Linear Methods for Regression: Shrinkage Methods (Regularization)

Statistical learning – Part II

Alberto Castellini
University of Verona

Motivation

- **Subset selection** is a **discrete** process (variables are retained or discarded).
- It often exhibits **high variance**, thus it does **not always reduce the prediction error** of the full model.
- **Shrinkage methods** are more **continuous** and they **do not suffer** as much from **high variability.**

Ridge regression

- **Ridge regression** shrinks the regression coefficients imposing an L_2 penalty on their size

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

Complexity parameter: controls the amount of shrinkage

Lagrangian form

- The **larger** the value of **λ**, the **greater** the amount of **shrinkage**.
- Coefficients are **shrunk towards zero**.
- Penalization of the sum-of-squares of parameters is used also in *neural networks* (**weight decay**).

Equivalent way to write the Ridge problem

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

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

The **size constraint** t on parameters is **explicit.**

In case of **many correlated variables,** coefficients of OLS may become poorly determined (high variance).

- A large positive coefficient in one variable can be **canceled** by a negative coefficient of a correlated variable
- This problem is alleviated by the above formulation (squared constraint penalizes large coefficients)

Assumptions

- Data **standardization** is needed since solutions are not equivalent under scaling.
- The **intercept** β_0 is not shrunk
- The computation of β^{ridge} can be separated in **two steps**:
  - 1. $\boldsymbol{\beta_0}$ is estimated by $\bar{y} = \frac{1}{N} \sum_1^N y_i$
  - 2. **all coefficients except $\boldsymbol{\beta_0}$** are computed from centered x and without intercept by ridge regression

Matrix form for the ridge RSS

- Residual Sum of Squares:

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

- Ridge regression solution:

Identity matrix

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

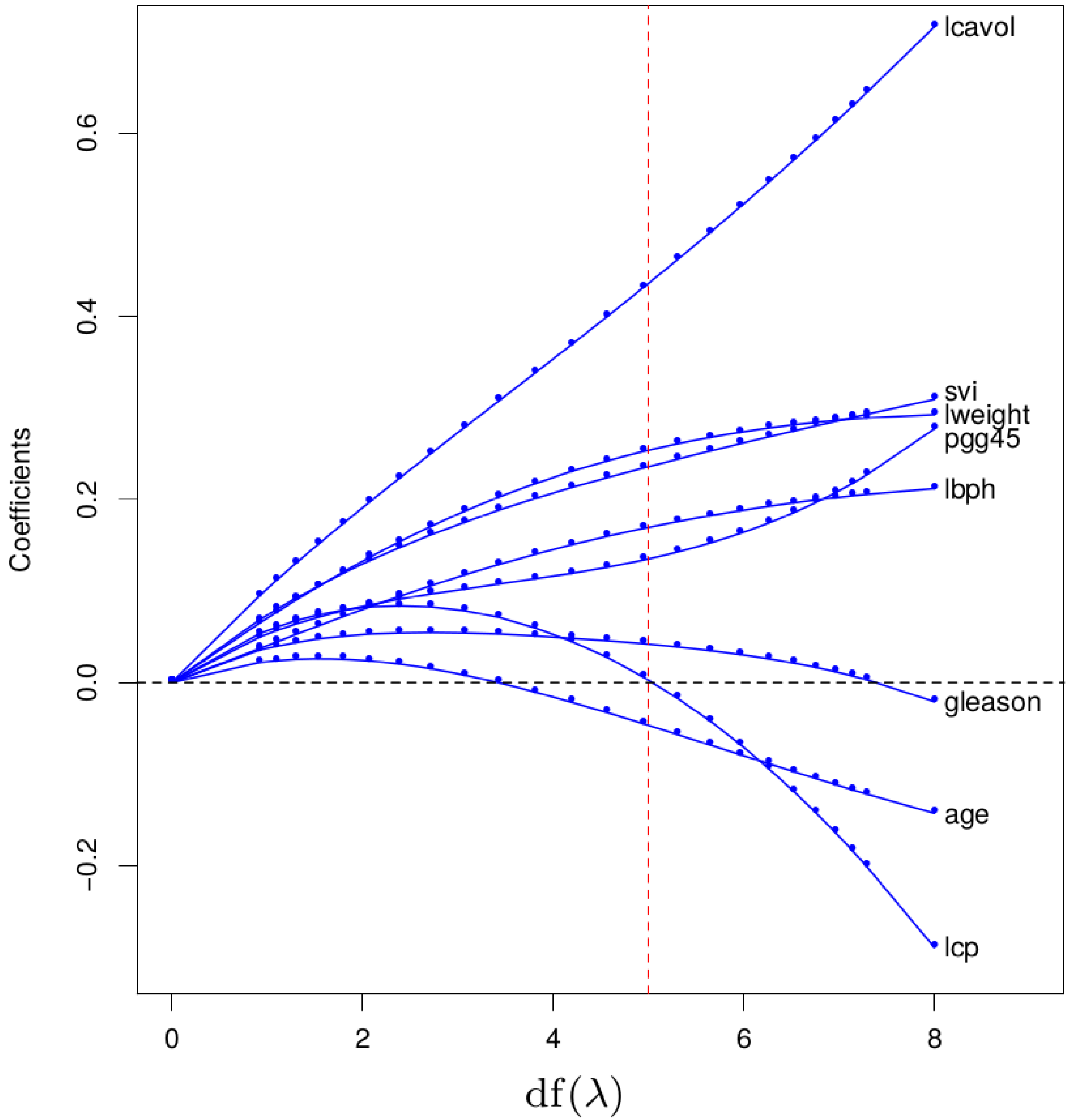
$\approx$ Covariance matrix

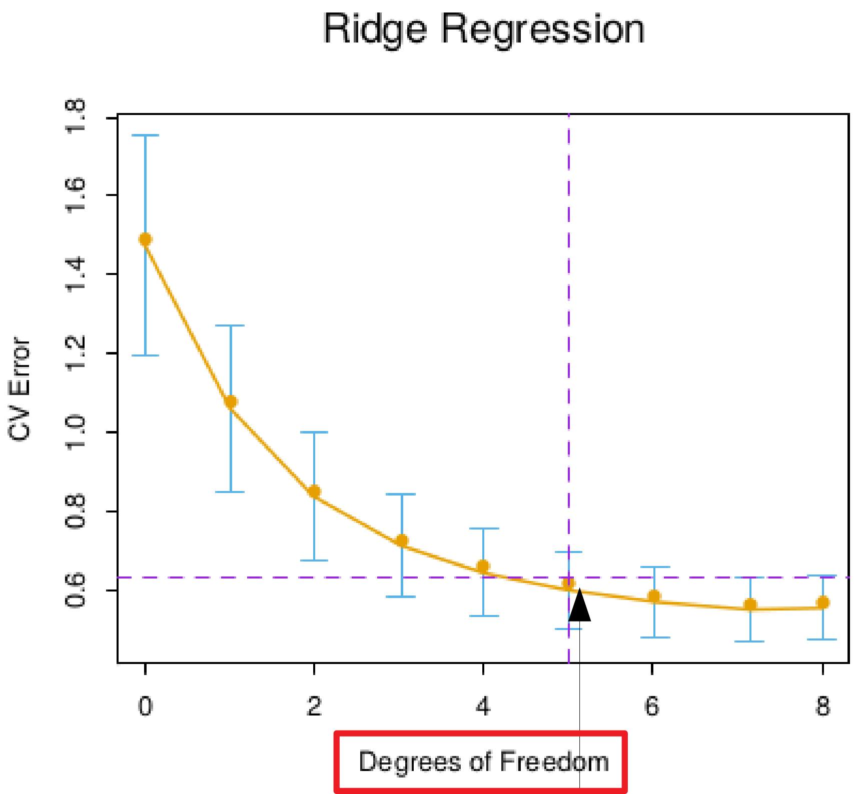
- The quadratic penalty $\beta^T\beta$ ensures that **ridge** regression solution is a **linear** function of **y**.

- The solution **adds a positive constant** to the **diagonal** of X^TX before inversion → **nonsingular problem even if X has not full rank**

Main motivation for ridge regression when it was introduced (Hoerl and Kennard, 1970)

Ridge coefficient estimate for prostate cancer example





Selection based on
1-standard error rule

In case of **orthonormal inputs** $\hat{\beta}^{\text{ridge}} = \hat{\beta}/(1+\lambda)$

Singular Value Decomposition (SVD) and Ridge regression

The **SVD** of the centered matrix X provides additional **insight** into the nature of the ridge regression.

The SVD of the N x p matrix **X** can be written as:

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

- U and V **orthogonal** matrices
- Columns of U (Nxp) span the **column space** of X
- Columns of V (pxp) span the **row space** of X
- D is a p x p diagonal matrix with entries d1 >= d2 >= … >= dp >=0 **singular values** of X.
- If one or more dj=0 then X is **singular**

Singular Value Decomposition (SVD) and Ridge regression

Using the SVD the **least squares fitted vector** can be written as:

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

Similar to the OLS case

$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}$$

(QR decomposition)

and the **ridge solutions** can be expressed as:

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

Projection of Y on column space of X

where $\mathbf{u}_j$ are the columns of U and $d_j^2 / (d_j^2 + \lambda) <= 1$.

- As in OLS, ridge regression computes the coordinates of y as **linear combinations of the orthonormal basis U**. Then it **shrinks** the coordinates by the factor $d_j^2 / (d_j^2 + \lambda)$.
- **The smaller d_j^2 the larger the amount of shrinkage.**

Singular Value Decomposition (SVD) and Ridge regression

Using the SVD the **least squares fitted vector** can be written as:

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

Similar to the OLS case

$$\hat{\mathbf{y}} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}$$

(QR decomposition)

and the **ridge solutions** can be expressed as:

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

Projection of Y on column space of X

where u_j are the columns of U and $d_j^2 / (d_j^2 + \lambda) <= 1$.

- As in OLS, ridge regression computes the coordinates of y as **linear combinations of the orthonormal basis U**. Then it **shrinks** the coordinates by the factor $d_j^2 / (d_j^2 + \lambda)$.
- **The smaller d_j^2 the larger the amount of shrinkage.**

What are the d_j?

Principal component interpretation

The **SVD** of the centered matrix X is a way of expressing the **principal component** of the variables in X.

Using the SVD, the **covariance matrix** can be written as:

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

which is the **eigen decomposition of X^TX.**

- The **eigenvectors** v_j (columns of V) are the **principal component** (Karhunen–Loeve) directions of X.
- The **first** principal component has the property that $\mathbf{z}_1 = X^*\mathbf{v}_1$ has the **largest sample variance**

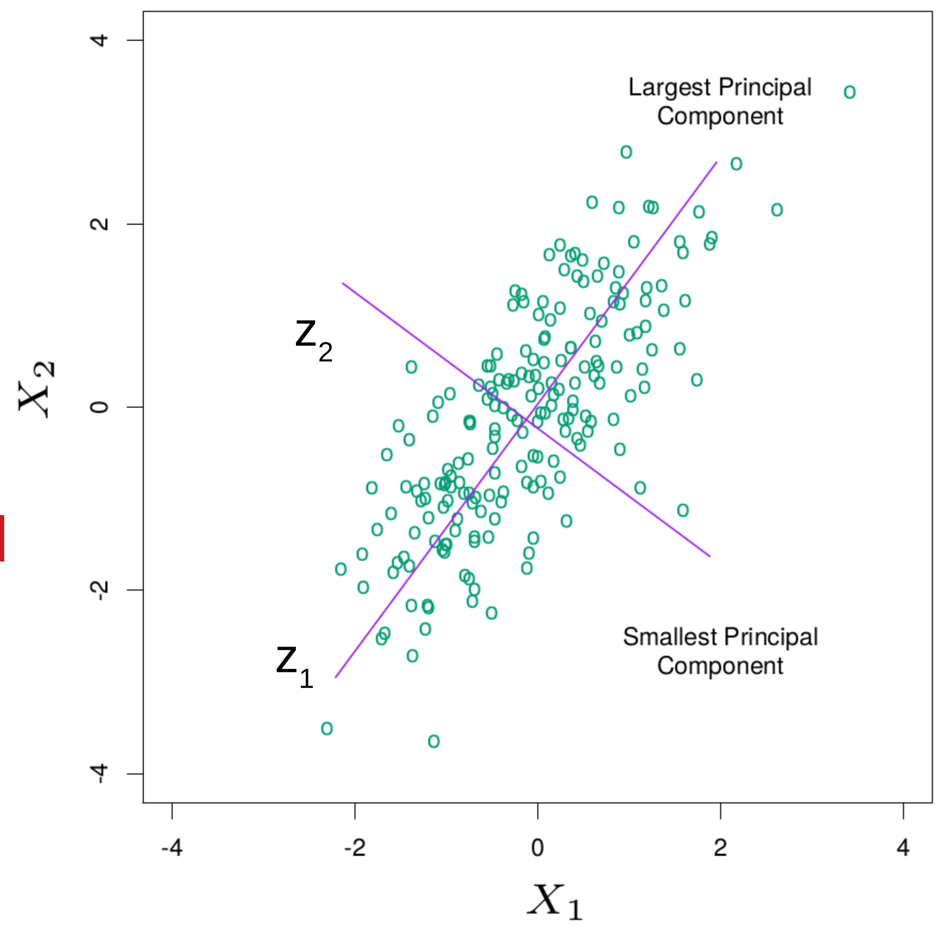
$$\mathrm{Var}(\mathbf{z}_1) = \mathrm{Var}(\mathbf{X}v_1) = \boxed{\frac{d_1^2}{N}}$$

- Similar for other d_j

Principal component interpretation

Subsequent principal components
z_j have maximum variance $\mathbf{d_j^2/N}$, subject to being **orthogonal** to the earlier ones

- The **last principal component** has **minimum variance**
- **Small singular values d_j** correspond to **directions** in the column space of X having **small variance**
- **Ridge** regression **shrinks** these directions **the most**



- **Implicit assumption:** the **response** will tend to **vary most** in the directions of **high variance** of the inputs
- Often reasonable but need not hold in general

Effective degrees of freedom

- Although **all p coefficients in a ridge fit will be non-zero**, they are **fit in a restricted fashion** controlled by λ.

- The **effective degree of freedom** of the ridge regression fit is:

$$
\begin{aligned}
\mathrm{df}(\lambda) &= \mathrm{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T] \\
&= \mathrm{tr}(\mathbf{H}_\lambda) \\
&= \boxed{\sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.}
\end{aligned}
$$

- df(λ) = p when λ = 0 (no regularization)

- df(λ) → 0 as λ → ∞.

Ridge coefficient estimate for prostate cancer example

Term	LS	Best Subset	Ridge
Intercept	2.465	2.477	2.452
lcavol	0.680	0.740	0.420
lweight	0.263	0.316	0.238
age	−0.141		−0.046
lbph	0.210		0.162
svi	0.305		0.227
lcp	−0.288		0.000
gleason	−0.021		0.040
pgg45	0.267		0.133
Test Error	0.521	0.492	0.492
Std Error	0.179	0.143	0.165

Ridge regression **reduces the test error** of the full least squares estimates by a **small amount**

LASSO regression

- The **lasso estimate** is defined by

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \Bigg\{ \underbrace{\frac{1}{2} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2}_{\textbf{Goodness-of-fit}} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\textbf{Penalty}} \Bigg\}$$

Complexity parameter: controls the amount of shrinkage

Lagrangian form

- The $\mathbf{L_2}$ ridge penalty $\sum_1^p \beta_j^2$ is **replaced** by the $\mathbf{L_1}$ lasso penalty $\sum_1^p |\beta_j|$
- The nature of the shrinkage causes some of the **coefficients to be exactly zero** (kind of **continuous subset selection**)

LASSO regression

- Alternative (non-Lagrangian) form of the lasso problem:

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

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

- If t is chosen lager than $t_0 = \sum_1^p |\hat{\beta}_j|$ then no shrinkage is performed.

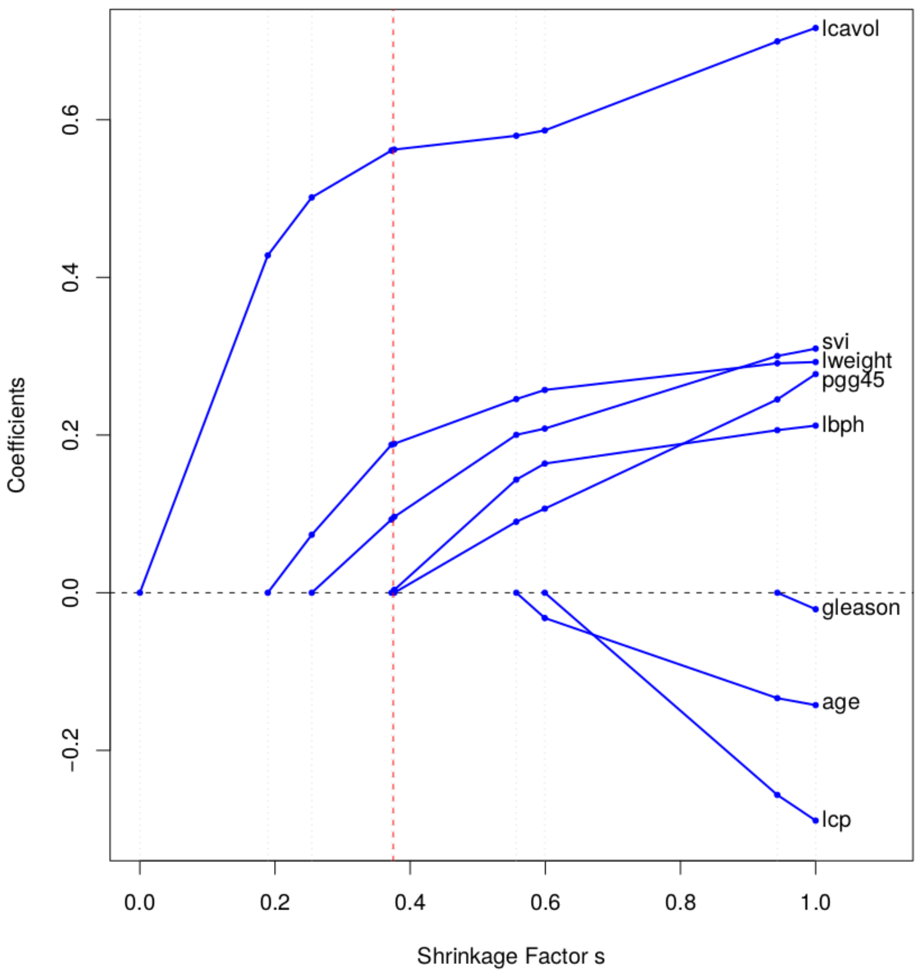
- For $t = t_0/2$ for instance, OLS coefficients are shrunk of 50% on average.

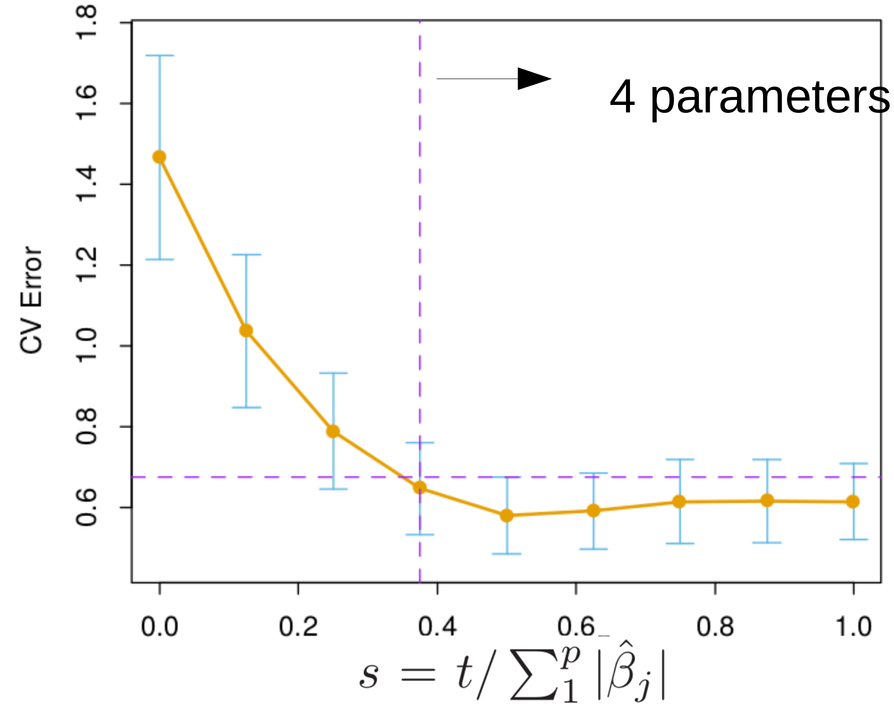
- The **nature of shrinkage** is not obvious.

Complexity

- The LASSO constraint makes the solution **nonlinear** in the y_i
- **No closed form** expression as in ridge regression
- **Quadratic programming** problem
- The **complexity parameter** should be chosen to **minimize** an **estimate** of the **expected prediction error** (cross validation)

Coefficient estimate for prostate cancer example

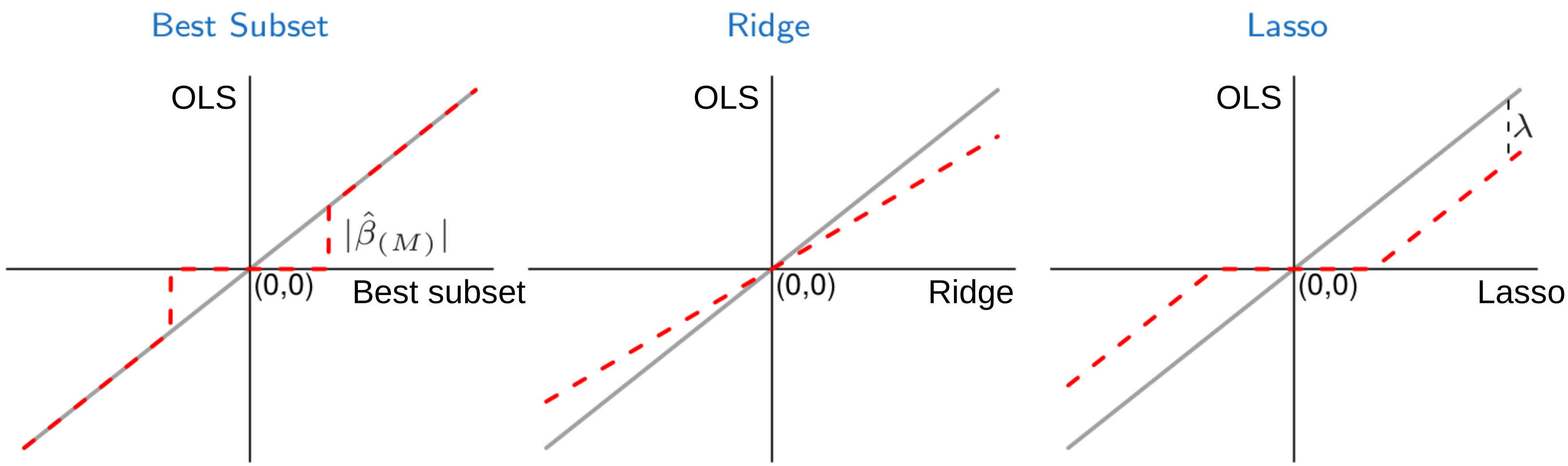




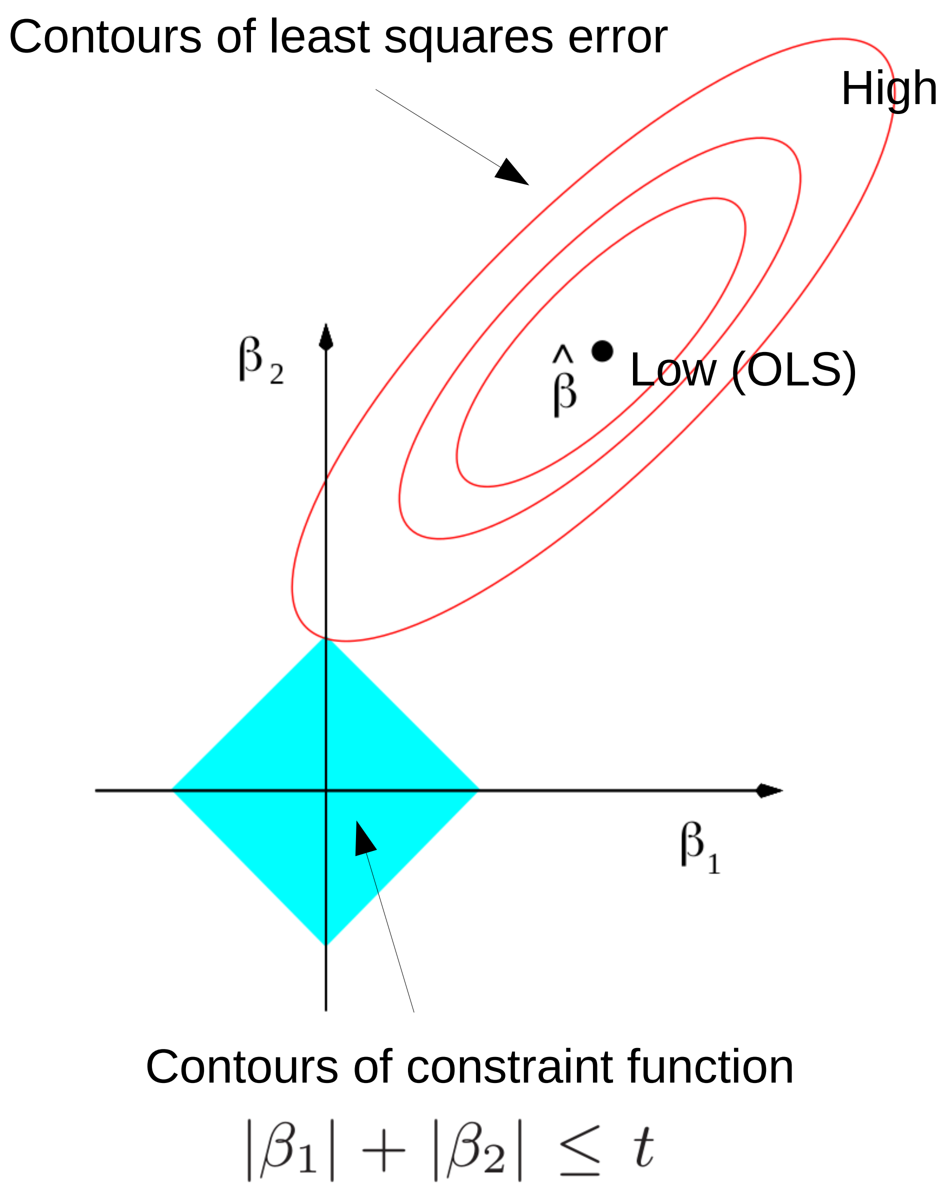
Term	LS	Best Subset	Ridge	Lasso
Intercept	2.465	2.477	2.452	2.468
lcavol	0.680	0.740	0.420	0.533
lweight	0.263	0.316	0.238	0.169
age	−0.141		−0.046	
lbph	0.210		0.162	0.002
svi	0.305		0.227	0.094
lcp	−0.288		0.000	
gleason	−0.021		0.040	
pgg45	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

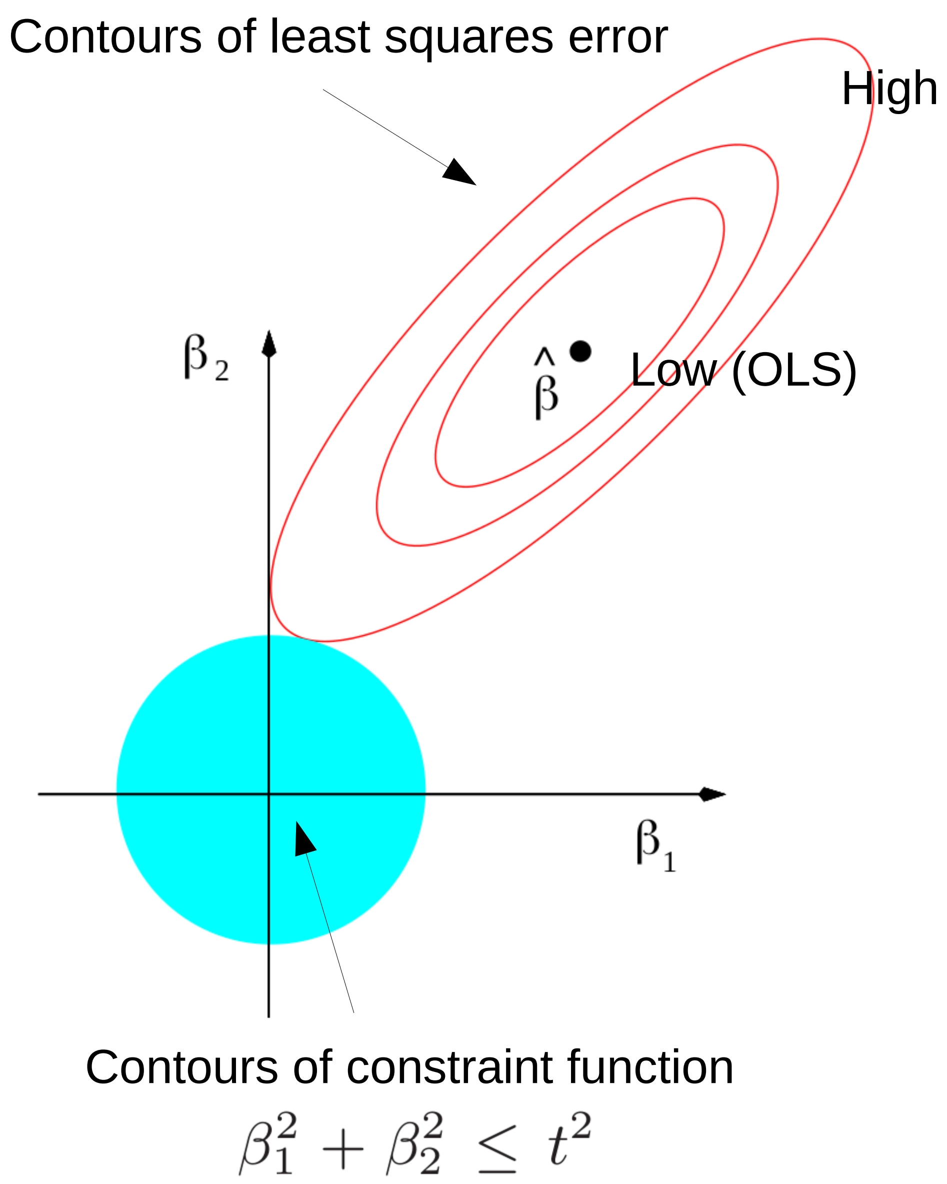
"Nature of shrinkage": comparison (1/2)

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



“Nature of shrinkage”: comparison (2/2)





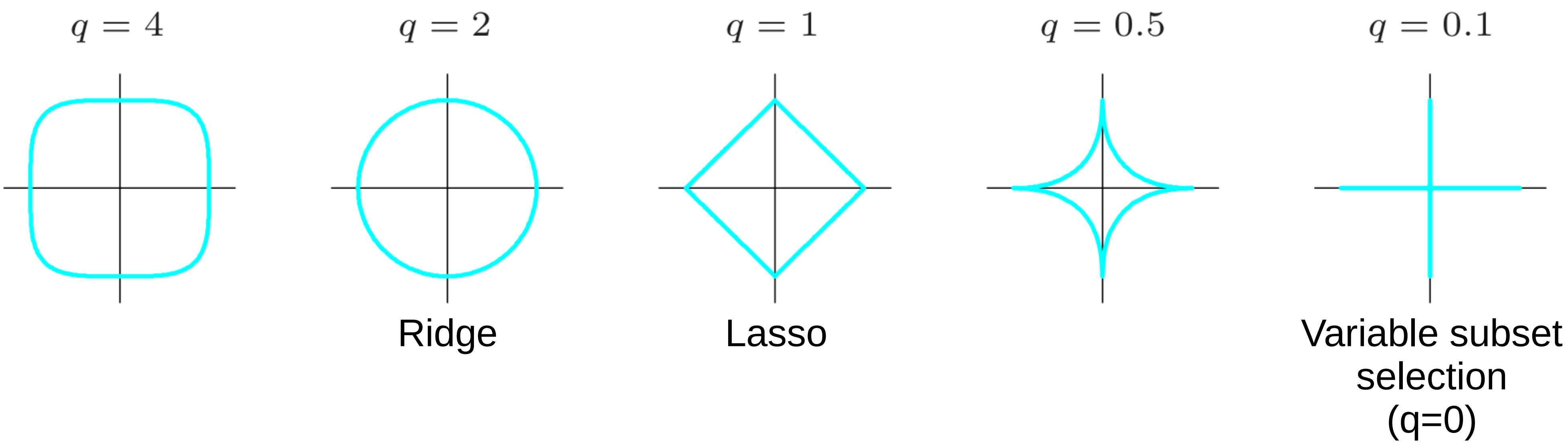
Generalizations of ridge and lasso regression

- Ridge regression and lasso can be generalized by

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

where q>=0.

- The contours of $\sum_j |\beta_j|^q$ for different q are shown in the following:



- **Lasso** sets coefficients to zero because its $|\beta|^1$ is not differentiable at 0
- **Ridge** shrinks together coefficients of correlated variables
- How to put these two effects together?

Elastic net regression

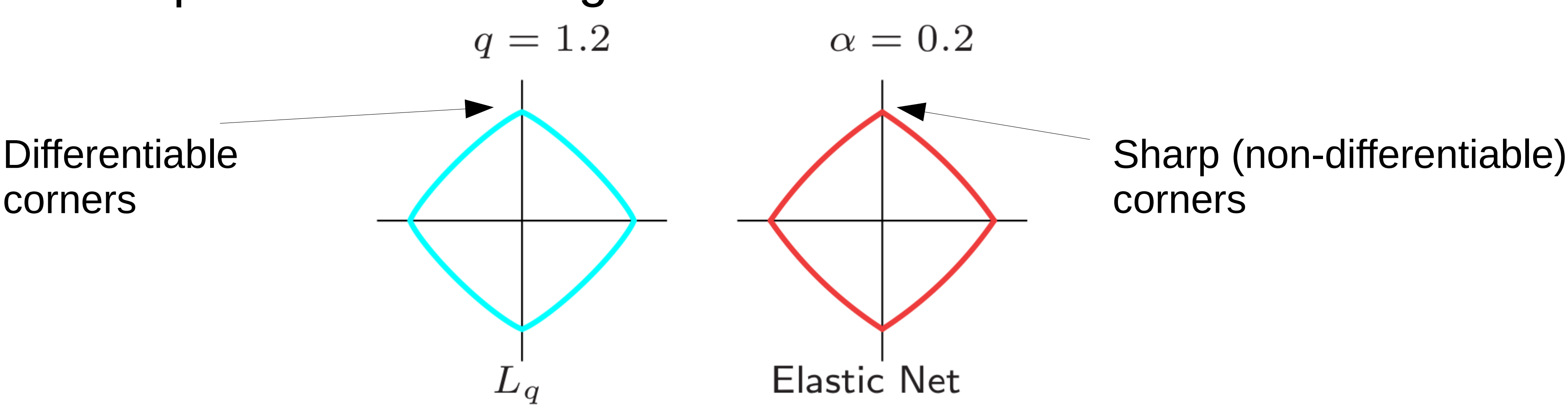
- One possibility is to use q in (1,2), such as q=1.2
- The elastic net penalty (Zou and Hastie, 2005)

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right),$$

Ridge Lasso

is a different compromise

- It **selects variable** like lasso, and shrinks together the coefficients of correlated predictors like ridge



Contours of constraint function

Exercise: Prediction on the prostate cancer dataset

See text of Exercise 4

References

[Hastie 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition). Springer. 2009.