Introduction to Machine Learning CS182

Lu Sun

School of Information Science and Technology

ShanghaiTech University

October 17, 2023

Today:

- Linear Methods for Regression II
  - Ridge Regression
  - The Lasso
  - Discussion

Readings:

- The Elements of Statistical Learning (ESL), Chapter 3
- Pattern Recognition and Machine Learning (PRML), Chapter 3

Introduction

- Subset selection 扔掉不必要的子集
    - retain a subset of the predictors, and discard the rest
    - accuracy and interpretation
    - discrete process
        - variable are either retained or discarded
        - high variance
- Shrinkage methods 是连续的过程
    - continuous process 方差会小
        - don’t suffer much from high variability
    - ridge regression, lasso, …

Linear Methods for Regression

--- Ridge Regression

截距，因为惩罚β_0无意义，

正则化不包括截距项，故需先优化 β_0

岭回归

Shrinkage Methods – Ridge Regression

可写成矩阵：$\|Y-\beta_0-\beta X\|^2$ → 都是二范数；半正定矩阵

λ=∞时惩罚过大，故classifier简单即可

- Shrink the regression coefficients
  - impose a penalty on the size

RSS　　$\hat{y}_i$　　λ=0时直接RSS拟定

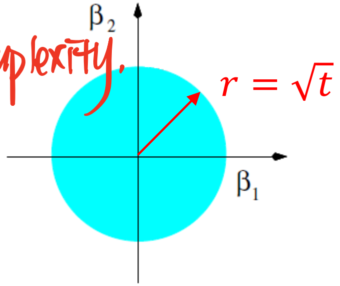
P1
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2}_{RSS} + \lambda \boxed{\sum_{j=1}^{p} \beta_j^2} \right\}$$

- Squared ℓ_2-norm on β
$$\|\beta\|_2^2 = \beta^T\beta = \sum_{j=1}^{p} \beta_j^2$$
- Other possible constraints?

  - the larger the value of λ, the greater the amount of shrinkage
  - the coefficients are shrunk toward zero
- An equivalent expression

P2
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2$$
$$\text{subject to } \boxed{\sum_{j=1}^{p} \beta_j^2} \le t,$$

t 控制 complexity



  - One-to-one correspondence between λ and t

Shrinkage Methods – Ridge Regression *

- Equivalence between P1 and P2

P1: $\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$

P2: $\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\beta\|_2^2, \ \text{s.t.} \|\beta\|_2^2 \le t$

- Goal: $\forall\lambda, \exists t \ge 0$: $\hat{\beta} = \tilde{\beta}$ (Step 1)
- $\forall t, \exists\lambda \ge 0$: $\hat{\beta} = \tilde{\beta}$ (Step 2)

Proof:

- Step 1: assume that P1 is solved

$$\boxed{-\boldsymbol{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\beta}) + \lambda\hat{\beta} = 0}$$

- Lagrange form of P2

$$L(\beta, \mu) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \mu(\|\beta\|_2^2 - t)$$

- KKT conditions
  1. $\nabla_\beta L(\tilde{\beta}, \tilde{\mu}) = 0$ $\Rightarrow$ $\boxed{-\boldsymbol{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\beta}) + \tilde{\mu}\tilde{\beta} = 0}$
  2. $\tilde{\mu}\left(\|\tilde{\beta}\|_2^2 - t\right) = 0$
  3. $\tilde{\mu} \ge 0$
  4. $\|\tilde{\beta}\|_2^2 \le t$

- Thus,
  - if

$$t = \|\hat{\beta}\|_2^2$$

  - Then

$$\tilde{\mu} = \lambda, \qquad \tilde{\beta} = \hat{\beta}$$

  - Satisfy the KKT conditions.

- Step 2: conversely, assume that P2 is solved
- The optimal solution $(\tilde{\beta}, \tilde{\mu})$ must satisfies KKT conditions. Therefore, let $\lambda = \tilde{\mu}$, we always have $\hat{\beta} = \tilde{\beta}$.

Strong duality holds for P2:

$(\tilde{\beta}, \tilde{\mu})$ is the optimal solution of P2 $\Leftrightarrow$ $(\tilde{\beta}, \tilde{\mu})$ satisfies KKT conditions

Shrinkage Methods – Ridge Regression

Important notes

- ridge solutions are not equivalent under scaling of inputs
  - *standardize* the inputs before solving it
- the intercept β_0 should be left out of the penalty term
  - Ex. 3.5 → once $x_{ij} - \bar{x}_j$, β_0 is estimated by $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ $\beta_0 = \bar{y}$
  - the rest parameters are estimated by the centered data   ↳把方去掉、
- Henceforth we assume the data has been standardized
  - $\mathbf{X}$ has p rather than $p+1$ columns

Standardization

$$x' = \frac{x - \bar{x}}{\sigma}$$

Training: preprocess: $x' = \frac{x - \bar{x}}{\sigma}$ (对每列, σ是列方差)

Prediction?

Centering: $y \to y - \bar{y}$ $\beta_0 \to \bar{y}$

train: $\min(\|y - X\beta\|^2 + \lambda\|\beta\|^2)$

Test: preprocess: $x_0 = \frac{x_0 - \bar{x}}{\sigma}$ predict: $\hat{y}_0 = \hat{\beta}^T X + \beta_0$

Shrinkage Methods – Ridge Regression

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \Big(y_i - \cancel{\beta_0} - \boxed{\sum_{j=1}^{p} x_{ij}\beta_j} \Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$x_i^T \beta$

- Ridge regression in matrix form

$$\hat{\beta}^{ridge} = \operatorname{argmin}_{\beta} \mathrm{PRSS}(\lambda, \beta) = \operatorname{argmin}_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

$\beta^T \beta$

- We can rewrite $\mathrm{PRSS}(\lambda, \beta)$ as follows

$$\mathrm{PRSS}(\lambda, \beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

$$= \mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda\beta^T\beta$$

- Differentiating $\mathrm{PRSS}(\lambda, \beta)$ w.r.t. β

$$\frac{\partial \mathrm{PRSS}(\lambda, \beta)}{\partial \beta} = -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)\beta = 0$$

- The closed form solution $\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \boxed{\lambda\mathbf{I}_p})^{-1}\mathbf{X}^T\mathbf{y}$

  - $\mathrm{rank}(\mathbf{I}_p) = p$
  - make the problem nonsingular, even if $\mathrm{rank}(\mathbf{X}) < p$

一定为非奇异

I 是 p×p 的矩阵.

SVD 分解之后对 X 加上非负的 λI_p 一定为非奇异.

Shrinkage Methods – Ridge Regression

$X^TX = (UDV^T)^T(UDV^T) = (VDU^T)(UDV^T) = VD^2V^T$ 则 $X^TX+\lambda I_p$

对角矩阵, $D^T=D$.

$= VD^2V^T + \lambda VV^T$

$= V(D^2+\lambda I_p)V^T$

一定是非奇异的.

Additional insight into ridge regression

- Singular value decomposition (SVD)

$\mathbf{U}^T\mathbf{U} = \mathbf{I}_p, \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$

对角矩阵.

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

  - $\mathbf{U} \in \mathbb{R}^{N\times p}$: its columns span the column space ($\mathbb{R}^N$) of $\mathbf{X}$
  - $\mathbf{V} \in \mathbb{R}^{p\times p}$: its columns span the row space ($\mathbb{R}^p$) of $\mathbf{X}$
  - $\mathbf{D} \in \mathbb{R}^{p\times p}$: diagonal matrix $(d_1 \geq d_2 \geq \cdots \geq d_p \geq 0)$

- Singular values of $\mathbf{X}$
- if $\exists d_j = 0$, $\mathbf{X}$ is singular

Least squares

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ &= \mathbf{U}\mathbf{U}^T\mathbf{y},\\ &= \sum_{j=1}^{p} \boxed{\mathbf{u}_j}\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

The j-th column of $\mathbf{U}$

Ridge regression

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}\\ &= \mathbf{U}\,\mathbf{D}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\,\mathbf{U}^T\mathbf{y}\\ &= \sum_{j=1}^{p}\mathbf{u}_j\boxed{\frac{d_j^2}{d_j^2+\lambda}}\mathbf{u}_j^T\mathbf{y},\end{aligned}$$

- shrinkage factor
- smaller d_j leads to a larger shrinkage

Shrinkage Methods – Ridge Regression

- Prostate cancer example
  - #training(N) = 67, #testing=30
  - #variables(p)=8 8维
  - ridge coefficient estimates
- *Effective degree of freedom*

$$\mathrm{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \in (0, p]$$

有效自由度 ⇒ 复杂度

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$$

$$\begin{aligned}\mathrm{df}(\lambda) &= \mathrm{Tr}\left(\mathrm{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\right) \\ &= \mathrm{Tr}\left(\boxed{\mathbf{U}\mathbf{D}\left(\mathbf{D}^2 + \lambda\mathbf{I}_p\right)^{-1}\mathbf{D}\mathbf{U}^T}\right) \\ &= \boxed{\sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}}\end{aligned}$$

Trace equals to sum of eigenvalues

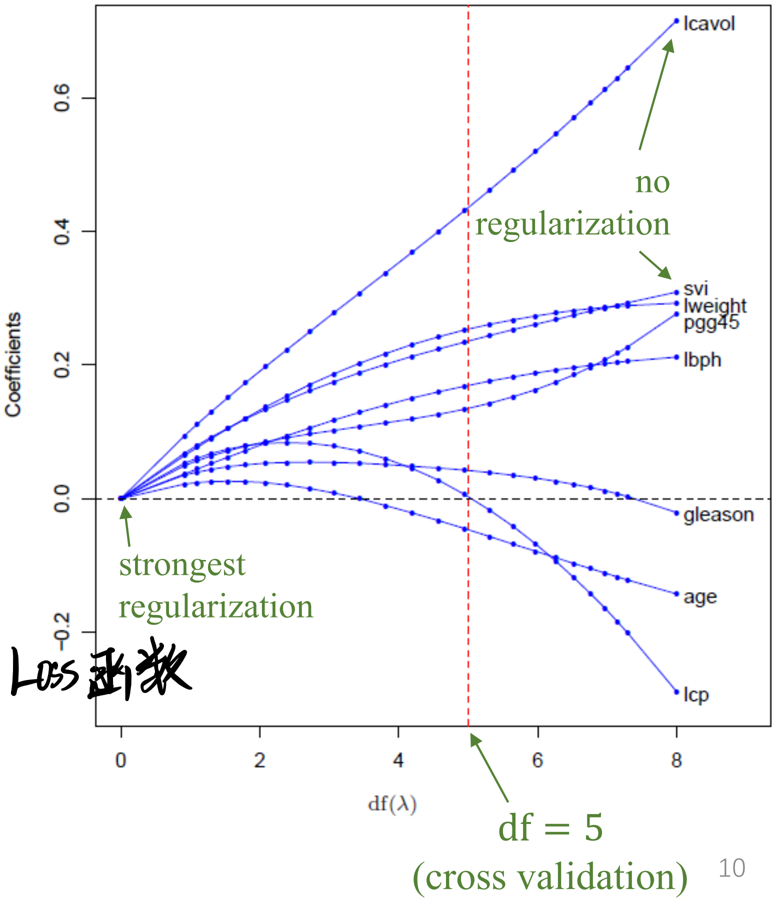
Shrinkage Methods – Ridge Regression

- Prostate cancer example
  - #training(N) = 67, #testing=30
  - #variables(p)=8
  - ridge coefficient estimates
- *Effective degree of freedom*

$$\mathrm{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \in (0, p]$$

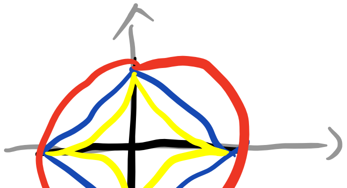
  - $\lambda \to 0, \mathrm{df}(\lambda) = p$ ← no regularization
  - $\lambda \to \infty, \mathrm{df}(\lambda) \to 0$

↳把自由度罚没了，不再关心Loss函数



Linear Methods for Regression

--- The Lasso



$\|\cdot\|_2$ 二范数 (convex)

$\|\cdot\|_1$ 一范数 (convex) (距零范数最近的 convex 是一范数)

$\|\cdot\|_0$ 零范数 (非凸)

$\|\cdot\|_p$ 范数, $0<p<1$ (非凸)

稀疏化方法

Shrinkage Methods – The Lasso

→最理想(稀疏化),但极复杂.

→ $\|\beta\|_0$:非零元的个数

→从平方和 → 绝对值之和

- The lasso estimate:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}}\left\{\underbrace{\frac{1}{2}\sum_{i=1}^{N}\Big(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2}_{\text{training error}} + \lambda\underbrace{\boxed{\sum_{j=1}^{p}|\beta_j|}}_{\text{model complexity}}\right\}.$$

ℓ_1-norm on β

$$\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$$

  - the ℓ_2 ridge penalty is replaced by ℓ_1 lasso penalty.
  - no closed-form solution (ℓ_1 penalty is nondifferentiable)

  ↪在顶点处不可导,故用梯度下降逼近.

- Or equivalently,

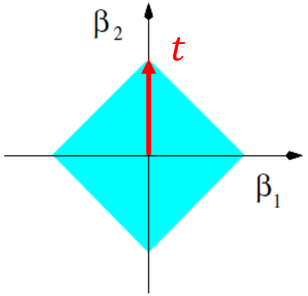
Constraint optimization

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}}\sum_{i=1}^{N}\Big(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2$$

$$\text{subject to } \sum_{j=1}^{p}|\beta_j| \le \boxed{t.}$$

- if $t \ge \|\hat{\beta}^{ls}\|_1$, $\hat{\beta}^{lasso} = \hat{\beta}^{ls}$
- if $t = \frac{1}{2}\|\hat{\beta}^{ls}\|_1$, $\hat{\beta}^{ls}$ is shrunk about 50% on average

  - making t sufficiently small → some coefficients equal to 0



Shrinkage Methods – The Lasso

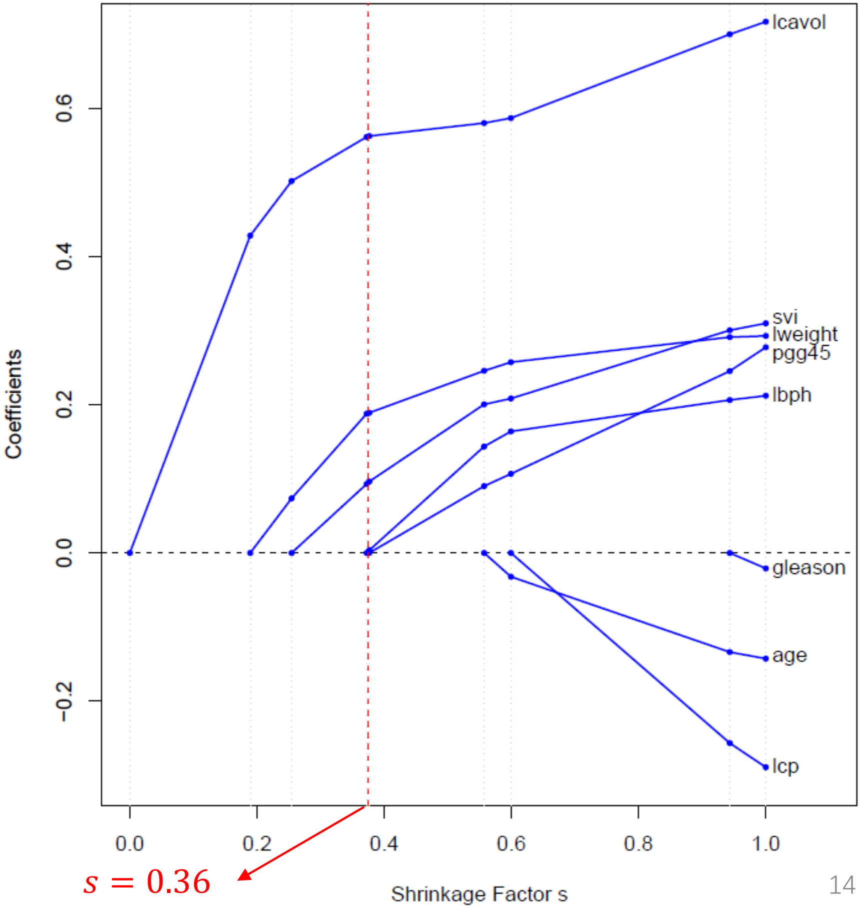
- The lasso in matrix form

$$\hat{\beta}^{lasso} = \text{argmin}_{\beta} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

- Prostate cancer example
- The standardized parameter

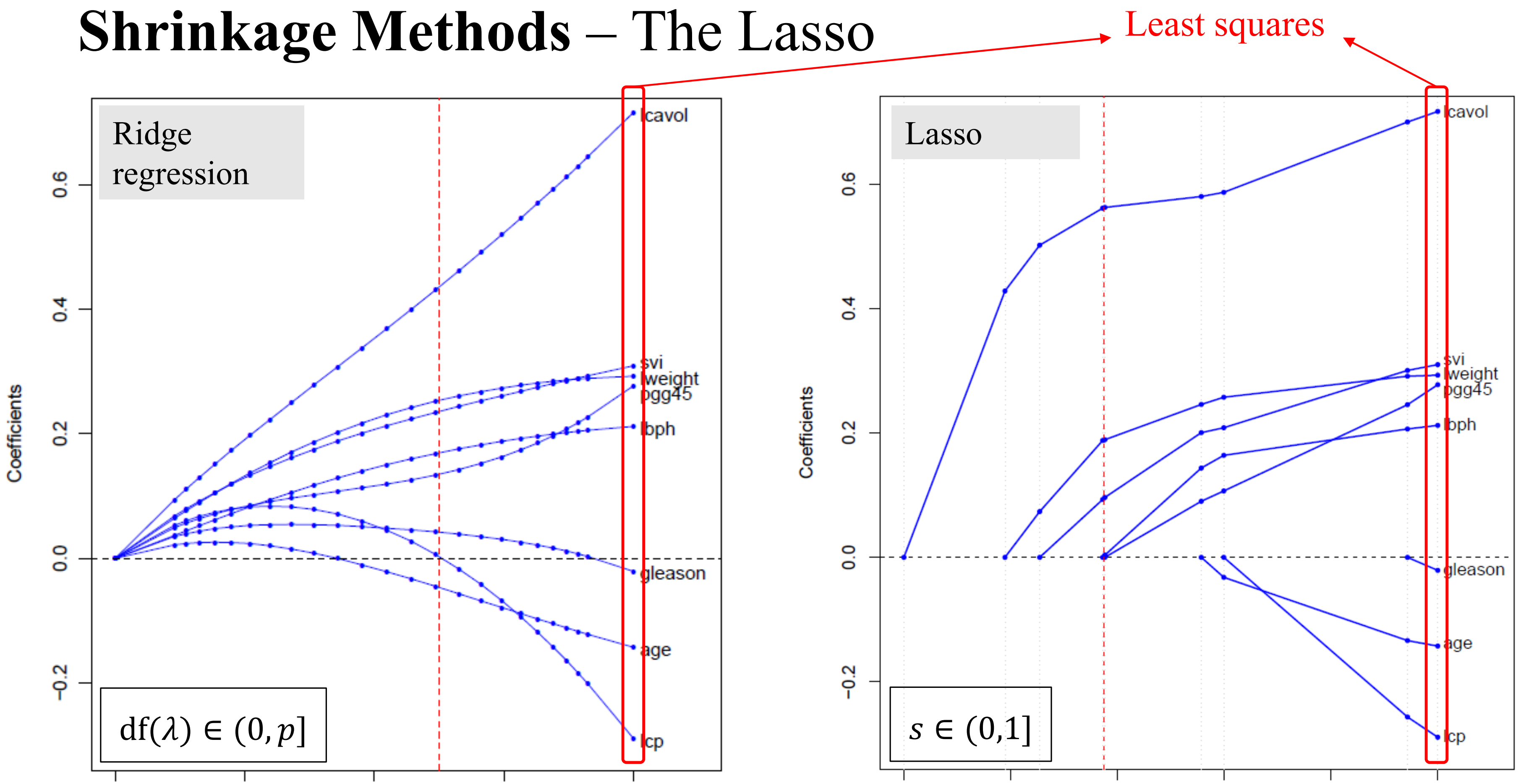
$$s = t / \left\| \hat{\beta}^{ls} \right\|_1 \in (0,1]$$

  - $s = 1, \hat{\beta}^{lasso} = \hat{\beta}^{ls}$
  - $s \to 0, \hat{\beta}^{lasso} \to 0$
  - $s \in (0,1), \hat{\beta}_j^{lasso} \in \left(0, \hat{\beta}_j^{ls}\right), \forall j$

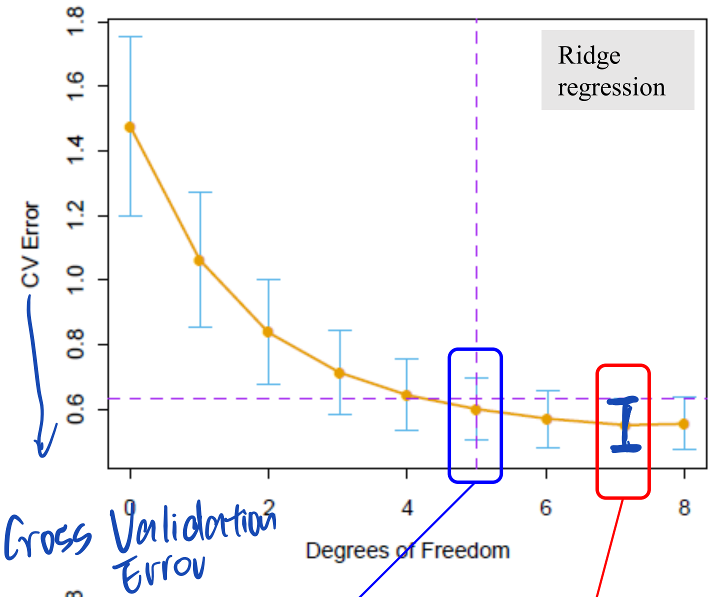


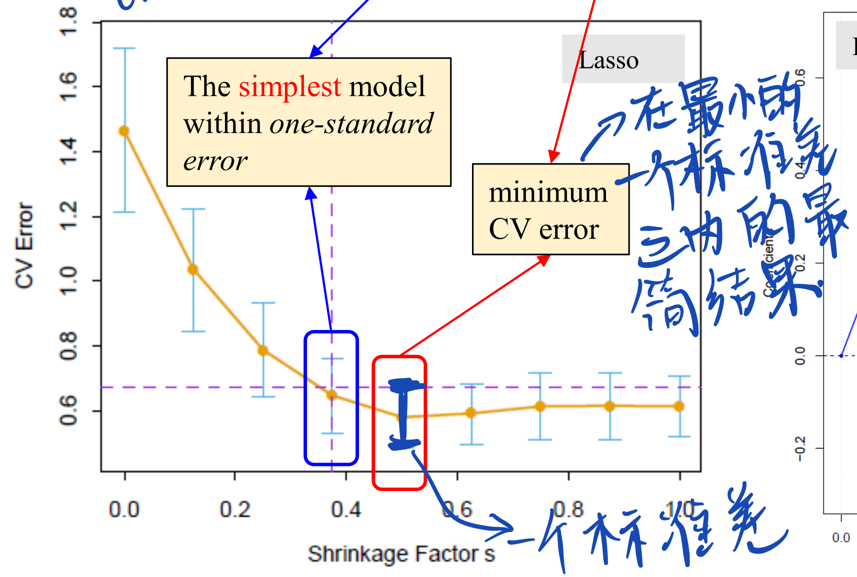
$s = 0.36$
selected by cross validation

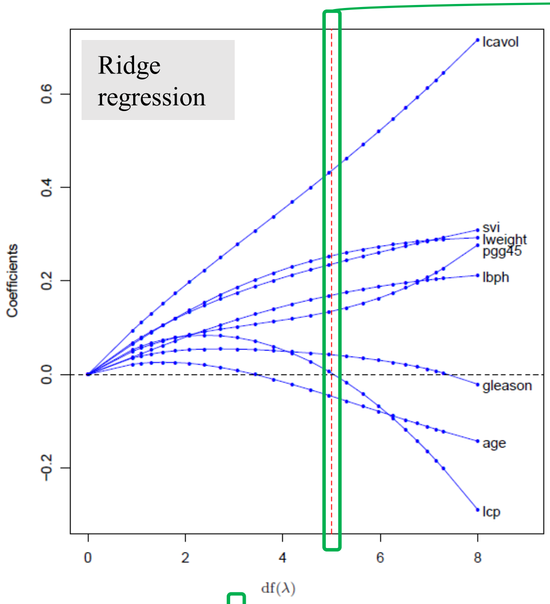
Shrinkage Methods – The Lasso

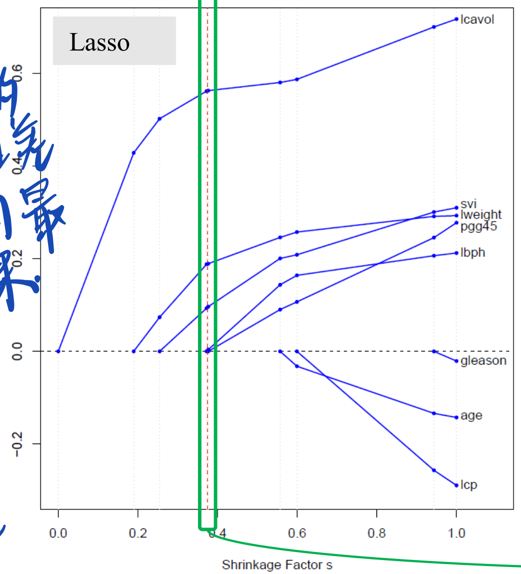


Difference: the lasso profiles hit zero, while those for ridge do not.





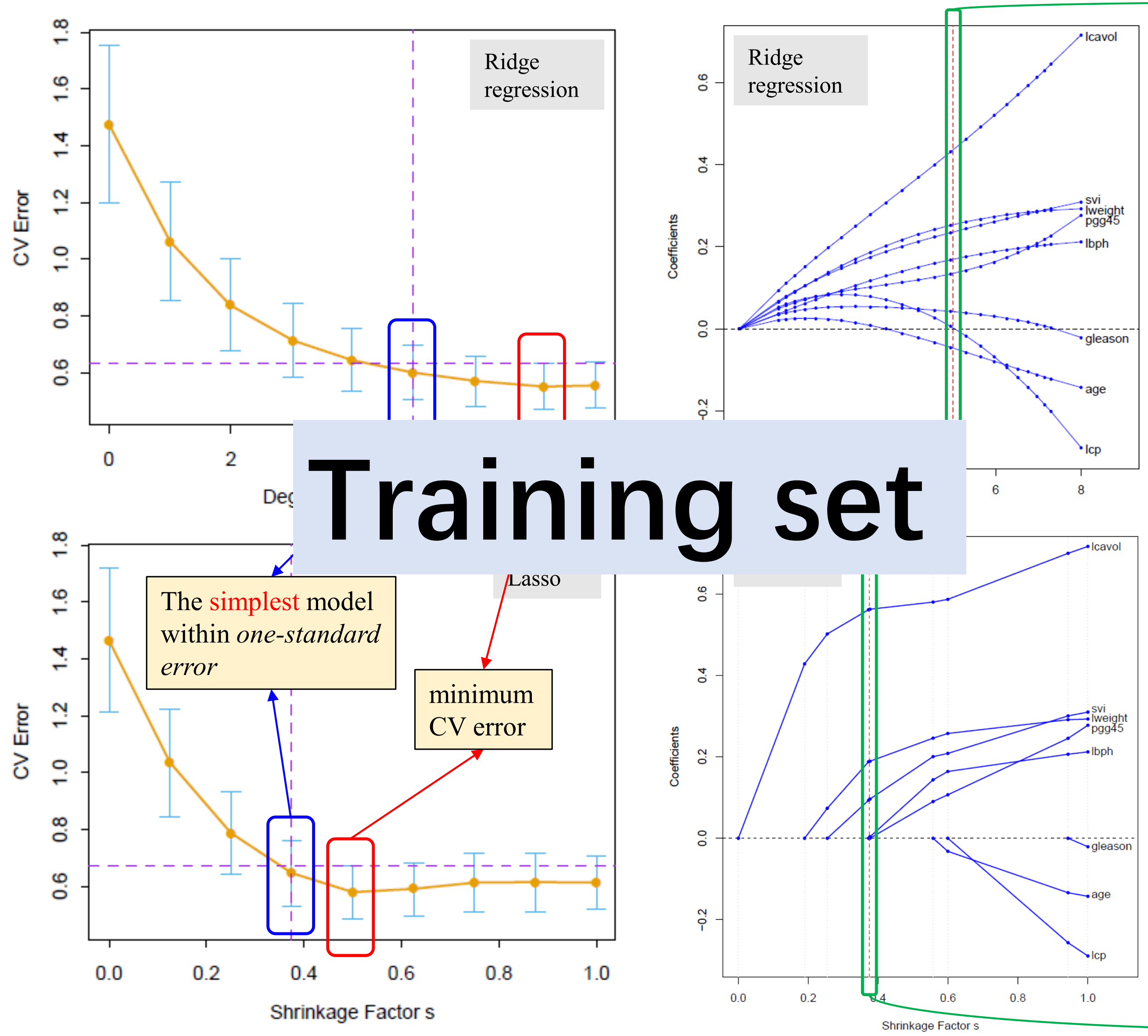




$\mathrm{df}(\lambda) = 5$

Term	LS	Ridge	Lasso
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
age	−0.141	−0.046	
lbph	0.210	0.162	0.002
svi	0.305	0.227	0.094
lcp	−0.288	0.000	
gleason	−0.021	0.040	
pgg45	0.267	0.133	
Test Error	0.521	0.492	0.479
Std Error	0.179	0.165	0.164

$s = 0.36$



Training set

Testing set

$\mathrm{df}(\lambda) = 5$

Term	LS	Ridge	Lasso
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
gleason	−0.021	0.040	
pgg45	0.267	0.133	
Test Error	0.521	0.492	0.479
Std Error	0.179	0.165	0.164

- Biased linear methods achieved a better var-bias trade-off
- CV is usually time-consuming
  - e.g. given $s \in [0.1: 0.1: 1]$, we need to train the lasso by $10 \times 10 = 100$ times in 10-fold CV.

Linear Methods for Regression

--- Discussion

Shrinkage Methods – Geometric Interpretation

Orthonormal case ($\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$)

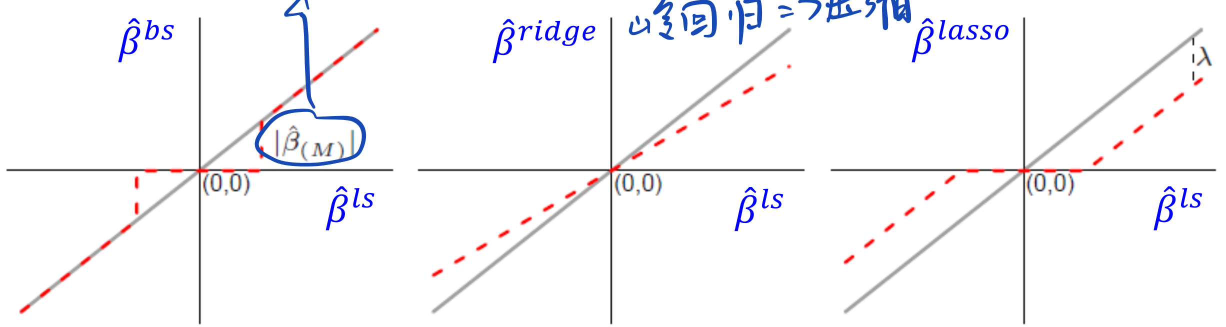
- Best-subset
  - hard-thresholding
  - discontinuity
- Ridge regression
  - proportional shrinkage
- Lasso
  - soft-thresholding

(M个子集)

有一个阈值

软阈值

岭回归 => 压缩



Estimator	Formula
Best subset (size M)	$\hat\beta_j \cdot I(\lvert\hat\beta_j\rvert \geq \lvert\hat\beta_{(M)}\rvert)$
Ridge	$\hat\beta_j/(1+\lambda)$
Lasso	$\text{sign}(\hat\beta_j)(\lvert\hat\beta_j\rvert - \lambda)_+$

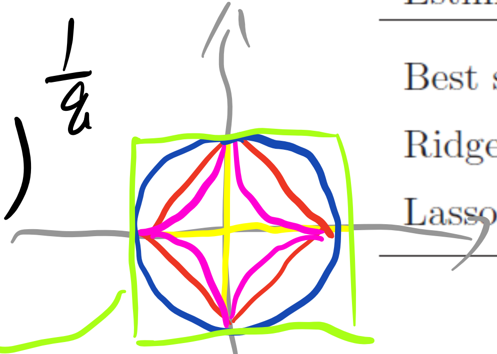
取β的符号

表示取+的部分

In this table $\hat\beta_j$ represents $\hat\beta_j^{ls}$

$$\|\beta\|_q = \left(\sum_{j=1}^{q} |\beta_j|^q\right)^{\frac{1}{q}}$$

$$\|\beta\|_\infty = \max_{j} |\beta_j|$$



若小于0则取0,

Shrinkage Methods – Geometric Interpretation

Orthonormal case ($\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$)

- Least squares

$$\hat{\beta}^{ls} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}$$

- Ridge regression 正则化

$$\hat{\beta}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y}$$

$$= \frac{1}{1+\lambda}\mathbf{X}^T\mathbf{y} = \frac{1}{1+\lambda}\hat{\beta}^{ls}$$

- Best subset

$$\hat{\beta}_j^{bs} = \mathbf{x}_j^T\mathbf{y}, \qquad \forall j$$

只需把 $\hat{\beta}_j^{bs}$ 排序 找前M个

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

- Lasso

$$\text{PRSS}(\beta,\lambda) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

常数,可扔

$$= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{y} + \frac{1}{2}\beta^T\mathbf{X}^T\mathbf{X}\beta + \lambda\|\beta\|_1$$

$$= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \beta^T\hat{\beta}^{ls} + \frac{1}{2}\beta^T\beta + \lambda\|\beta\|_1$$

- Minimizing $\text{PRSS}(\beta,\lambda)$ is equivalent to

$$\min_{\beta_j}\frac{1}{2}\beta_j^2 - \hat{\beta}_j^{ls}\beta_j + \lambda\,|\beta_j|, \qquad \forall j$$

需保证同号

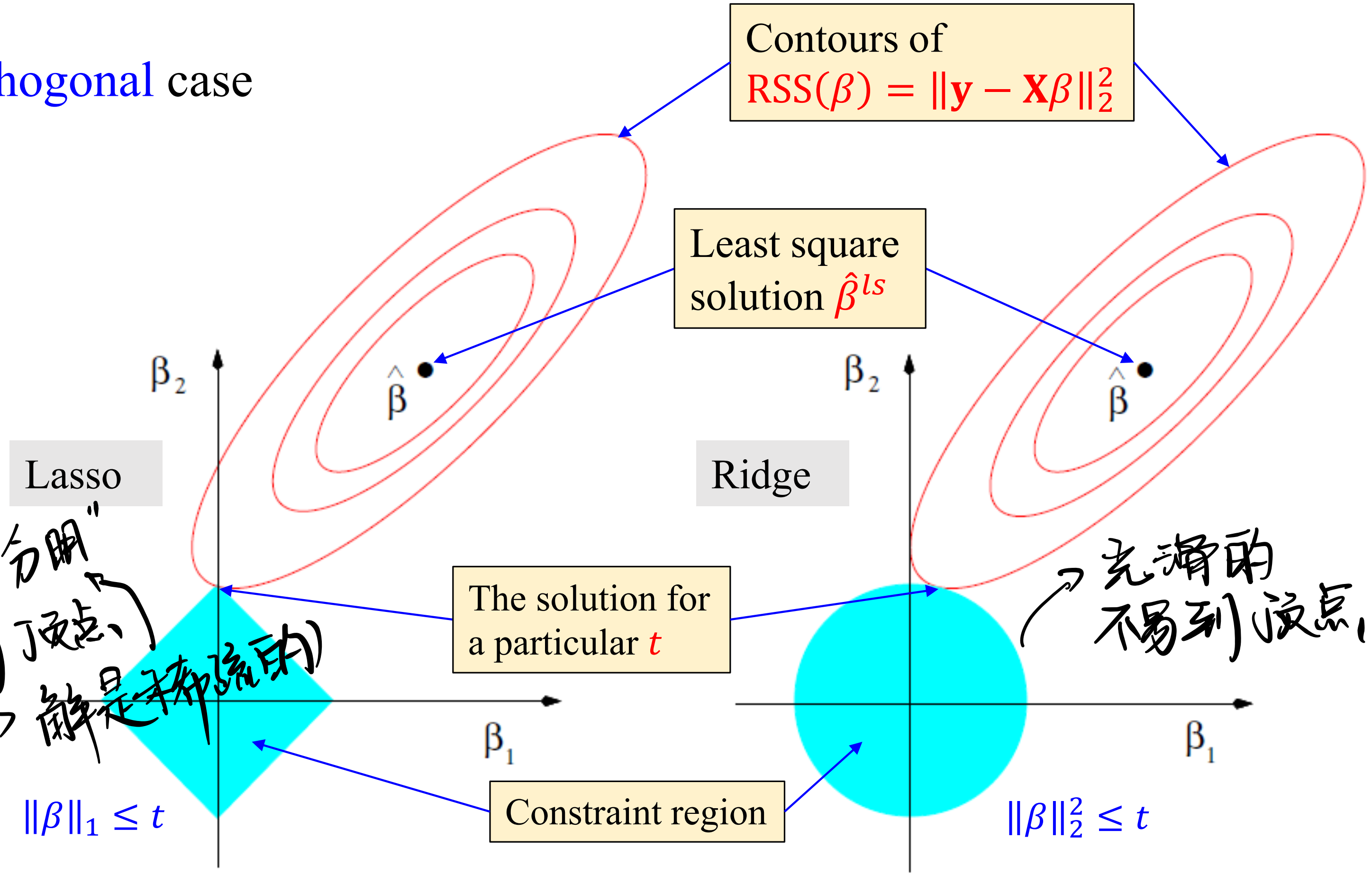
- Signs of $\hat{\beta}_j$ and $\hat{\beta}_j^{ls}$ must be the same.
  - $\hat{\beta}_j > 0 \to \hat{\beta}_j = \hat{\beta}_j^{ls} - \lambda$
  - $\hat{\beta}_j \leq 0 \to \hat{\beta}_j = \hat{\beta}_j^{ls} + \lambda$
- $\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j^{ls})\left(|\hat{\beta}_j^{ls}| - \lambda\right)_+$

Gradient.

proximal operator : 优化 $\min_{\beta} \|y-\beta\|_2 + \lambda\|\beta\|_1$ 用来优化 Lasso

Shrinkage Methods – Geometric Interpretation

Nonorthogonal case



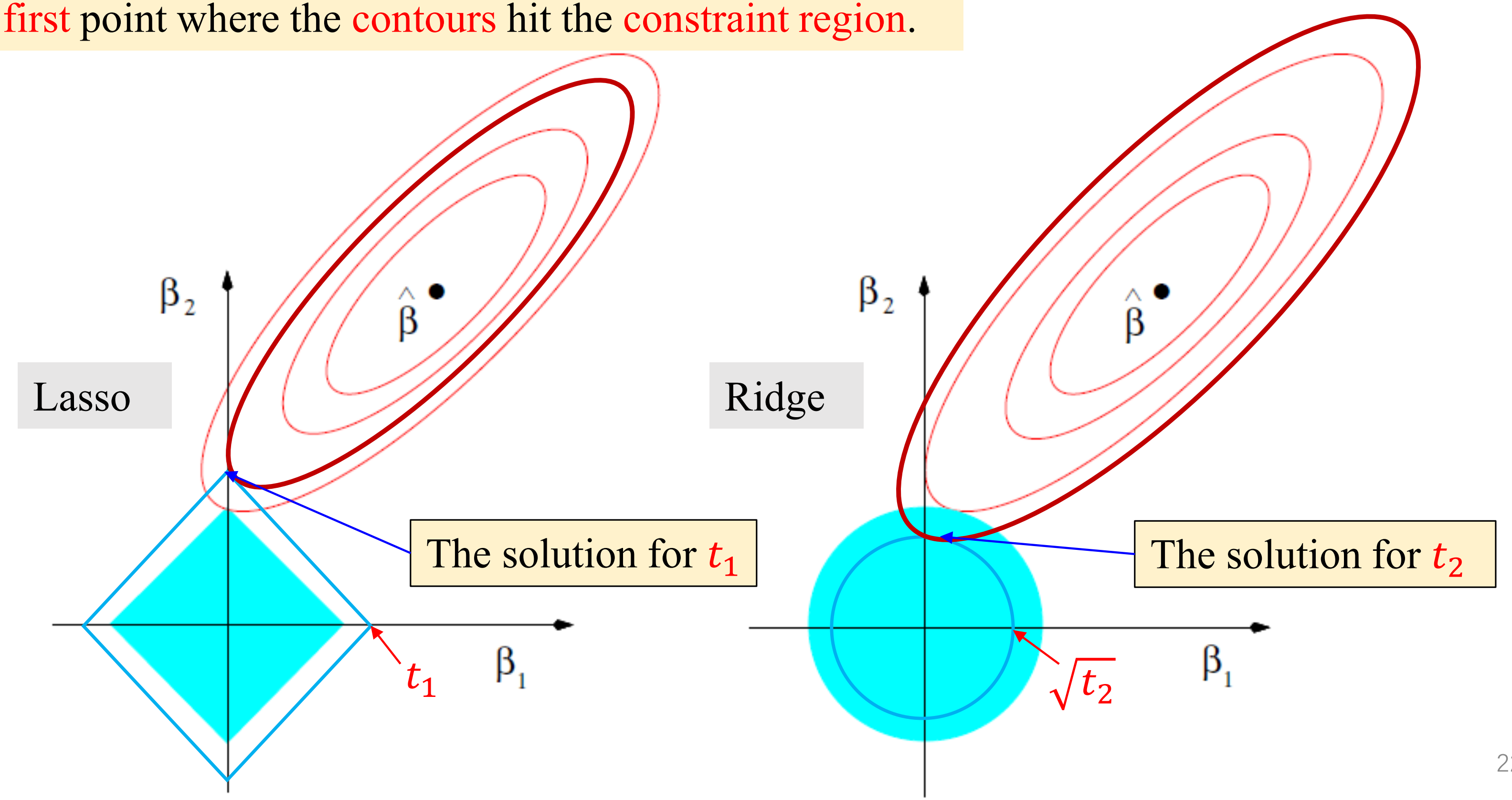
"棱角分明"

更易取到顶点，(某个 $\beta=0$ => 解是稀疏的)

光滑的 不易到顶点，

Shrinkage Methods – Geometric Interpretation

Lasso & Ridge regression:
Find the first point where the contours hit the constraint region.



Shrinkage Methods – Probabilistic Interpretation

Ridge and Lasso in the Bayes framework

- Suppose a Gaussian conditional distribution

$$\Pr(Y|X,\beta) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{Y-X^T\beta}{\sigma})^2)$$

$$\Pr(Y|X,\beta) = \mathcal{N}(X^T\beta, \sigma^2)$$

均值 方差.

- Log-likelihood

$$\begin{aligned}\ell(\beta) &= \ln \Pr(\mathbf{y}|\mathbf{X},\beta) \\ &= \sum_{i=1}^{N} \ln \Pr(y_i|x_i,\beta) \\ &= \underbrace{\boxed{-\frac{N}{2}\log(2\pi) - N\log\sigma}}_{\text{Constant}} - \boxed{\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2}\end{aligned}$$

MLE:

$$\begin{aligned}\hat{\beta}^{ls} &= \operatorname{argmax}_\beta \ell(\beta) \\ &= \operatorname{argmin}_\beta \|\mathbf{y}-\mathbf{X}\beta\|_2^2\end{aligned}$$

- Maximum a posterior (MAP)

$$\hat{\beta} = \operatorname{argmax}_\beta \underbrace{\boxed{\Pr(\beta|\mathbf{X},\mathbf{y})}}_{\text{Posterior}} = \operatorname{argmax}_\beta \frac{\underbrace{\boxed{\Pr(\mathbf{y}|\mathbf{X},\beta)}}_{\text{Likelihood}}\underbrace{\boxed{\Pr(\beta)}}_{\text{Prior}}}{\underbrace{\boxed{\Pr(\mathbf{X},\mathbf{y})}}_{\text{Irrelevant with }\beta}}$$

Posterior $\propto$ Likelihood $\times$ Prior

Shrinkage Methods – Probabilistic Interpretation

Ridge and Lasso in the Bayes framework

MLE: $\hat{\beta}^{MLE} = \text{argmax}_{\beta}\Pr(\mathbf{y}|\mathbf{X},\beta)$ ⟵ Least squares

MAP: $\hat{\beta}^{MAP} = \text{argmax}_{\beta}\Pr(\mathbf{y}|\mathbf{X},\beta)\Pr(\beta)$ ⟵ Ridge & Lasso

- Ridge regression
  - MAP with a prior $\Pr(\beta) = \mathcal{N}(\beta|0,\frac{1}{\lambda}\mathbf{I}_p)$ Gaussian distribution

$$\hat{\beta}^{ridge} = \text{argmax}_{\beta} \ln\big(\Pr(\mathbf{y}|\mathbf{X},\beta)\Pr(\beta)\big)$$
$$= \text{argmax}_{\beta} \ln\left(\prod_{i=1}^{N}\mathcal{N}(y_i|x_i^T\beta,\sigma^2)\times\mathcal{N}(\beta|0,\tfrac{1}{\lambda}\mathbf{I}_p)\right)$$

- Lasso
  - MAP with a prior $\Pr(\beta) = \frac{\lambda}{2}e^{-\lambda\|\beta\|_1}$ Laplacian distribution

$$\hat{\beta}^{lasso} = \text{argmax}_{\beta} \ln\left(\prod_{i=1}^{N}\mathcal{N}(y_i|x_i^T\beta,\sigma^2)\times\frac{\lambda}{2}e^{-\lambda\|\beta\|_1}\right)$$

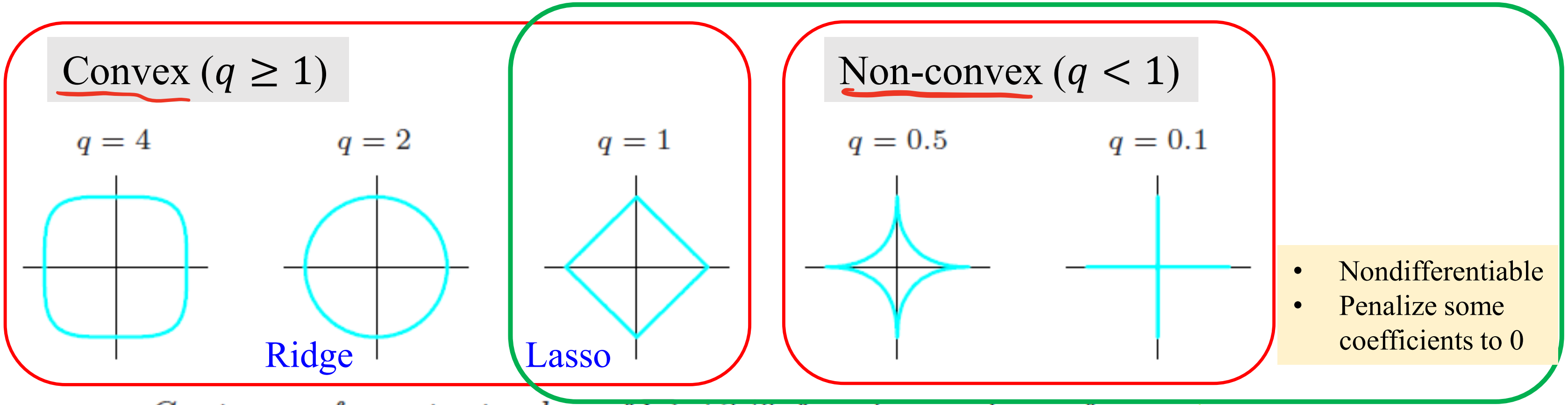
Shrinkage Methods – Generalization

Generalization of Ridge and Lasso

- Consider the criterion ($q \geq 0$)

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

- $q = 0$, best subset
- $q = 1$, lasso
- $q = 2$, ridge regression



Contours of constant value of $\sum_j |\beta_j|^q$ *for given values of* q.

若q<1则无距离上的意义 (如：欧氏距离：q=2)

因为不满足三角形不等式 $\|x+y\| \le \|x\|+\|y\|$

Shrinkage Methods – Generalization

岭回归：若相关则成一个组. Lasso：无关的置零.
相关的随机抽一个，剩下的置0.

Generalization of Ridge and Lasso

- Consider the criterion ($q \geq 0$)

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

- $q = 0$, best subset
- $q = 1$, lasso
- $q = 2$, ridge regression

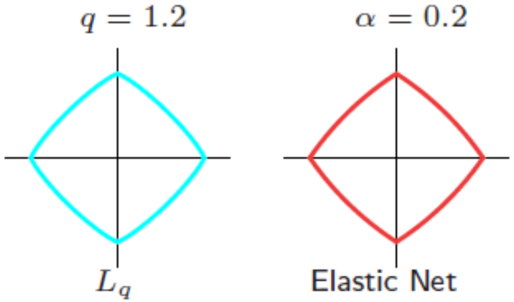
  - $q \in (1,2)$: a compromise between lasso and ridge regression
    - $|\beta_j|^q$ is differentiable at 0 $\rightarrow$ hard to set $\beta_j = 0, \forall j$

- Elastic-net

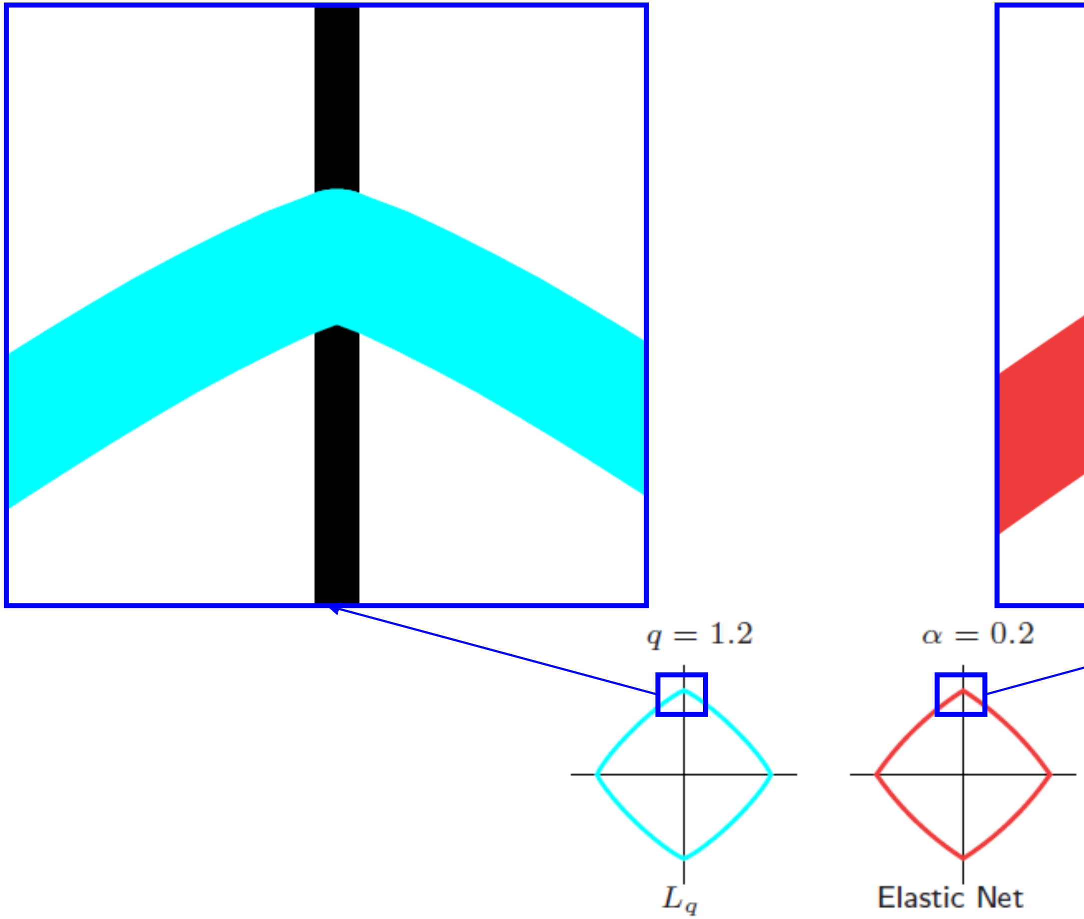
$\|\cdot\|_2^2 + \lambda \|\cdot\|_1$

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

  - ℓ_2 shrinks the coefficients of correlated predictors
  - ℓ_1 selects groups of correlated predictors



Shrinkage Methods – Generalization



The elastic-net has sharp (non-differentiable) corners