Optimization and Machine Learning SI151

Lu Sun

School of Information Science and Technology

ShanghaiTech University

March 4, 2021

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
  - 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
  - continuous process
    - don’t suffer much from high variability
  - ridge regression, lasso, …

Linear Methods for Regression

--- Ridge Regression

Shrinkage Methods – Ridge Regression

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

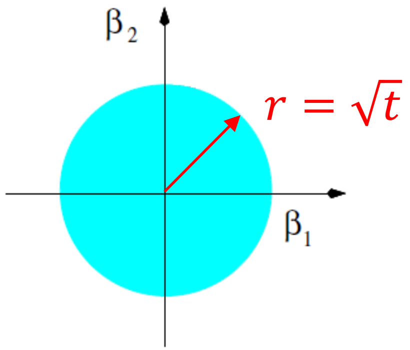
P1 $$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \boxed{\sum_{j=1}^{p}\beta_j^2}\right\}$$

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

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

  - One-to-one correspondence between λ and t

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



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: $\forall t, \exists \lambda \ge 0: \ \hat{\beta} = \tilde{\beta}$

Proof:

- Step 1: assume that P1 is solved

$$\boxed{\mathbf{y} - \mathbf{X}\hat{\beta} + \lambda\hat{\beta} = 0}$$

- Lagrange form of P2

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

- KKT conditions
  1. $\nabla_\beta L(\tilde{\beta}, \tilde{\mu}) = 0 \Longrightarrow \boxed{\mathbf{y} - \mathbf{X}\tilde{\beta} + \tilde{\mu}\tilde{\beta} = 0}$
  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 $\Longleftrightarrow$ $(\tilde{\beta}, \tilde{\mu})$ 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 $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
  - the rest parameters are estimated by the centered data
- Henceforth we assume the data has been standardized
  - $\mathbf{X}$ has p rather than $p+1$ columns

Standardization

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

Prediction?

Shrinkage Methods – Ridge Regression

$x_i^T\beta$

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

- Ridge regression in matrix form

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

$\beta^T\beta$

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

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

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

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

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

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

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

Shrinkage Methods – Ridge Regression

Additional insight into ridge regression

- Singular value decomposition (SVD)

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

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

  - $\mathbf{U} \in \mathbb{R}^{N\times p}$: 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

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

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

Q

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
  - ridge coefficient estimates
- *Effective degree of freedom*

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

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

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

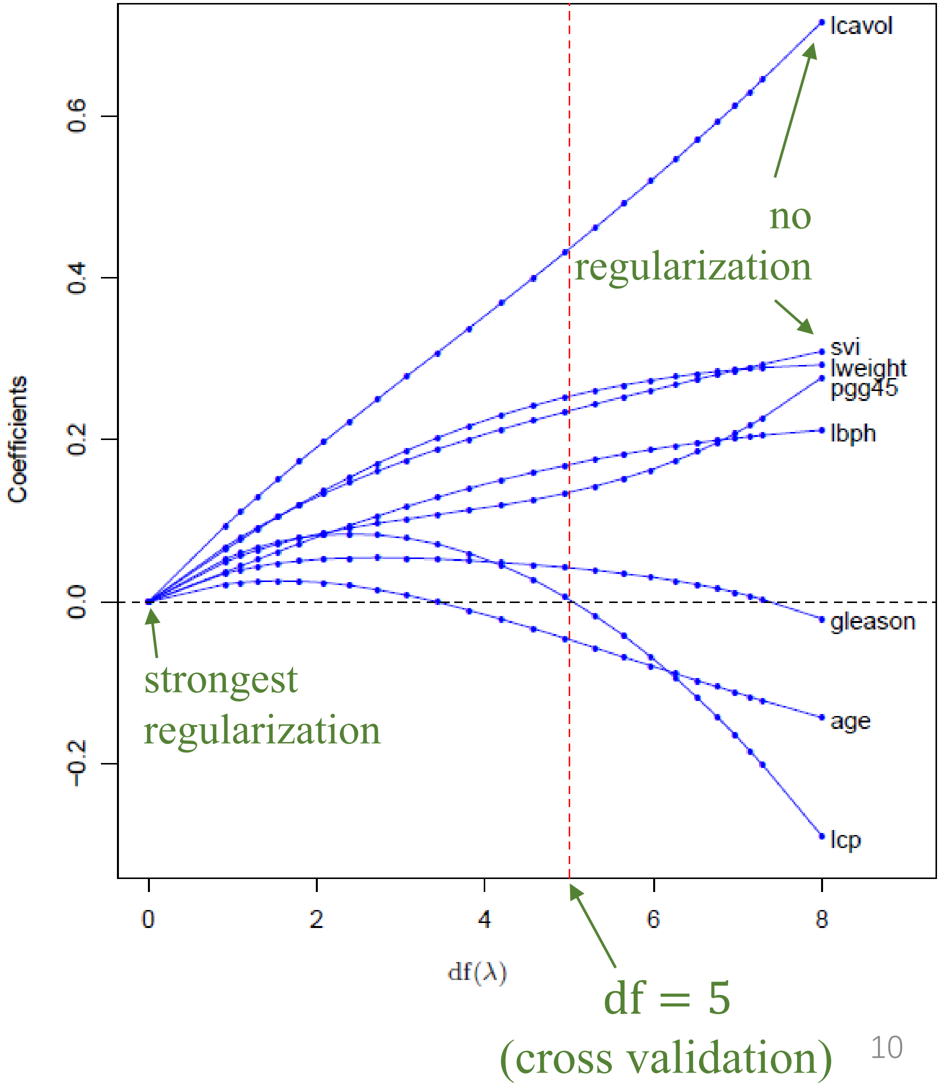
Trace equals to sum of eigenvalues

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
  - $\lambda \to \infty, \mathrm{df}(\lambda) \to 0$



Shrinkage Methods – Ridge Regression

- Principal components in $\mathbf{X}$
- Sample covariance

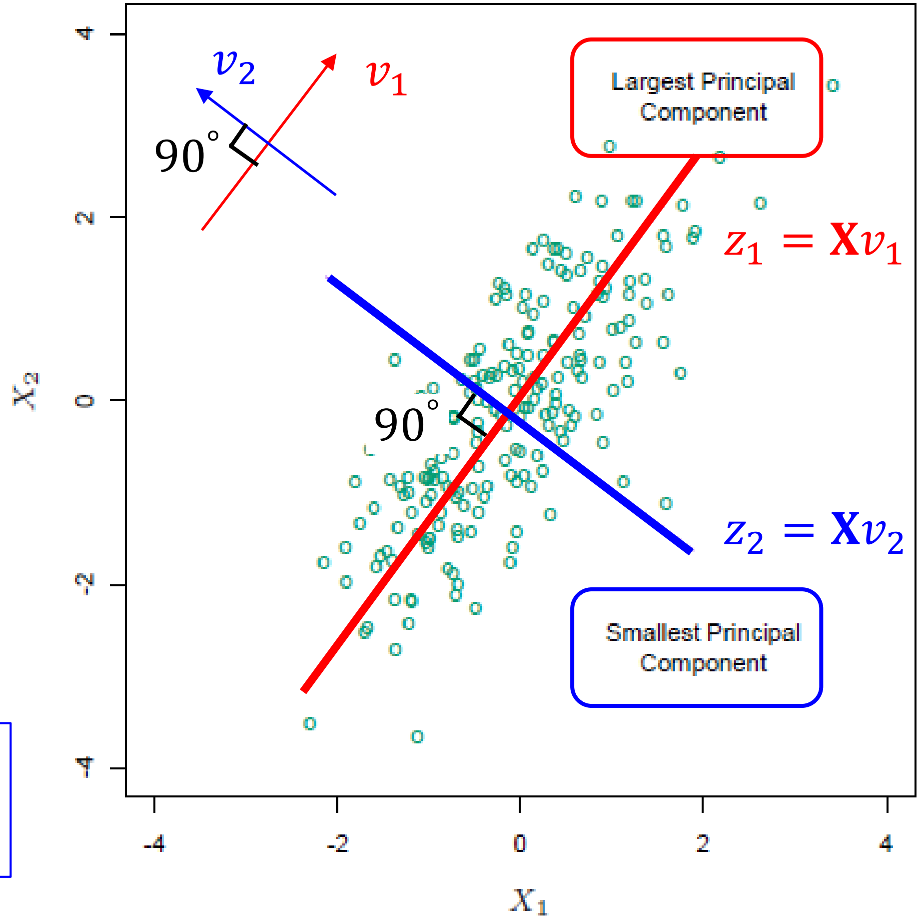
$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$

$$\mathbf{S} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X} = \frac{1}{N-1}\mathbf{V}\mathbf{D}^2\mathbf{V}^T$$

- *Eigen decomposition* of $\mathbf{X}^T\mathbf{X}$
- The eigenvector $\boxed{v_j}$ → The j-th column of $\mathbf{V}$
  - principal components directions of $\mathbf{X}$
  - $z_1 = \mathbf{X}v_1$: the first principal component

$$\begin{aligned} \mathrm{Var}(z_j) &= \mathrm{Var}(\mathbf{X}v_j) \\ &= \mathrm{Var}(\boldsymbol{u}_j d_j) \\ &= \frac{d_j^2}{N} u_j^T u_j \\ &= \frac{d_j^2}{N} \end{aligned}$$

- z_1 has the largest variance
- z_p has the smallest variance



shrinks the coefficients of the low-variance components more than the high-variance components.

Linear Methods for Regression

--- The Lasso

Shrinkage Methods – The Lasso

- The lasso estimate:

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

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)

- Or equivalently,

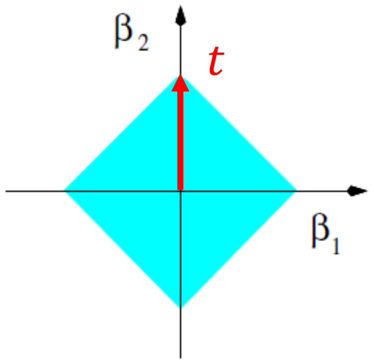
Constraint optimization

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

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

- if $t \geq \|\hat{\beta}^{ls}\|_1$, $\hat{\beta}^{lasso} = \hat{\beta}^{ls}$
- 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

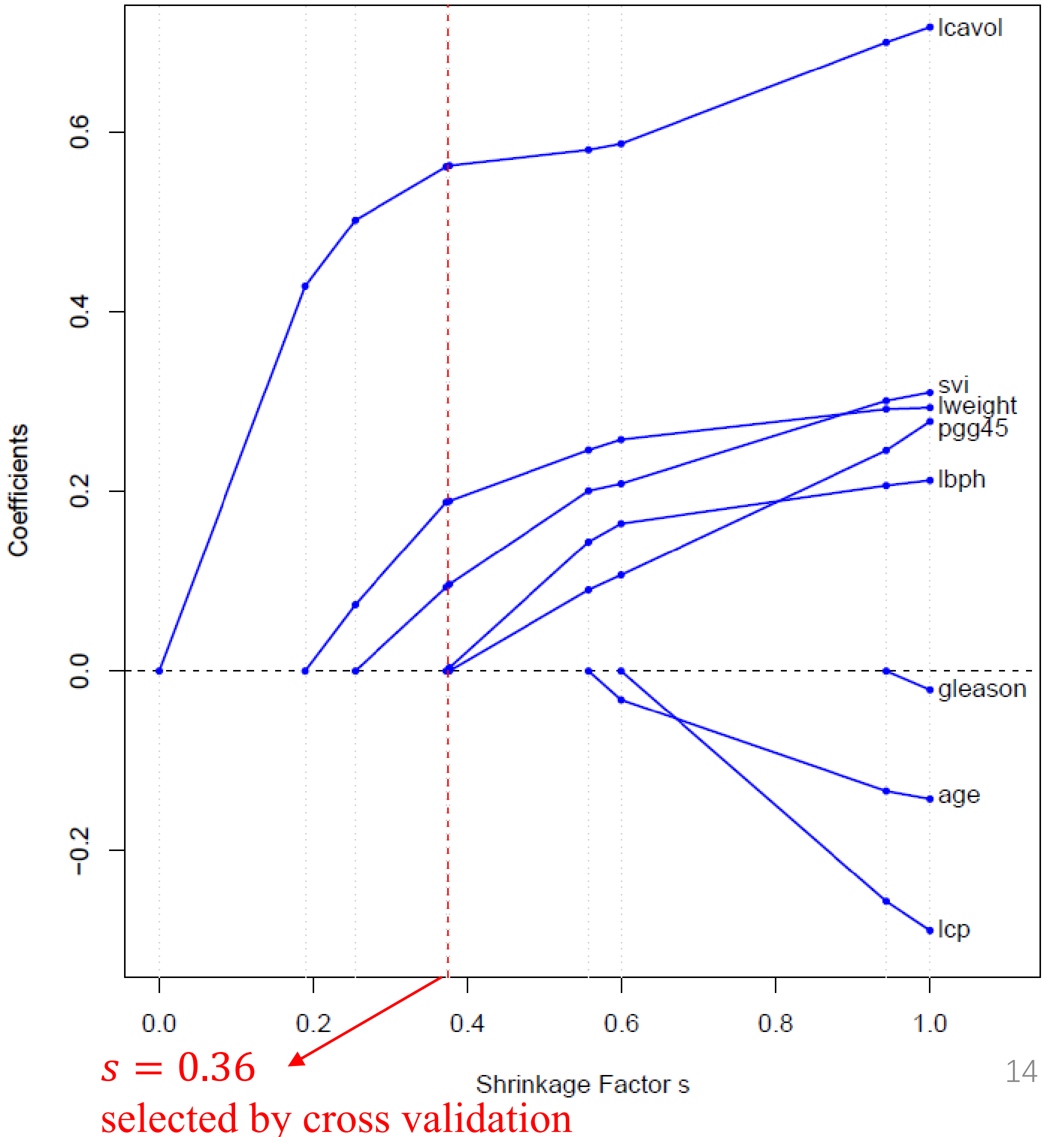
- The lasso in matrix form

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

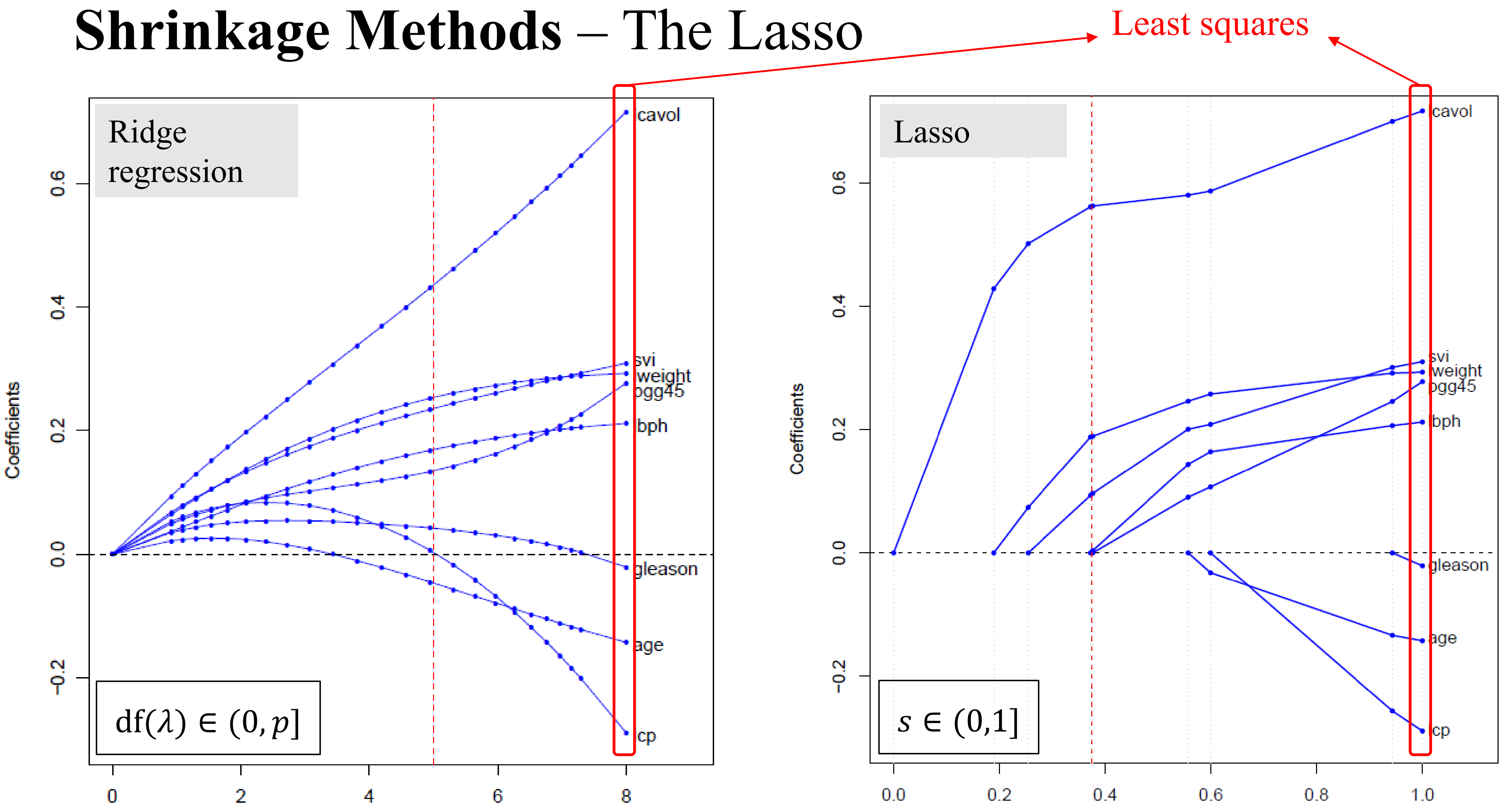
- Prostate cancer example
- The standardized parameter

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

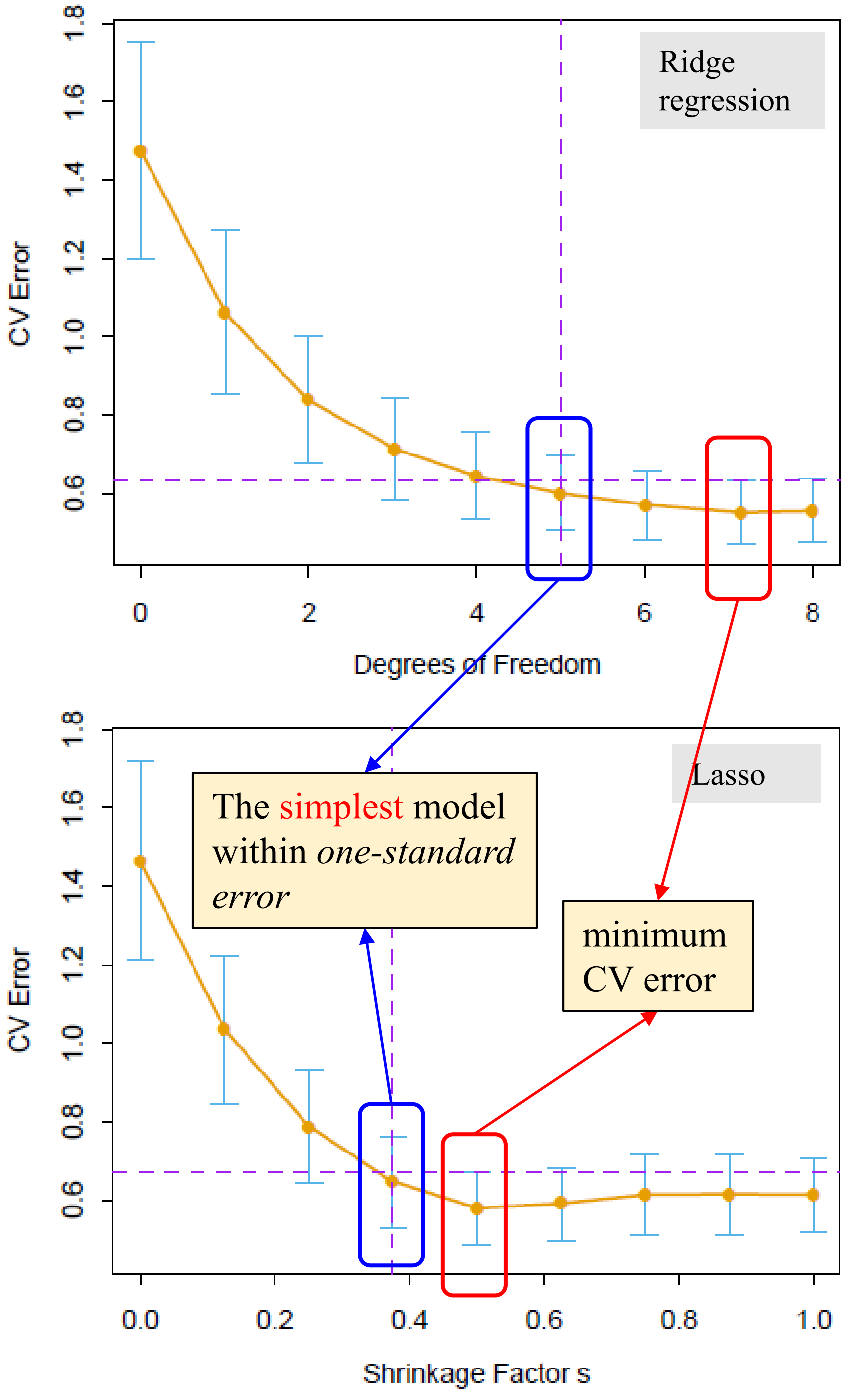
  - $s = 1, \hat{\beta}^{lasso} = \hat{\beta}^{ls}$
  - $s \to 0, \hat{\beta}^{lasso} \to 0$
  - $s \in (0,1), \hat{\beta}_j^{lasso} \in \left(0, \hat{\beta}_j^{ls}\right), \forall j$

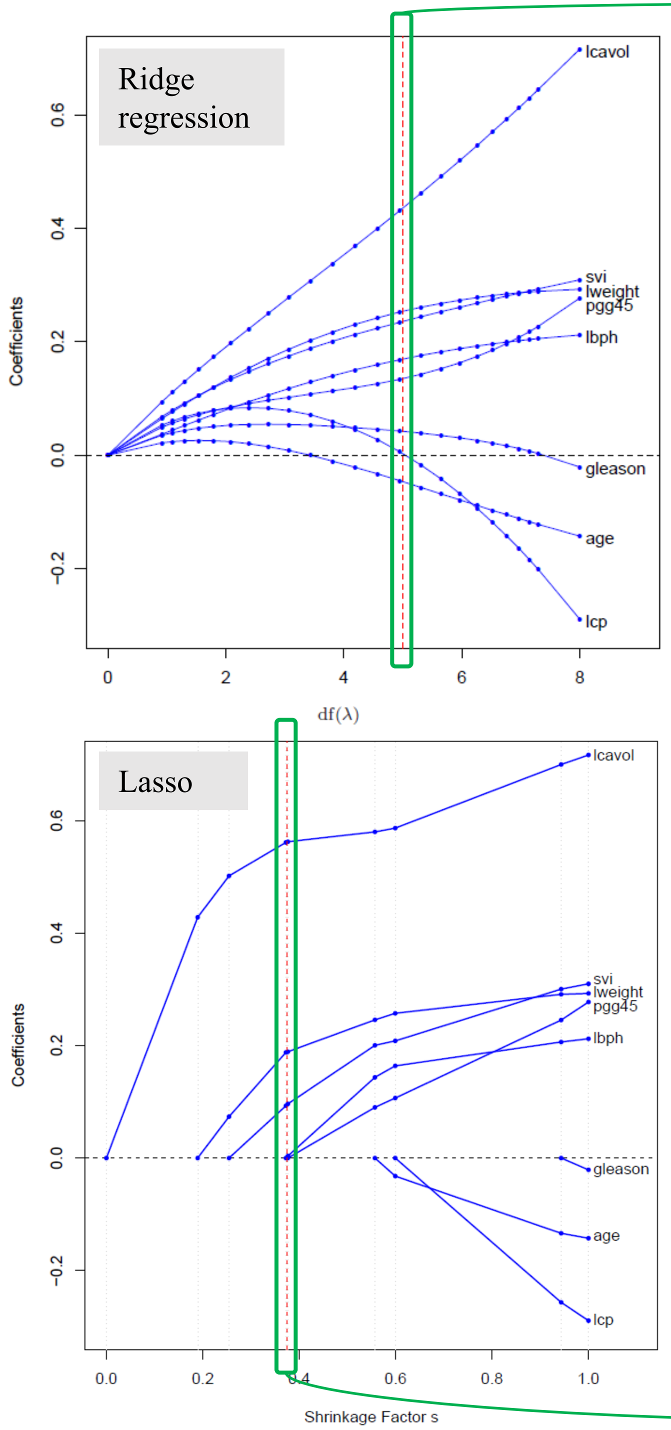


Shrinkage Methods – The Lasso



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

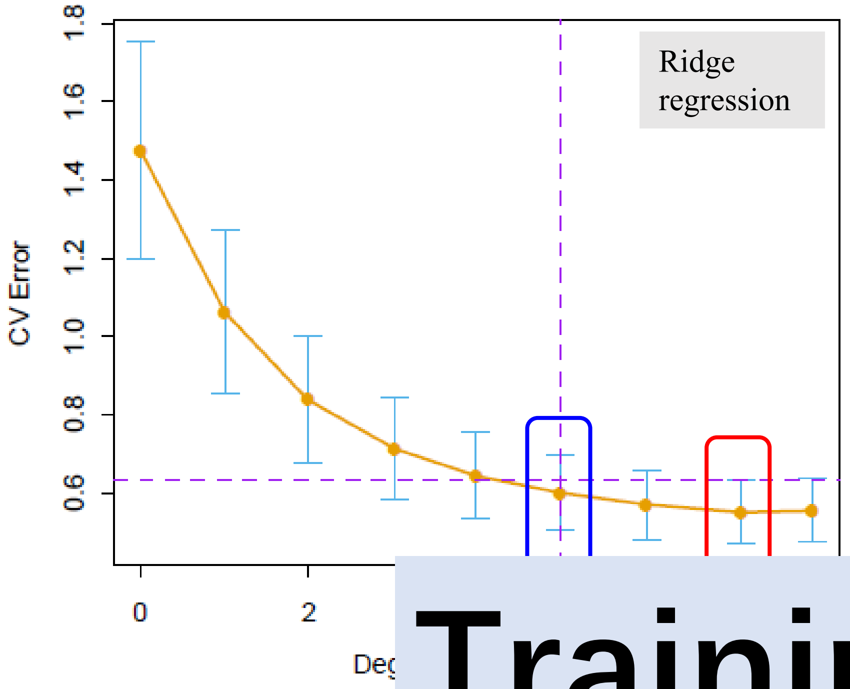


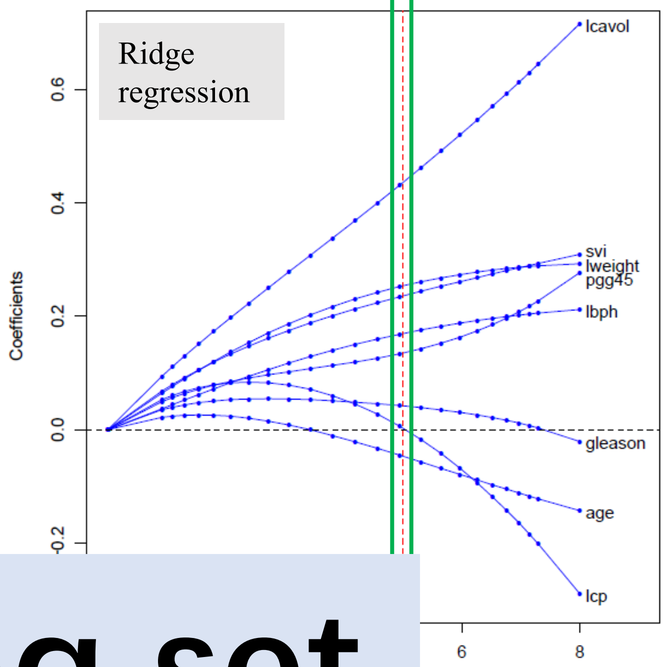


$\mathrm{df}(\lambda) = 5$

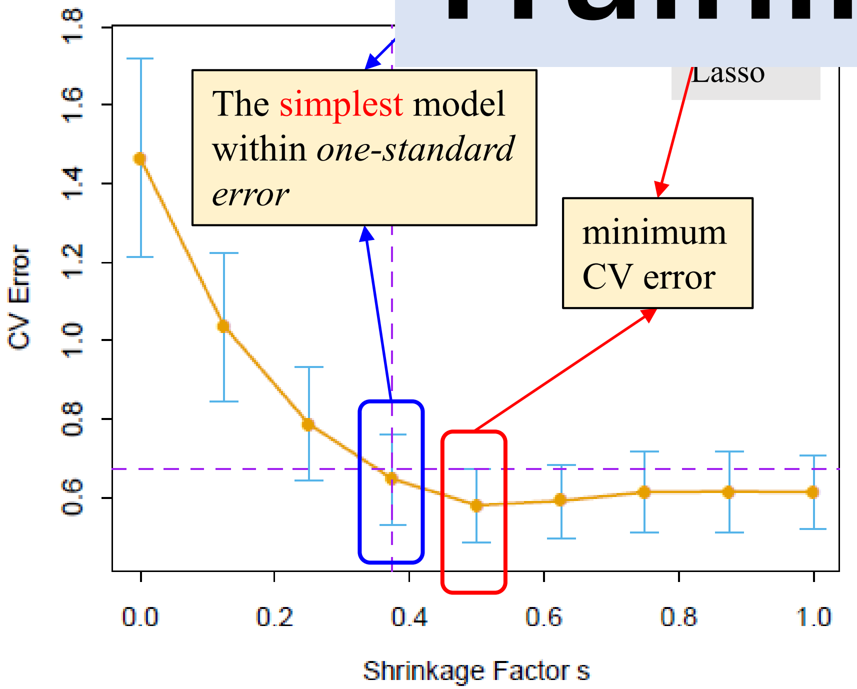
Term	LS	Ridge	Lasso
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
age	−0.141	−0.046	
lbph	0.210	0.162	0.002
svi	0.305	0.227	0.094
lcp	−0.288	0.000	
gleason	−0.021	0.040	
pgg45	0.267	0.133	
Test Error	0.521	0.492	0.479
Std Error	0.179	0.165	0.164

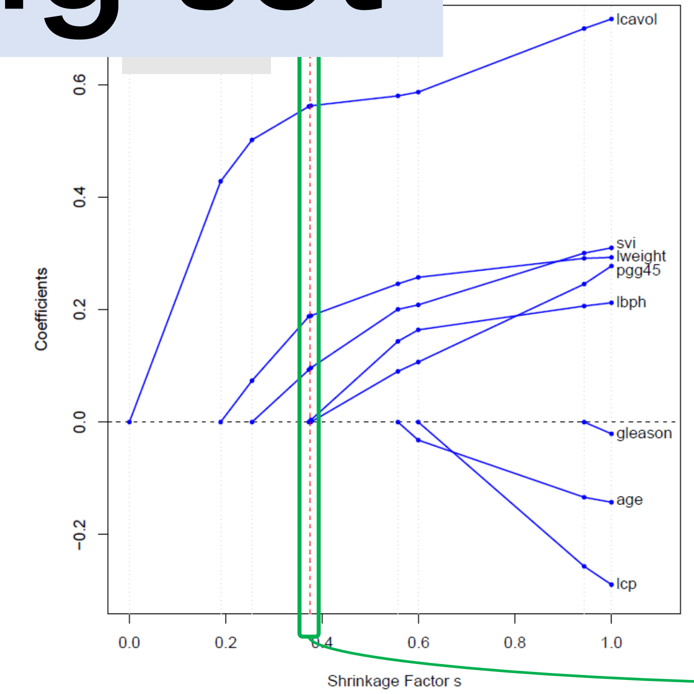
$s = 0.36$





Training set





Testing set

df(λ) = 5

Term	LS	Ridge	Lasso
lcavol	0.680	0.420	0.533
lweight	0.263	0.238	0.169
lcp	[illegible]	[illegible]	
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.

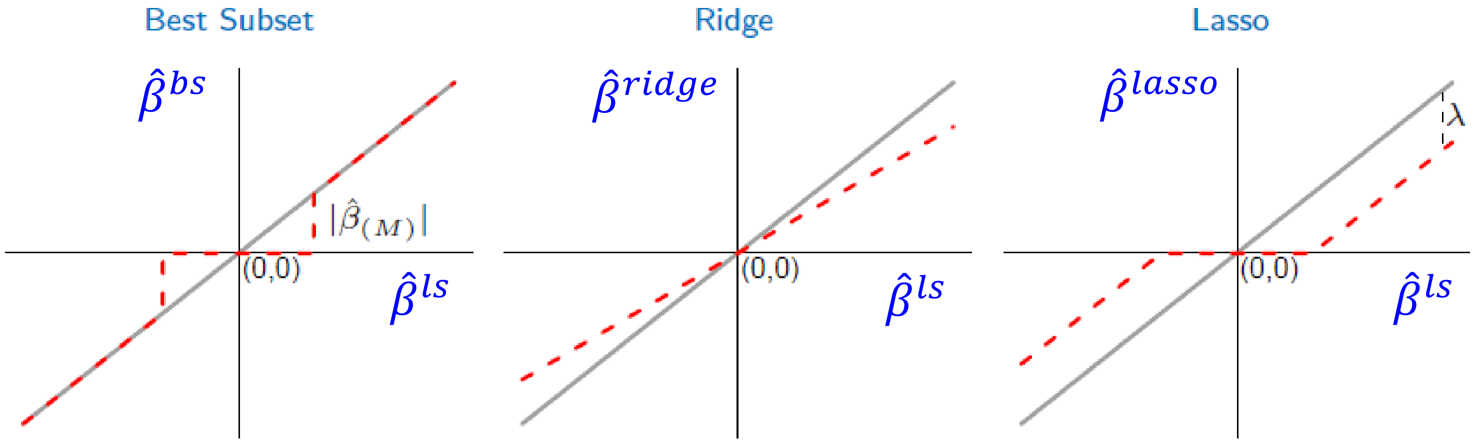
Linear Methods for Regression

--- Discussion

Shrinkage Methods – Discussion

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

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



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	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

In this table $\hat{\beta}_j$ represents $\hat{\beta}_j^{ls}$

Shrinkage Methods – Discussion

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	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

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

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

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

- Best subset

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

- Lasso

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

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

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

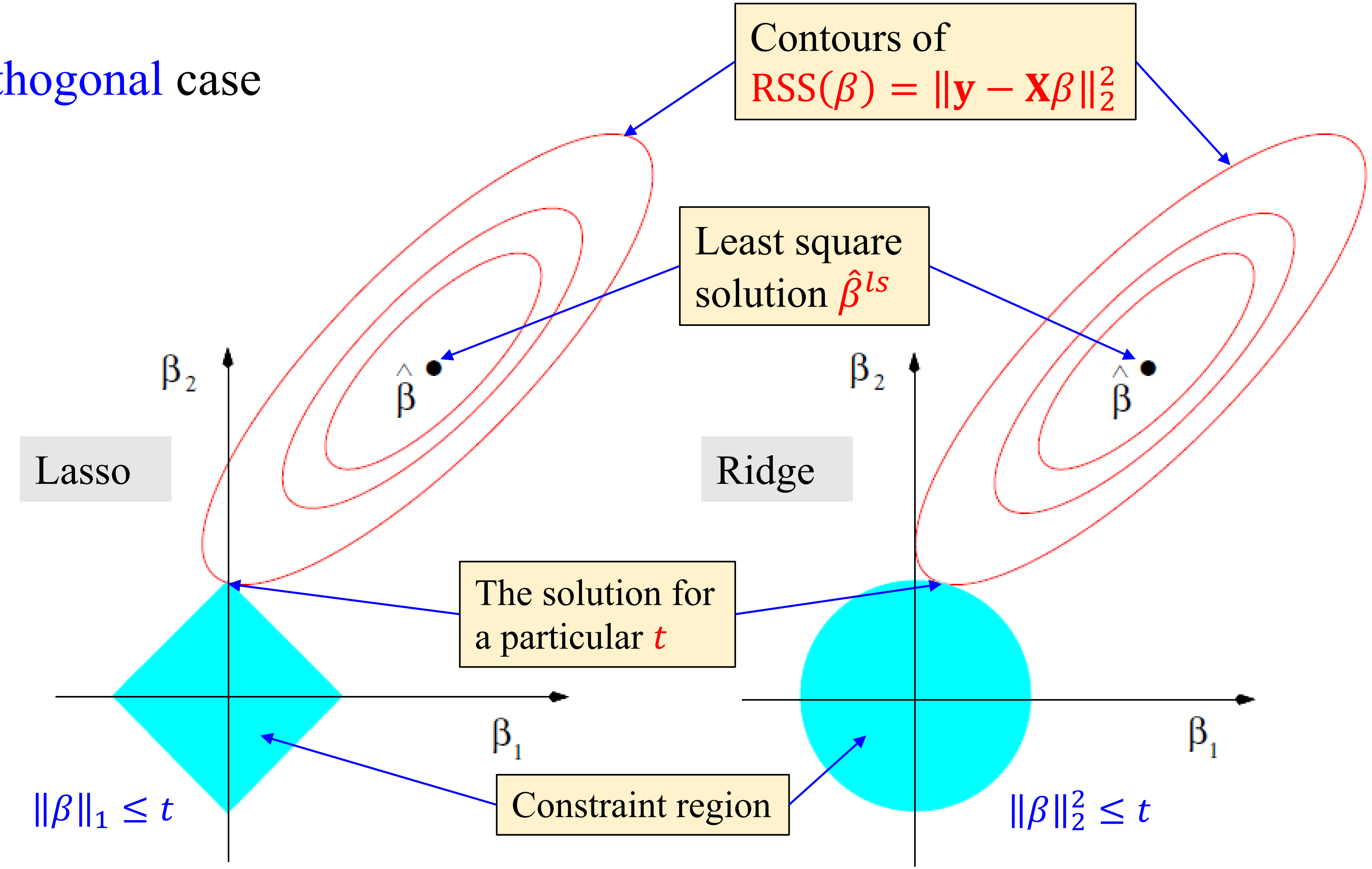
- Minimizing $\mathrm{PRSS}(\beta,\lambda)$ is equivalent to

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

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

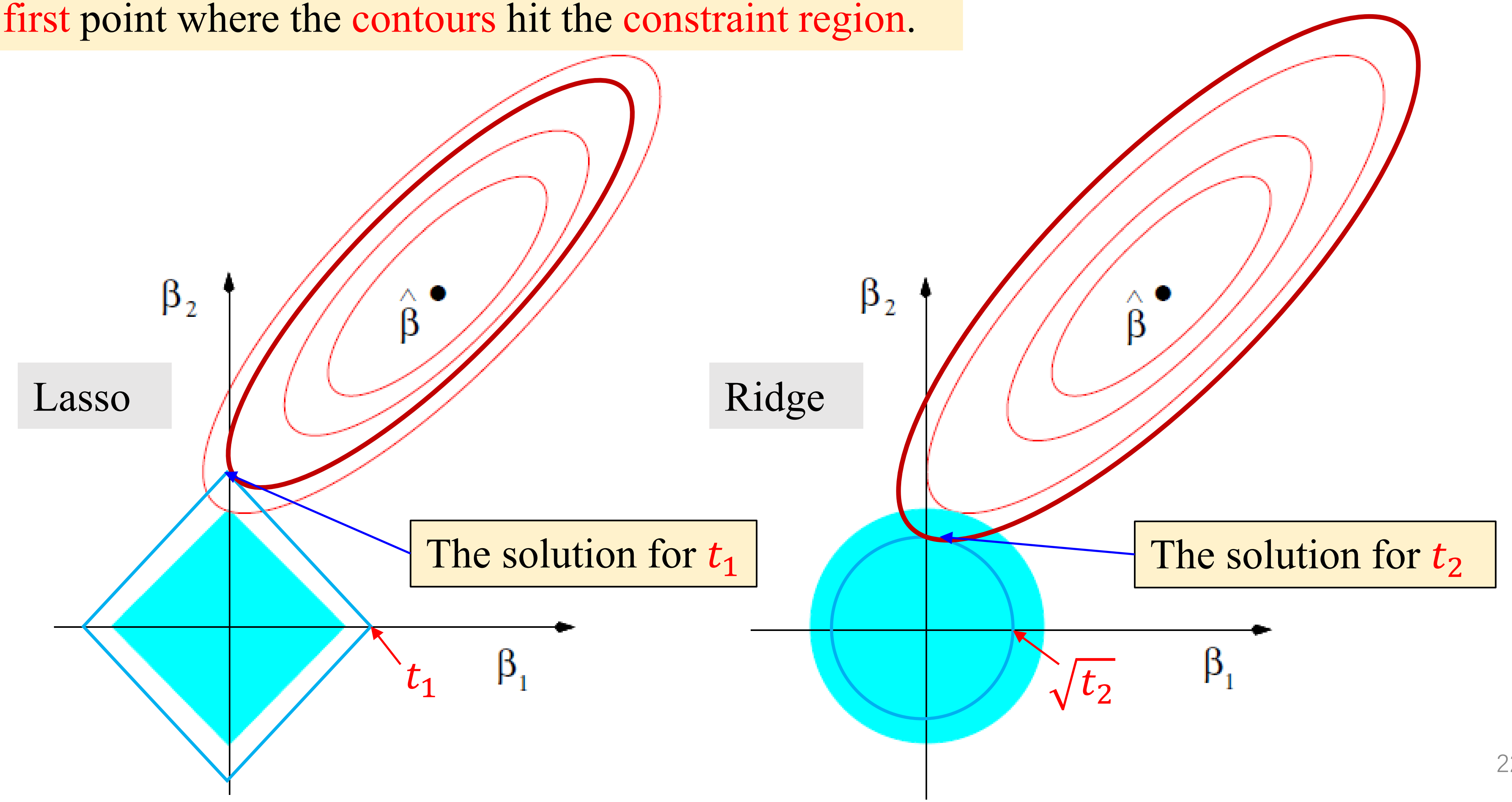
Shrinkage Methods – Discussion

Nonorthogonal case



Shrinkage Methods – Discussion

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



Shrinkage Methods – Discussion

Ridge and Lasso in the Bayes framework

- Suppose a Gaussian conditional distribution

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

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

- Log-likelihood

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

Constant

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)

Posterior

$$\hat{\beta} = \operatorname{argmax}_\beta \boxed{\Pr(\beta|\mathbf{X},\mathbf{y})} = \operatorname{argmax}_\beta \frac{\boxed{\Pr(\mathbf{y}|\mathbf{X},\beta)}\boxed{\Pr(\beta)}}{\boxed{\Pr(\mathbf{X},\mathbf{y})}}$$

Prior

Likelihood

Irrelevant with β

Posterior $\propto$ Likelihood $\times$ Prior

Shrinkage Methods – Discussion

Ridge and Lasso in the Bayes framework

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

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

- 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

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

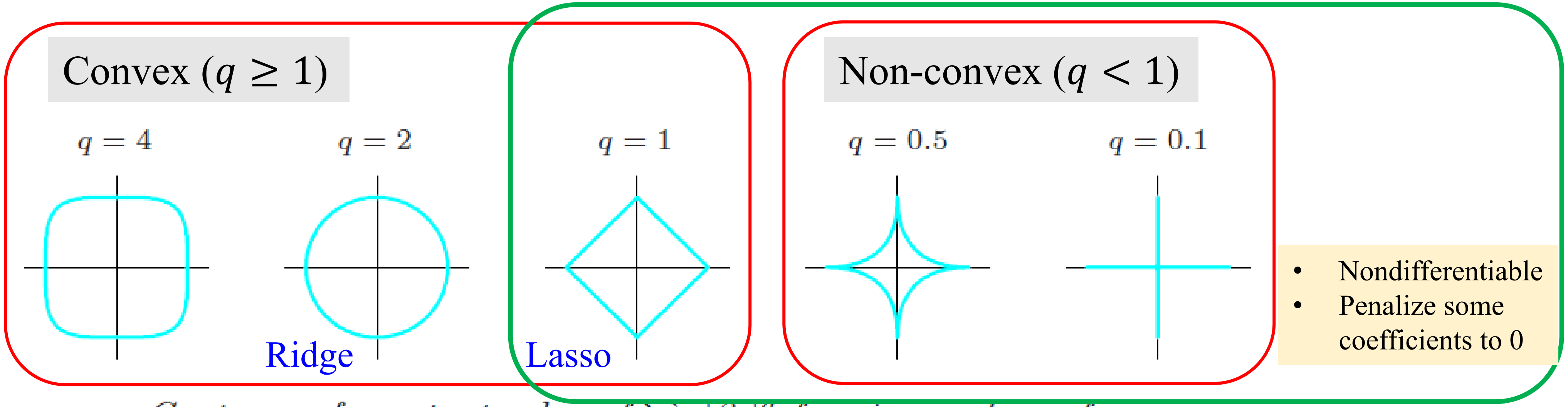
Shrinkage Methods – Discussion

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.

Shrinkage Methods – Discussion

Generalization of Ridge and Lasso

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

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

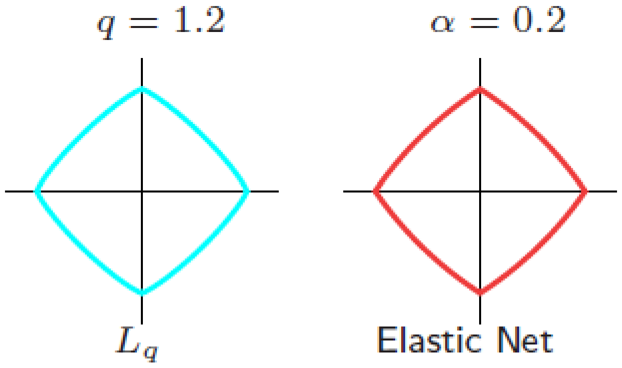
- $q = 0$, best subset
- $q = 1$, lasso
- $q = 2$, ridge regression

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

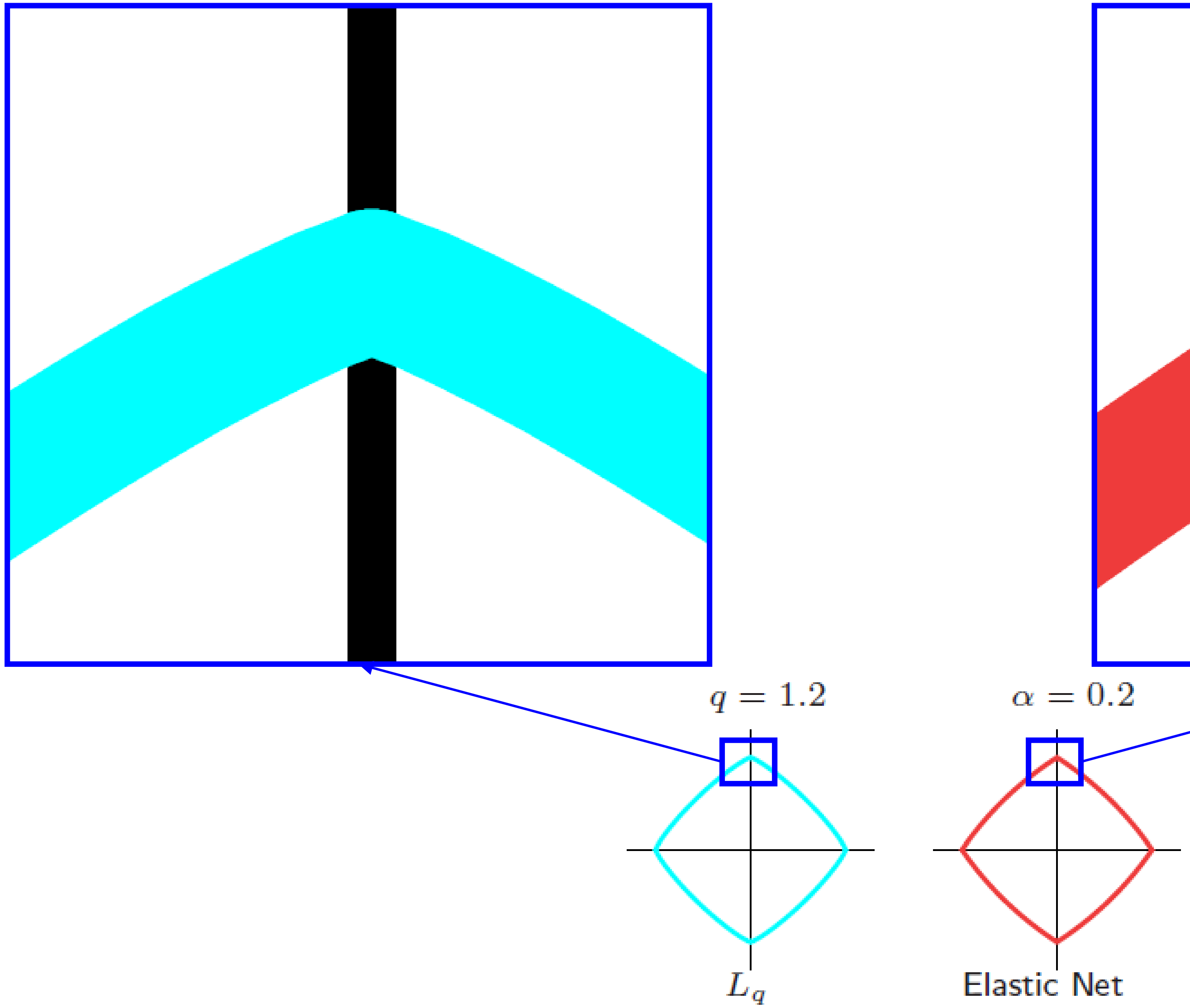
- Elastic-net

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

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



Shrinkage Methods – Discussion



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