

Optimization and Machine Learning SI151

Lu Sun

School of Information Science and Technology
ShanghaiTech University

March 11, 2020

Today:

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

Readings:

- The Element of Statistical Learning, Chapters 3
- Pattern Recognition and Machine Learning, 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
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

- the larger the value of λ , the greater the amount of shrinkage
- the coefficients are shrunk toward **zero**

- An equivalent expression

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

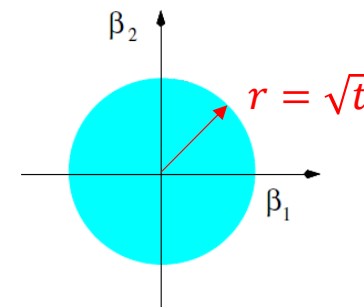
subject to $\sum_{j=1}^p \beta_j^2 \leq t,$

- 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

$$\begin{aligned} \text{P1:} \quad & \min_{\beta} f_{\lambda}(\beta) \\ & = \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \end{aligned}$$

$$\begin{aligned} \text{P2:} \quad & \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \\ & \text{s. t. } \|\beta\|_2^2 \leq t \end{aligned}$$

- Lagrange form of P2

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

- KKT conditions

1. $\nabla_{\beta} L(\beta^*, \mu^*) = \nabla_{\beta} f_{\mu^*}(\beta^*) = 0$
2. $\mu^*(\|\beta^*\|_2^2 - t) = 0$
3. $\mu^* \geq 0$

- Suppose that the solution of P1

$$\beta_{\lambda}^* = \operatorname{argmin}_{\beta} f_{\lambda}(\beta)$$

- If

$$t = \|\beta_{\lambda}^*\|_2^2$$

- then

$$\mu^* = \lambda, \quad \beta^* = \beta_{\lambda}^*$$

- Satisfy the KKT conditions.

- Conversely, suppose that P2 is solved, the setting leads to the same solution of P1

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

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

Shrinkage Methods – Ridge Regression

- Ridge regression in **matrix** form

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

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

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

$$\begin{aligned} \operatorname{PRSS}(\lambda, \beta) &= (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + \lambda \beta^T \beta \\ &= \mathbf{y}^T \mathbf{y} - \beta^T \mathbf{X}^T \mathbf{y} - \mathbf{y}^T \mathbf{X} \beta + \beta^T \mathbf{X}^T \mathbf{X} \beta + \lambda \beta^T \beta \end{aligned}$$

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

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

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

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

Shrinkage Methods – Ridge Regression

Additional insight into ridge regression

- Singular value decomposition (SVD)

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

- $\mathbf{U} \in \mathbb{R}^{N \times p}$: 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 ($d_1 \geq d_2 \geq \dots \geq d_p \geq 0$)

- 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}

Ridge regression

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

- shrinkage factor
- smaller d_j leads to a larger shrinkage

Q

Shrinkage Methods – Ridge Regression

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

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

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

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

Trace equals to sum of eigenvalues

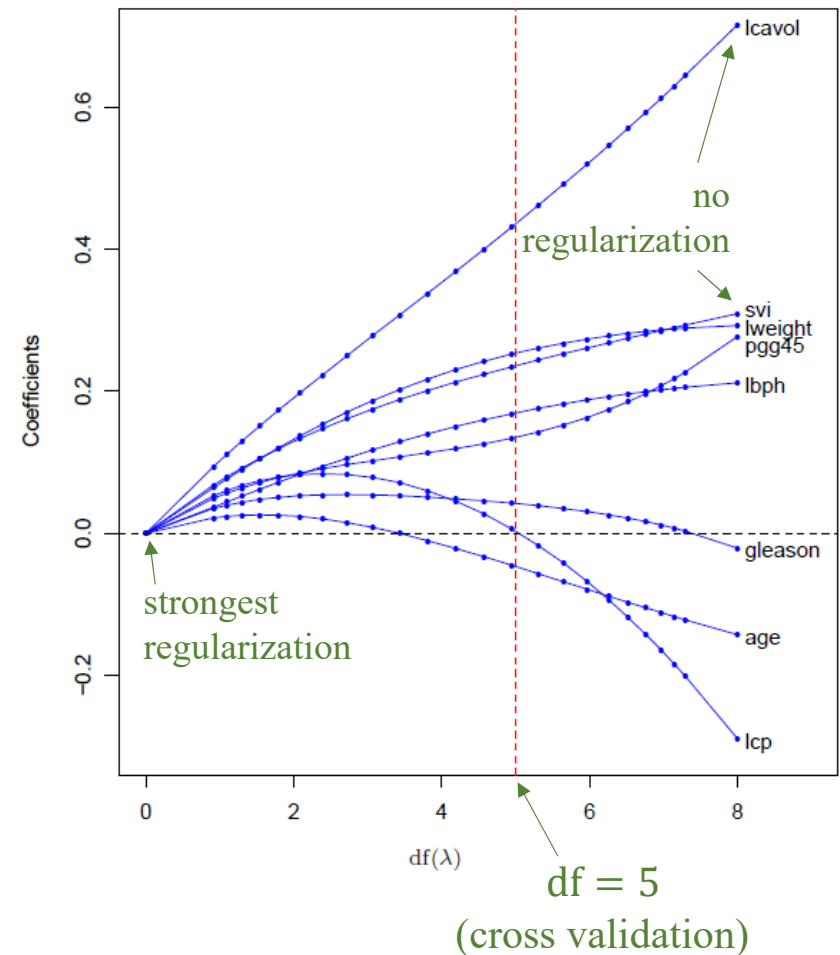
Shrinkage Methods – Ridge Regression

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

- *Effective degree of freedom*

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

- $\lambda \rightarrow 0$, $df(\lambda) = p$ ← no regularization
- $\lambda \rightarrow \infty$, $df(\lambda) \rightarrow 0$



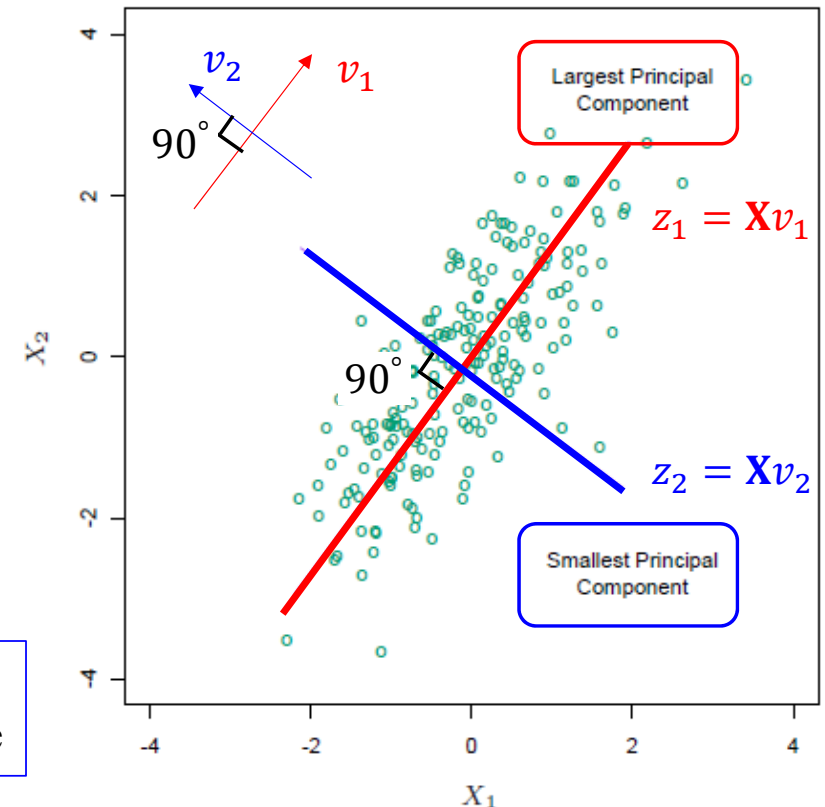
Shrinkage Methods – Ridge Regression

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

$$\mathbf{S} = \frac{1}{N} \mathbf{X}^T \mathbf{X} = \frac{1}{N} \mathbf{V} \mathbf{D}^2 \mathbf{V}^T$$
- Eigen decomposition of $\mathbf{X}^T \mathbf{X}$
- The eigenvector 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

$$\begin{aligned} \text{Var}(z_j) &= \text{Var}(\mathbf{X}v_j) \\ &= v_j^T \text{Var}(\mathbf{X}) v_j \\ &= \frac{1}{N} v_j^T \mathbf{V} \mathbf{D}^2 \mathbf{V}^T v_j \\ &= \frac{d_j^2}{N} \end{aligned}$$

- 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:

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

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**)

- Or equivalently,

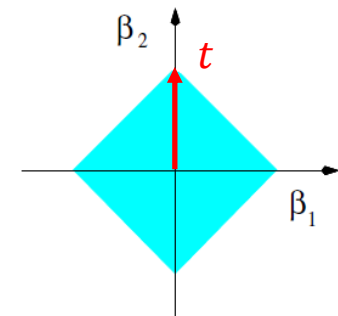
Constraint optimization

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

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

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

- making t sufficiently small \rightarrow some coefficients equal to 0



Shrinkage Methods – The Lasso

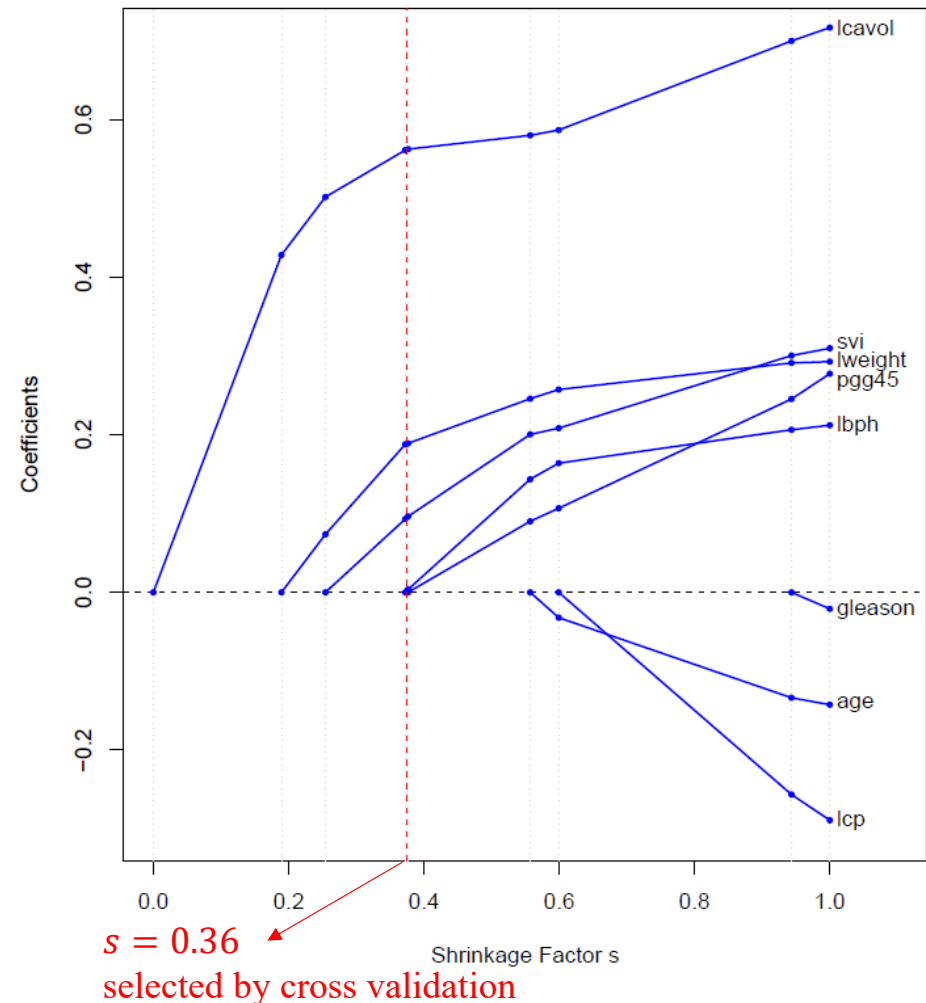
- The lasso in **matrix** form

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

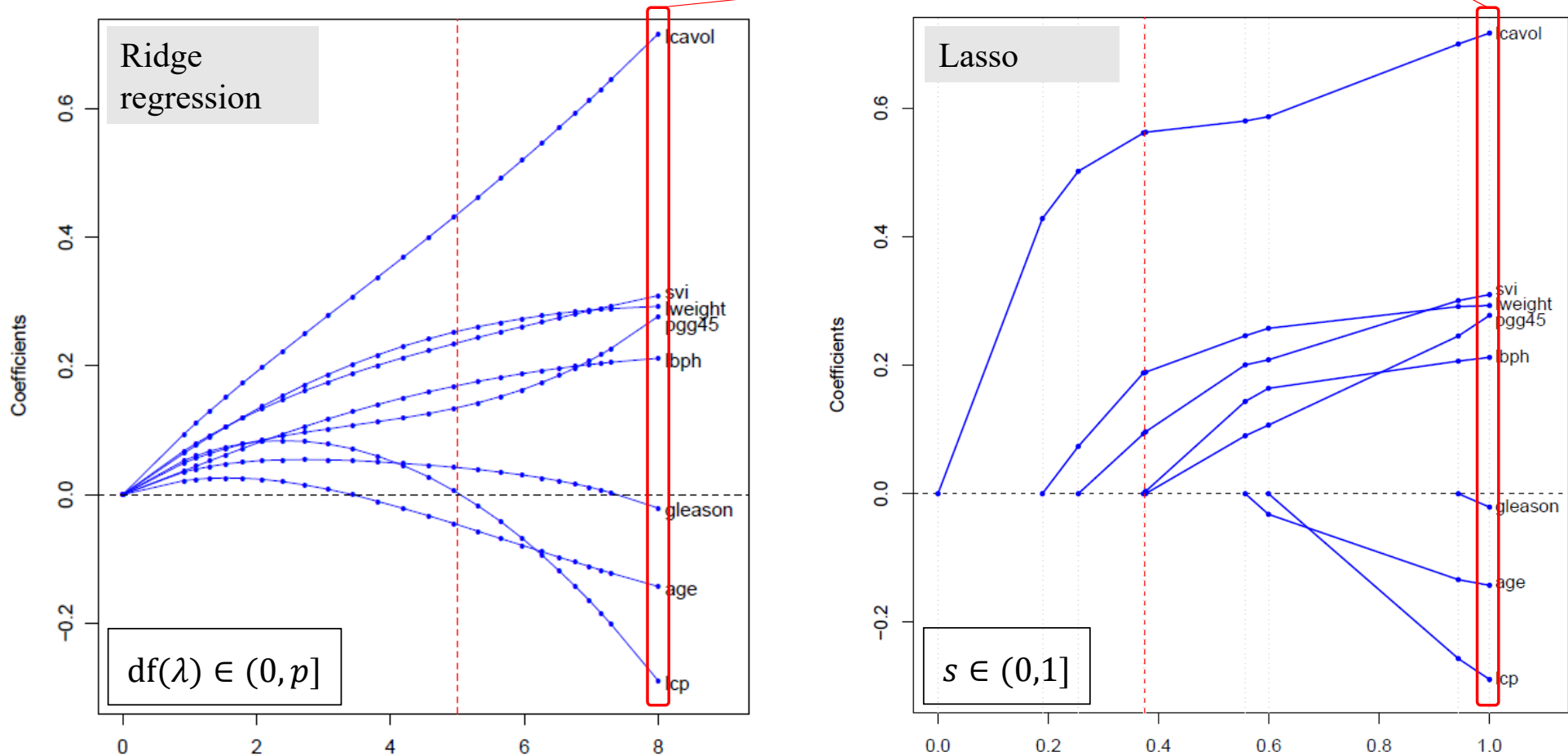
- Prostate cancer example**
- The standardized parameter

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

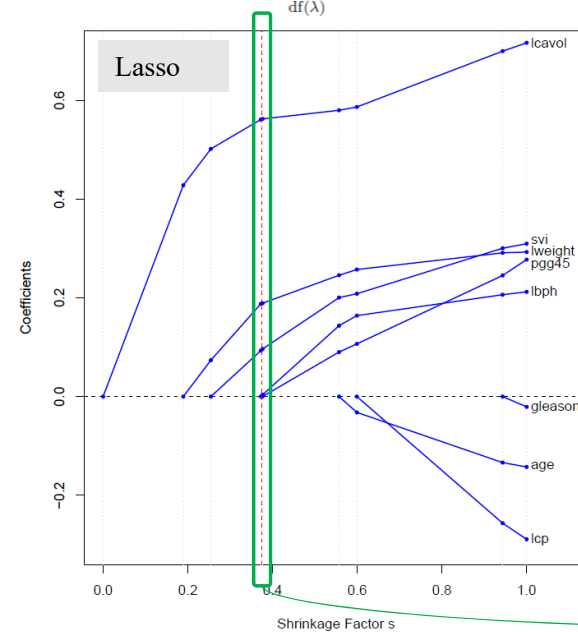
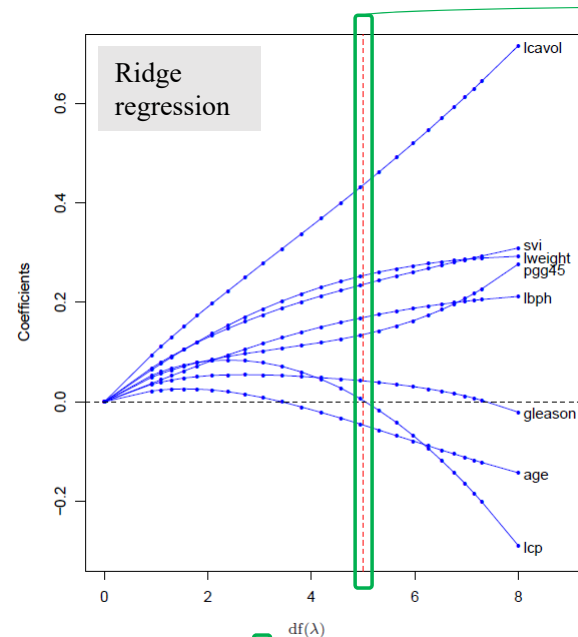
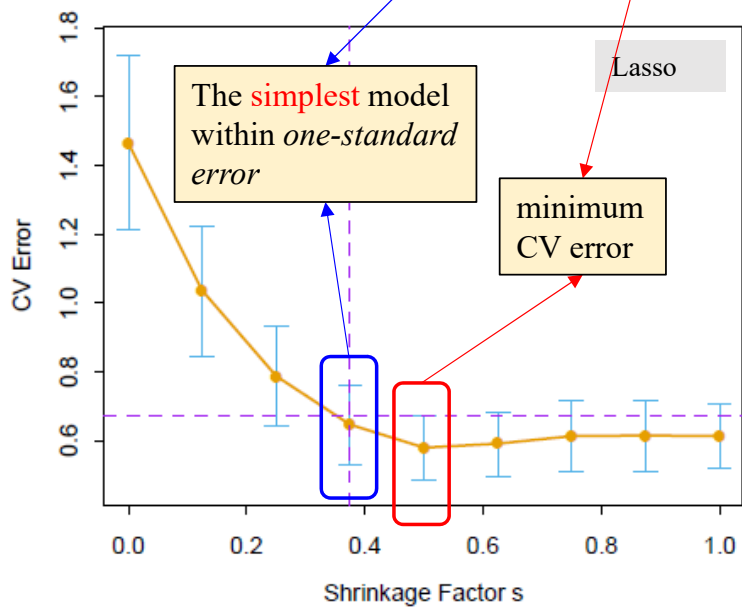
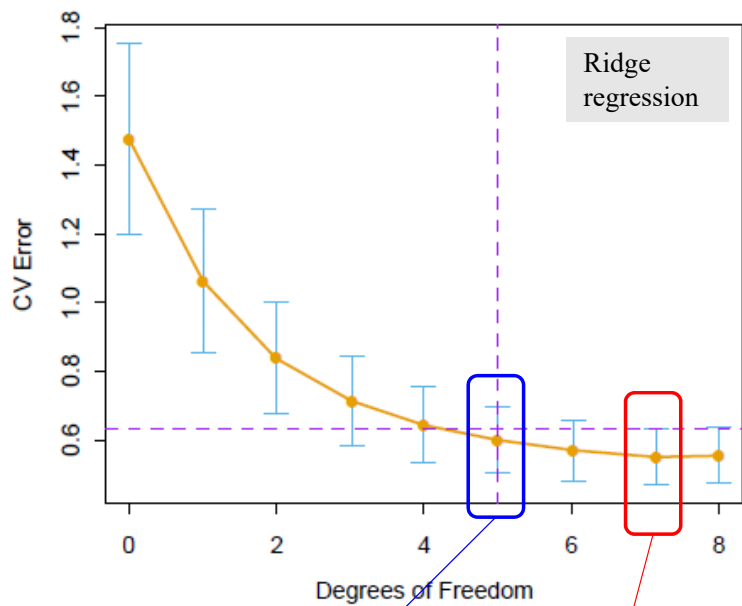
- $s = 1, \hat{\beta}^{lasso} = \hat{\beta}^{ls}$
- $s \rightarrow 0, \hat{\beta}^{lasso} \rightarrow 0$
- $s \in (0, 1), \hat{\beta}_j^{lasso} \in (0, \hat{\beta}_j^{ls}), \forall j$



Shrinkage Methods – The Lasso



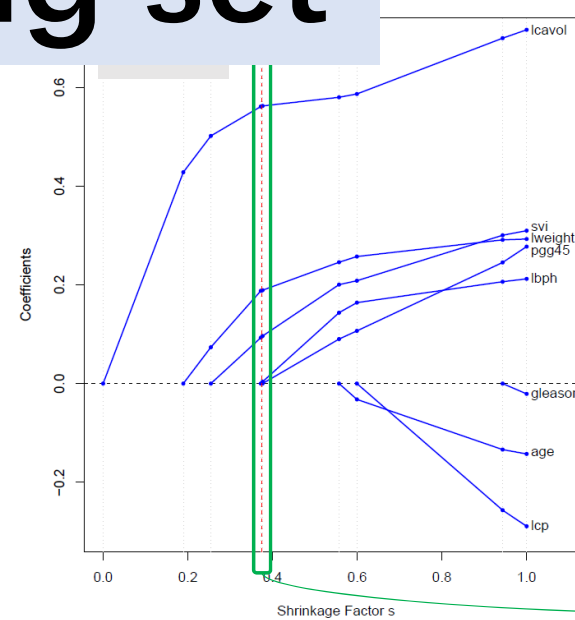
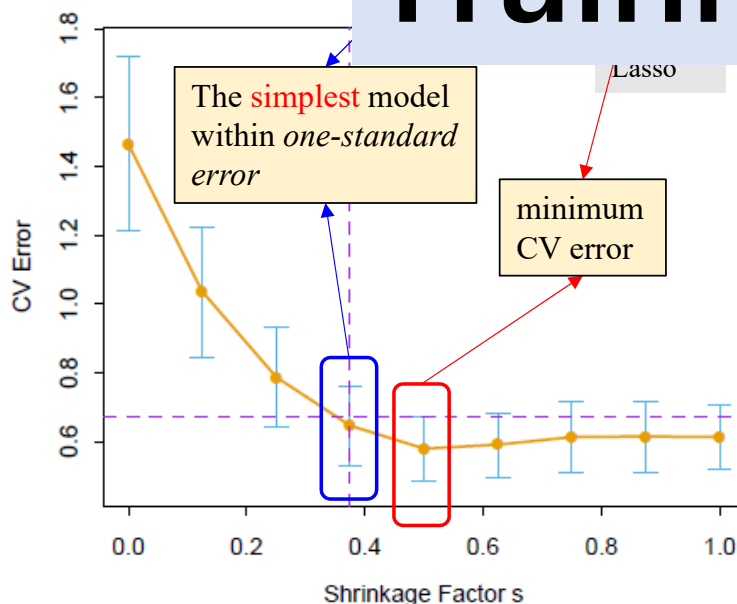
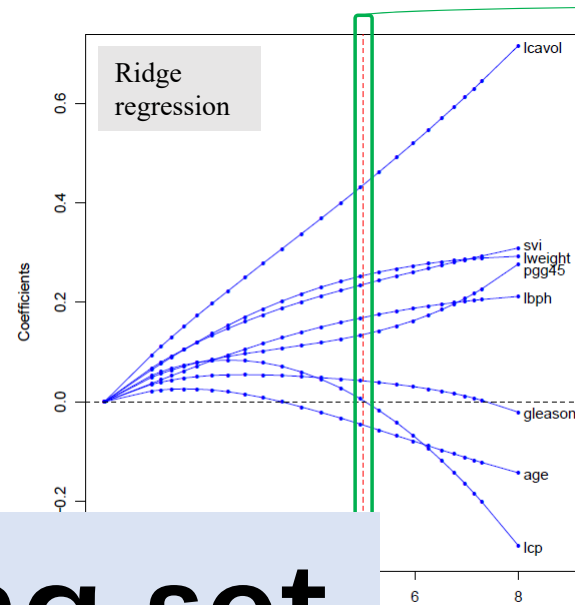
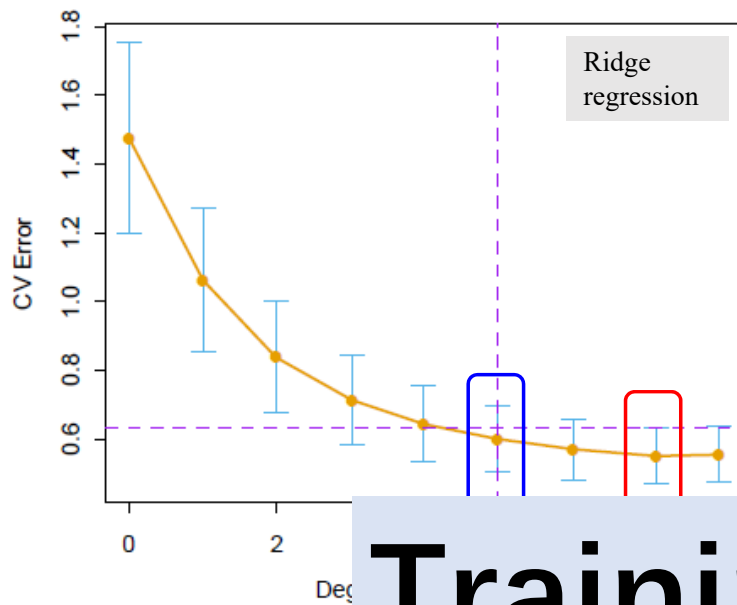
Difference: the lasso profiles hit zero, while those for ridge do not.



$df(\lambda) = 5$

Term	LS	Ridge	Lasso
lcp	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

Term	LS	Ridge	Lasso
lccavol	0.680	0.420	0.533
lweight	0.263	0.238	0.160
lcp	0.120	0.080	0.080
gleason	-0.021	0.040	0.040
pgg45	0.267	0.133	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.

$df(\lambda) = 5$

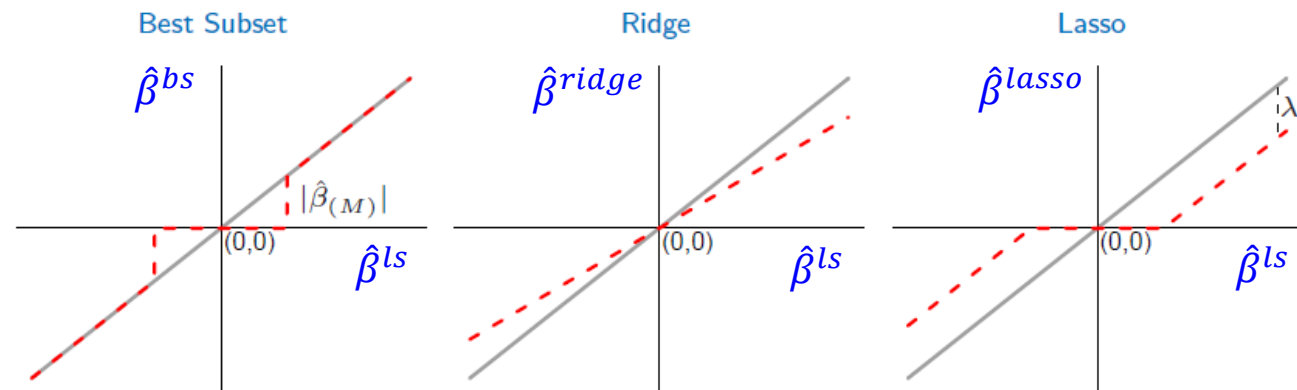
Linear Methods for Regression

--- 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)	$\hat{\beta}_j \cdot I(\hat{\beta}_j \geq \hat{\beta}_{(M)})$
Ridge	$\hat{\beta}_j / (1 + \lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\hat{\beta}_j - \lambda)_+$

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

Shrinkage Methods – Discussion

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

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

$$\begin{aligned} \hat{\beta}^{ridge} &= (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{y} \\ &= \frac{1}{1+\lambda} \mathbf{X}^T \mathbf{y} = \frac{1}{1+\lambda} \hat{\beta}^{ls} \end{aligned}$$

- Best subset

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

- Lasso

$$\begin{aligned} \text{PRSS}(\beta, \lambda) &= \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \\ &= \frac{1}{2} \mathbf{y}^T \mathbf{y} - \beta^T \mathbf{X}^T \mathbf{y} + \frac{1}{2} \beta^T \mathbf{X}^T \mathbf{X} \beta + \lambda \|\beta\|_1 \\ &= \frac{1}{2} \mathbf{y}^T \mathbf{y} - \beta^T \hat{\beta}^{ls} + \frac{1}{2} \beta^T \beta + \lambda \|\beta\|_1 \end{aligned}$$

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

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

- Signs of $\hat{\beta}_j$ and $\hat{\beta}_j^{ls}$ must be the same.

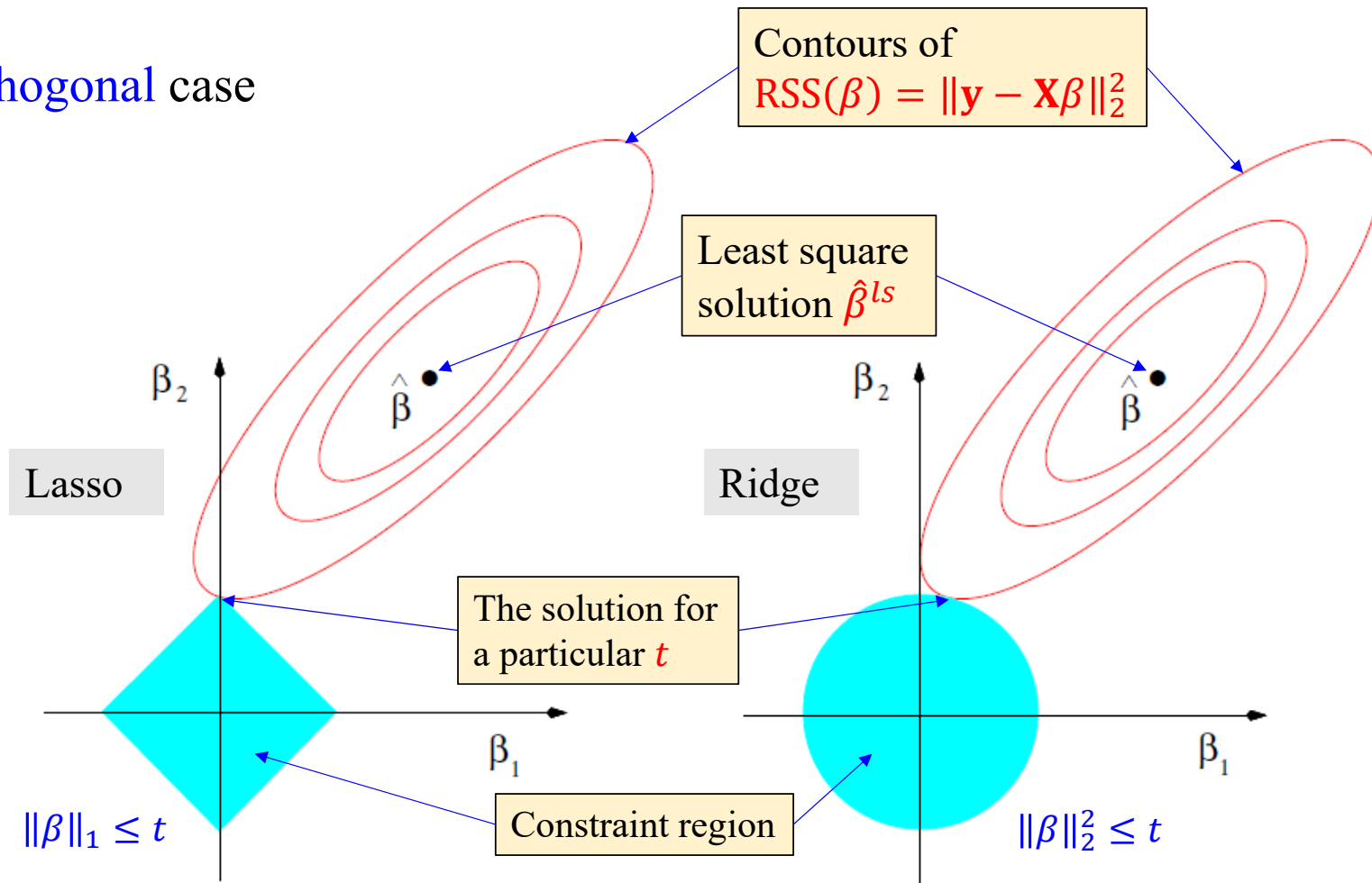
$$\square \quad \hat{\beta}_j > 0 \rightarrow \hat{\beta}_j = \hat{\beta}_j^{ls} - \lambda$$

$$\square \quad \hat{\beta}_j \leq 0 \rightarrow \hat{\beta}_j = \hat{\beta}_j^{ls} + \lambda$$

- $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{ls})(|\hat{\beta}_j^{ls}| - \lambda)_+$

Shrinkage Methods – Discussion

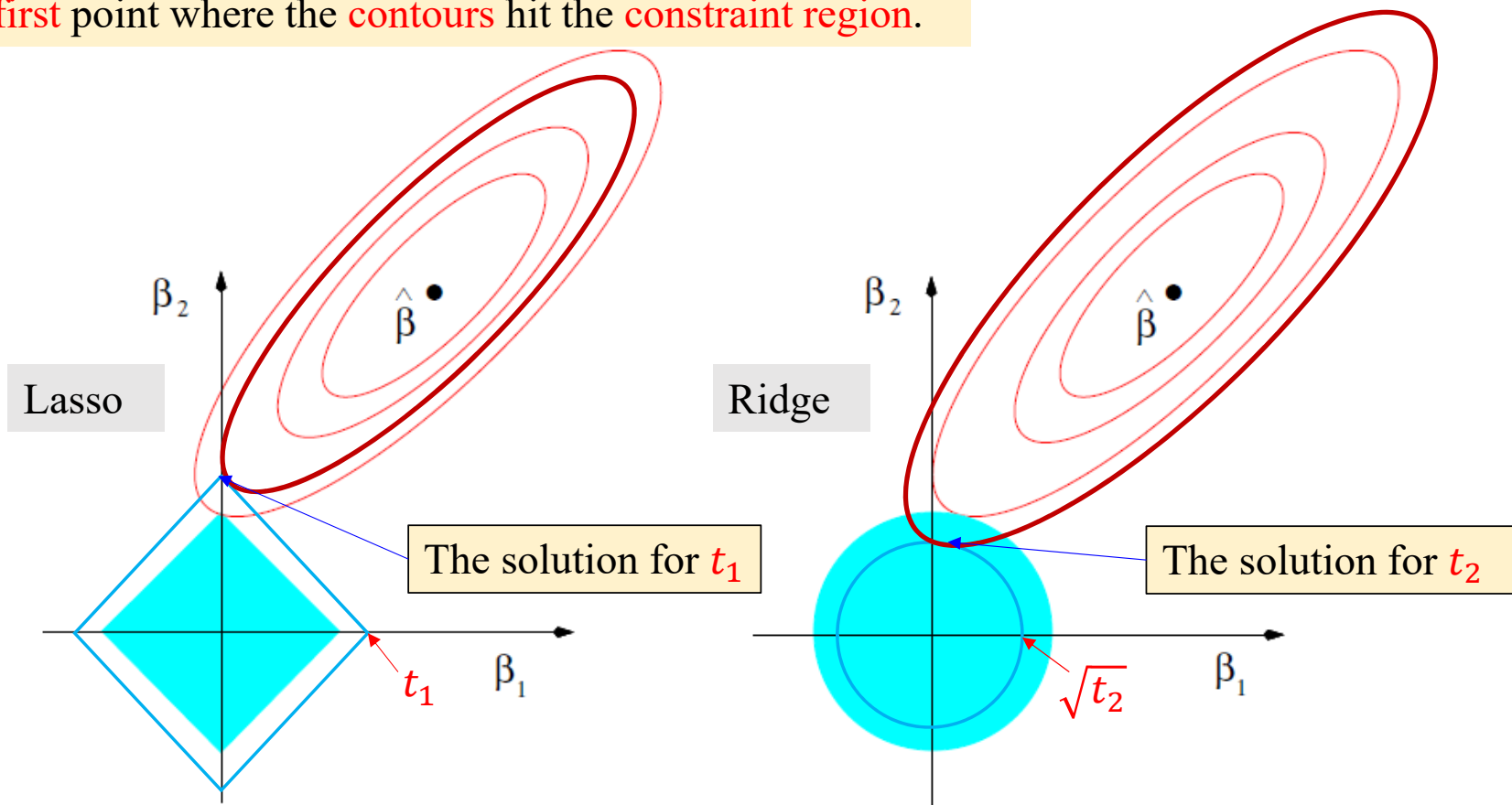
Nonorthogonal case



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

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

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

- Log-likelihood

$$\begin{aligned} L(\beta) &= \ln \Pr(\mathbf{y}|\mathbf{X}, \beta) \\ &= \sum_{i=1}^N \ln \Pr(y_i|x_i, \beta) \end{aligned}$$

MLE:

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

$$\text{Constant} \leftarrow = -\frac{N}{2} \log(2\pi) - N \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - x_i^T \beta)^2$$

- Maximum a posterior (**MAP**)

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

Posterior

Prior

Likelihood

Irrelevant with β

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

Shrinkage Methods – Discussion

Ridge and Lasso in the **Bayes** framework

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

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

- Ridge regression

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

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

- Lasso

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

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

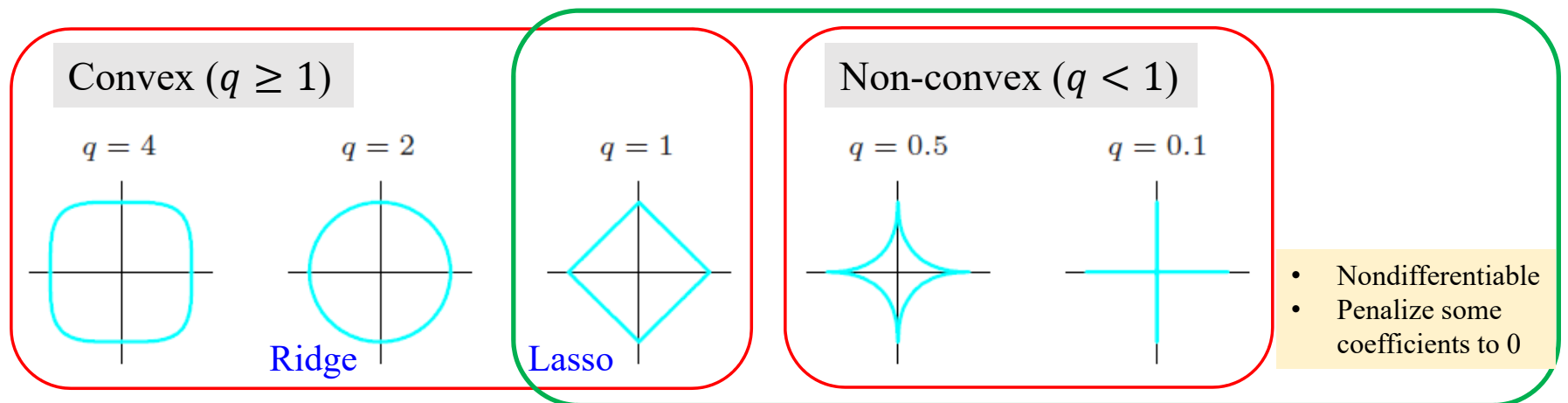
Shrinkage Methods – Discussion

Generalization of Ridge and Lasso

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

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

- $q = 0$, 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 ($q \geq 0$)

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

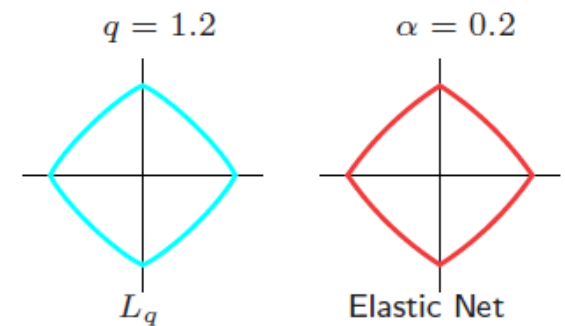
- $q = 0$, best subset
- $q = 1$, lasso
- $q = 2$, ridge regression

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

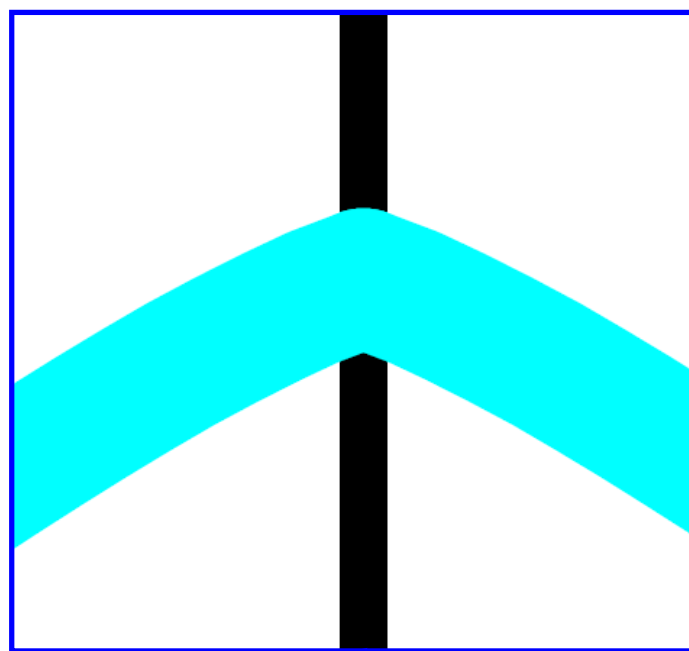
- Elastic-net

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

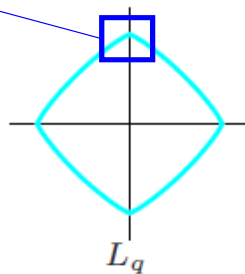
- ℓ_2 shrinks the coefficients of correlated predictors
- ℓ_1 selects groups of correlated predictors



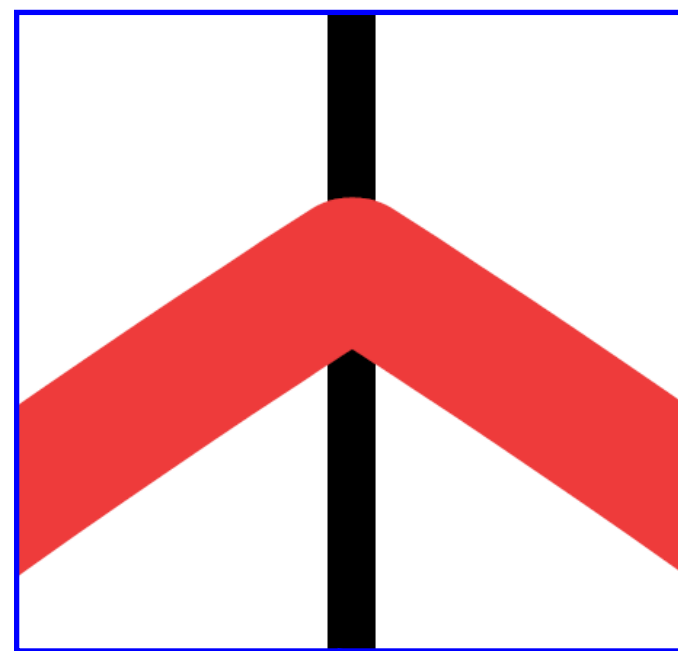
Shrinkage Methods – Discussion



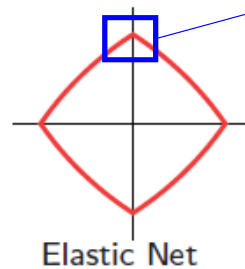
$q = 1.2$



L_q



$\alpha = 0.2$



Elastic Net

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