Section 4

Models with penalty

带罚项的模型

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.

$X^T X$ 不满秩

$\hat{\beta}$ 不能计算

Ridge Regression

Definition 4: Ridge Regression Estimator

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

$\hat{y}_i$

罚项

Ridge Regression Objective

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

Obviously, $\underline{\lambda \sum_{j=1}^{p} \beta_j^2}$ 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. $\square$

Ridge Regression is one kind of Shrinkage Methods.

Why Shrinkage?

收缩

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

Theorem 10

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

Proof

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

Prerequisite: Reduced-Rank Singular Value Decomposition

(低秩SVD)

- 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 $\Sigma_p = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$, 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 $\left(\mathbf{X}'\mathbf{X} + kI\right)^{-1}\mathbf{X}'\mathbf{X}$. By Reduced-Rank SVD and note that $\left(\mathbf{X}'\mathbf{X} + kI\right)^{-1}\mathbf{X}'\mathbf{X} = U(\Sigma^2 + \lambda I)^{-1}\Sigma^2 V^T$, 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.$\square$

Why Shrinkage?

Compare two estimators

收缩因子

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

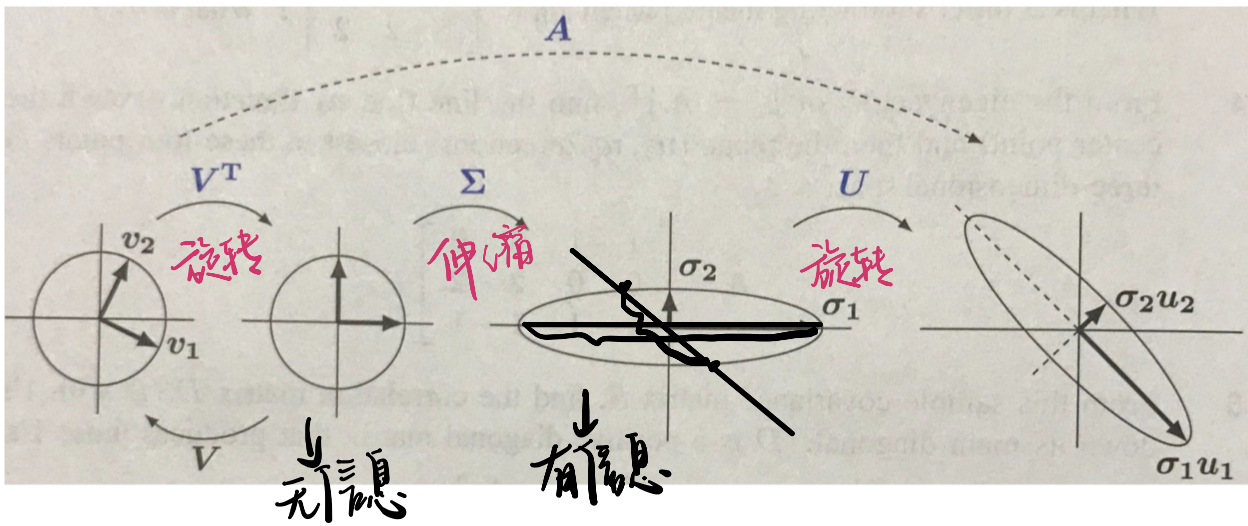
$$\hat{y}_2 = X\hat{\beta} = UU^T\mathbf{Y} = \sum_{j=1}^{p} u_j u_j^T \mathbf{Y}$$

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

Behavior

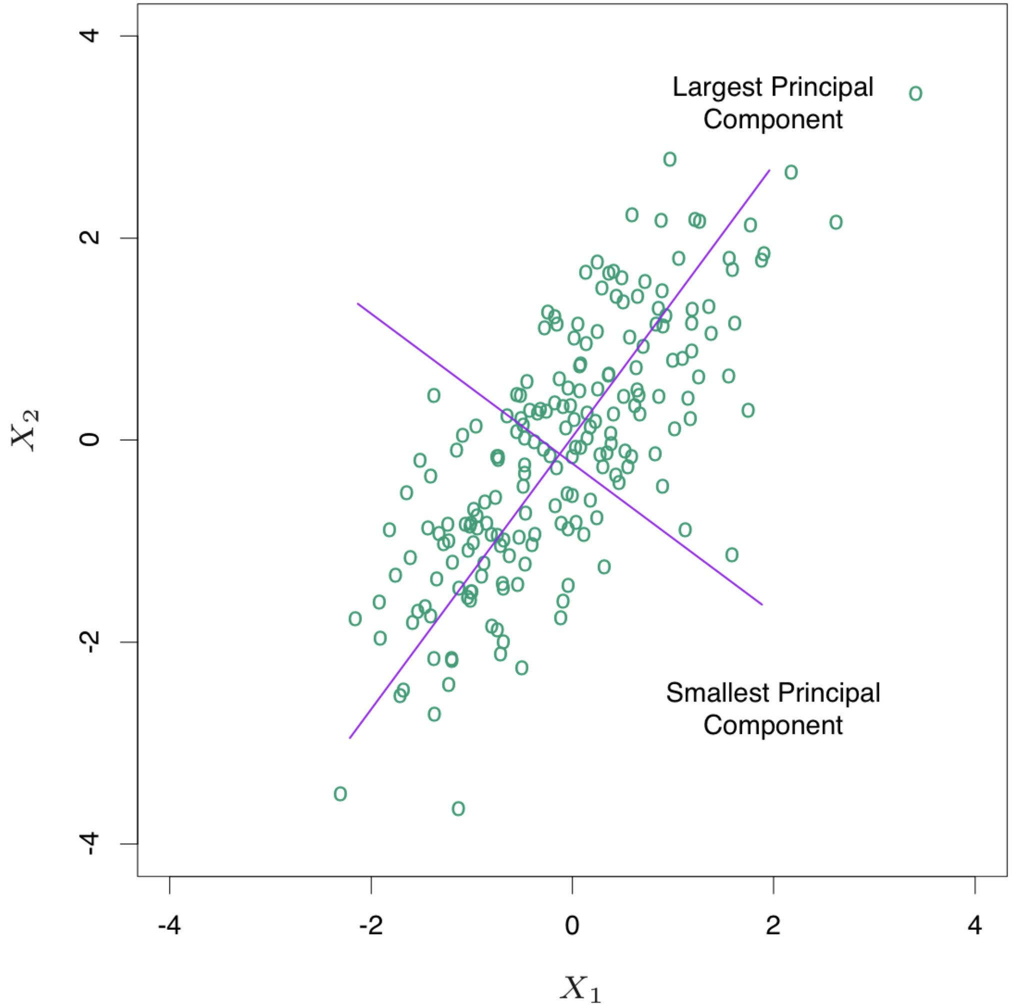
几何意义

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 $\mathbf{X}^T\mathbf{X} = V\Sigma^2V^T$ 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

套索回归

Introduction

- Could we change the penalty? Obviously could!
- From 2-norm to 1-norm
  - Least Absolute Selection and Shrinkage Operator (LASSO) 最小绝对选择与收缩算子
  - Have some unexpected properties.

LASSO Objective

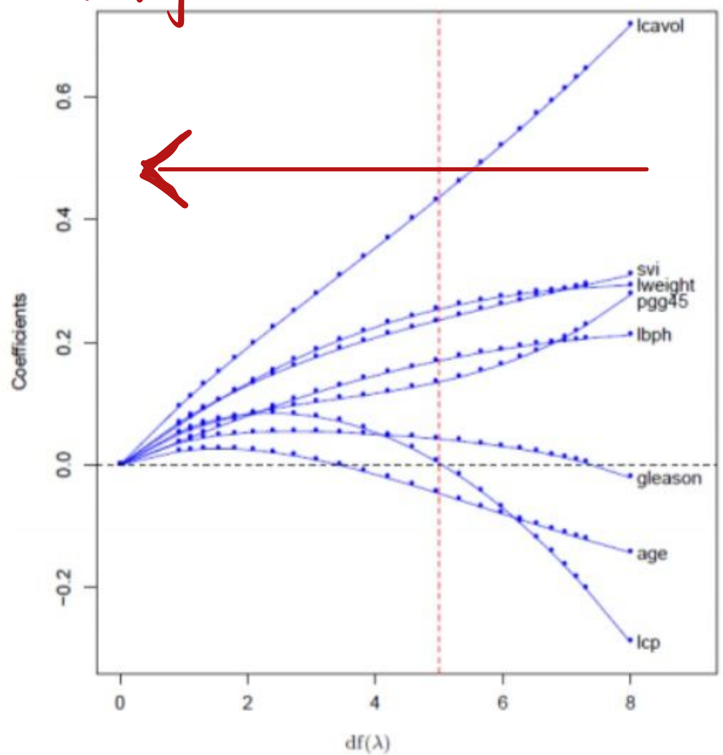
$$\hat{\beta} = \arg\min_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

罚项

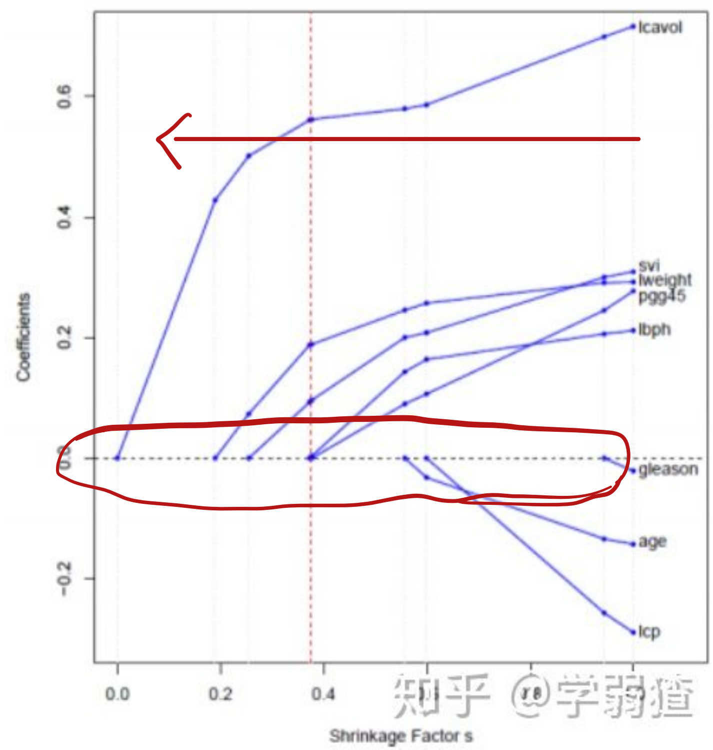
Comparison with Ridge Regression

容易碰到顶点

Ridge



LASSO



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 λ越大，自由度越小

Discussion

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

收缩　　子集选取

- Why?

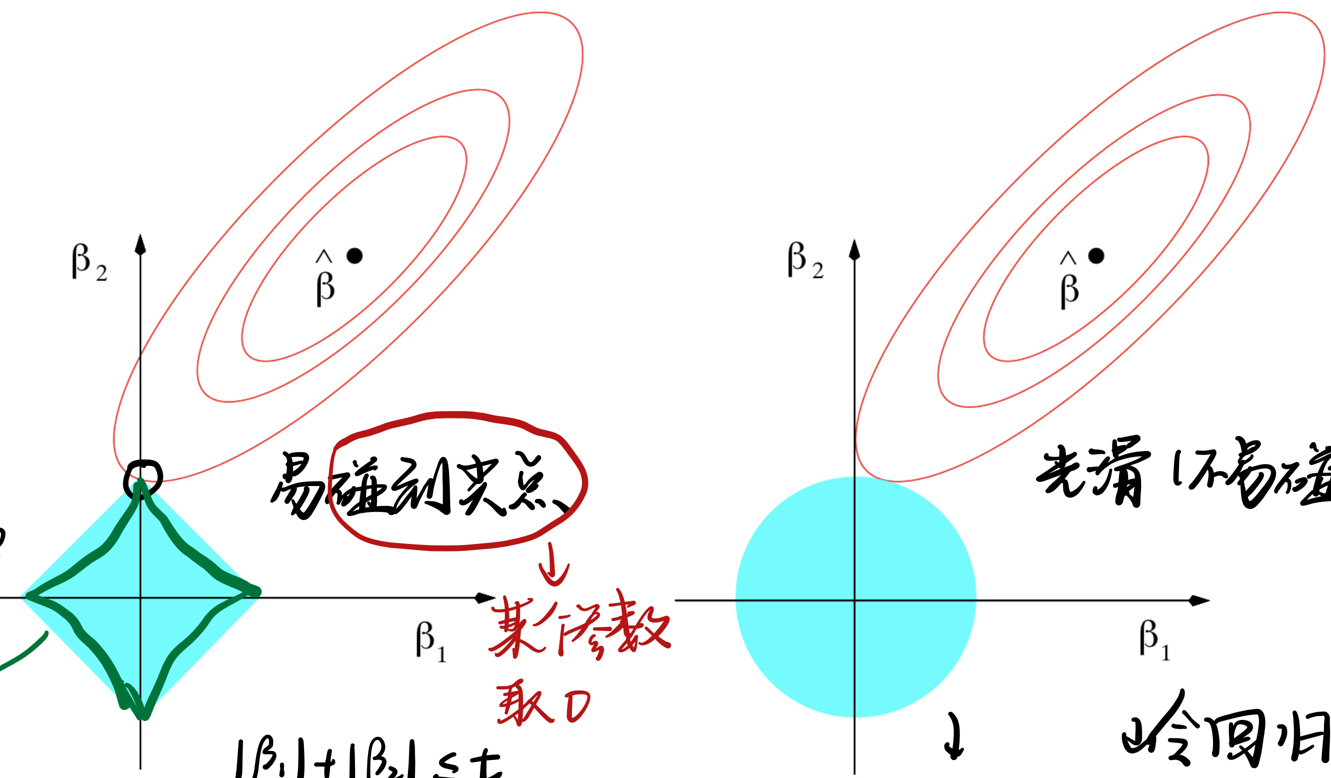
为了 部分系数不仅收缩，且会直接不存在

Discussion

$|\beta_j| \to$ LASSO

$\|\beta_j\|^2 \to$ Ridge

$\beta_j^{0.5}$



易碰到尖点

↓

某个参数取0

$|\beta_1| + |\beta_2| \le t$

光滑（不易碰到尖点）

↓

岭回归

$\sum \lambda \|\beta_j\|_2 \Rightarrow \beta_1^2 + \beta_2^2 \le t$

- Other penalties: Elastic Net $\sum \alpha |\beta_j| + (1-\alpha) \|\beta_j\|^2$

FYI

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