and Hessian.

Theorem 12

If $f : \mathbb{R}^n \to \mathbb{R}, C^1$, then $\mathrm{D} f(x)[v] = v^T \nabla f(x)$, where D means the directional derivative.

Theorem 13

If $f : \mathbb{R}^n \to \mathbb{R}, C^2$, then $\nabla^2 f(x) \cdot v = D(\nabla f(x))[v]$, where D means the directional derivative.

A Concrete Example

Example

Find the gradient and Hessian of $f(x) = \frac{1}{2}x'Ax$.

Solution (Part 1)

Firstly, note that $\forall v$, we have

$$\begin{aligned} \mathrm{D}f(x)[v] = v^T \nabla f(x) &= \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t} \\ &= \frac{1}{2} v^T Ax + \frac{1}{2} x^T Av = v^T Ax \end{aligned}$$

So we have $\nabla f(x) = Ax$.

A Concrete Example

Solution (Part 2)

Secondly, Note that $\forall v$, we have

$$\begin{aligned}\nabla^2 f(x) \cdot v &= \mathrm{D}\,\nabla f(x)[v] \lim_{t \to 0} \frac{\nabla f(x+tv) - \nabla f(x)}{t} \\ &= Av\end{aligned}$$

So we have $\nabla^2 f(x) = A$

Now you could use this trick to easily find the gradient and Hessian of the Ridge Regression Objective.

Subsection 2

Properties of Projection Matrix

Properties of Projection Matrix

Definition 5: Projection Matrix

For a matrix H, if $H^2 = H$, then it is called a projection matrix.

- An example: hat matrix.

Theorem 14

If H is a projection matrix, so is $I - H$.

- Why this name?

Subsection 3

Unbiased Estimator of σ^2 in Multivariate OLS

Unbiased Estimator of σ^2 in Multivariate OLS

Theorem 12

$\hat{\sigma}^2 = \frac{SSE}{n-p-1}$ is the unbiased estimate of σ^2.

Definition 6: SSE

SSE is the sum of squares for error. The formal definition is $SSE = \mathbf{e}^T\mathbf{e}$.

Proof (Part 1)

Note that $e^Te = [(I-H)\mathbf{Y}]^T(I-H)\mathbf{Y} = \mathbf{Y}^T(I-H)\mathbf{Y}$

Proof (Part 2)

$$= (X\beta + \epsilon)^T (I - H)(X\beta + \epsilon) = \left(\beta^T X^T + \epsilon^T\right)(I - H)(X\beta + \epsilon)$$
$$= \beta^T X^T (I - H) X\beta + \beta^T X^T (I - H)\epsilon + \epsilon^T (I - H) X\beta + \epsilon^T (I - H)\epsilon$$

For $(I - H)\mathbf{X} = 0$. Taking expectation yields

$$E(\mathbf{e}^T\mathbf{e}) = E\left(\epsilon^T (I - H)\epsilon\right) = E\left(\text{tr}\left(\epsilon^T (I - H)\epsilon\right)\right)$$
$$= \text{tr}\left(E\left((I - H)\epsilon\epsilon^T\right)\right) = \text{tr}\left((I - H)E\left(\epsilon\epsilon^T\right)\right)$$
$$= \sigma^2 \,\text{tr}(I - H) = \sigma^2(n - p - 1), \square$$

This result could be specialized for Univariate OLS.

Thank you!