STAT 432: Basics of Statistical Learning

Penalized Linear Regression

Shiwei Lan, Ph.D. <shiwei@illinois.edu>

http://shiwei.stat.illinois.edu/stat432.html

February 22, 2019

University of Illinois at Urbana-Champaign

Motivation

- Best subset selection
  - Computationally expensive; not feasible when p is large
- Forward/backward selection
  - No guarantee to find the best global sub-model
  - The selection process is discrete (“add” or “drop”). The result highly depends on the inclusion/exclusion criterion.

Motivation

- The OLS estimator is a linear function of $\mathbf{y}$, and it is the BLUE.
- Recall that the prediction accuracy is

$$\text{Irreducible Error} + \text{Bias}^2 + \text{Variance}$$

- Generally, by regularizing (shrinking, penalizing) the estimator in some way, we can create a new estimator
  - The estimator is biased
  - The variance is reduced
  - Overall, we can have a better prediction accuracy

Shrinkage Methods

Shrinkage Methods

- ℓ_2 penalty: Ridge regression
- ℓ_1 penalty: Lasso

Ridge Regression

Overview

- Definition of the Ridge regression
- How to derive the solution through connections with PCA?
- Effect of shrinkage and the degrees of freedom
- Selecting the tuning parameter

Ridge Regression

Penalizing the square of the coefficients

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

- proposed by Hoerl and Kennard (1970); Tikhonov (1943)
- $\lambda \geq 0$ is a tuning parameter (penalty level) that controls the amount of shrinkage
- penalizing the ℓ_2 norm of $\boldsymbol{\beta}$, hence is called the ℓ_2 penalty
- the coefficients $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ are shrunken towards 0

Solution for Ridge Regression

- We can also write the Ridge regression in matrix form:

$$\text{minimize} \quad (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}$$

- Similar to solving the linear regression, by taking the derivative of $\boldsymbol{\beta}$, we have the normal equation

$$\begin{aligned} \mathbf{0} &= -2\mathbf{X}^\mathsf{T}\mathbf{y} + 2\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \\ \implies \mathbf{X}^\mathsf{T}\mathbf{y} &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}\right)\boldsymbol{\beta} \\ \implies \quad \boldsymbol{\beta} &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \end{aligned}$$

- Why this helps fitting a linear model?

The Effect of Ridge Regression

- The Ridge regression is frequently used for addressing highly correlated variables
- When some variables are linearly correlated (e.g., $p > n$) $\mathbf{X}$ do not have full column rank
- This makes $\mathbf{X}^\mathsf{T}\mathbf{X}$ singular, hence inverting this matrix becomes impossible
- However, $\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}$ is always full ranked

The Effect of Ridge Regression

- Highly correlated variables makes the estimation unstable
- If $\mathbf{X}^\mathsf{T}\mathbf{X}$ is close to singular,

$$\det(\mathbf{X}^\mathsf{T}\mathbf{X}) \to 0 \quad \Rightarrow \quad \det((\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}) \to \infty$$

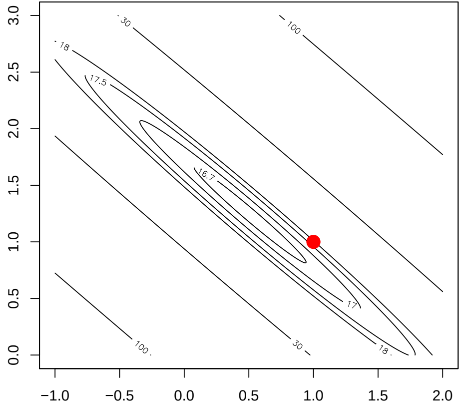
- Since $\text{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\sigma^2$, the variance of $\widehat{\boldsymbol{\beta}}$ (or certain combinations of $\widehat{\boldsymbol{\beta}}$) is extremely large.
- Trade that variance with some bias?

An Example

```
> library(MASS)
> set.seed(1)
> n = 30

> # highly correlated variables
> X = mvrnorm(n, c(0, 0), matrix(c(1,0.999, 0.999, 1), 2,2))
> y = rnorm(n, mean=1 + X[,1] + X[,2])

> # compare parameter estimates
> summary(lm(y~X))$coef
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   1.038007  0.1647551  6.300302 9.627026e-07
X1          -11.272638  4.6402098 -2.429338 2.205727e-02
X2           13.265586  4.6315269  2.864193 7.993486e-03

> # instead, the ridge regression
> lm.ridge(y~X, lambda=5)
                 X1        X2
1.1214448 0.8770568 0.9836474
```

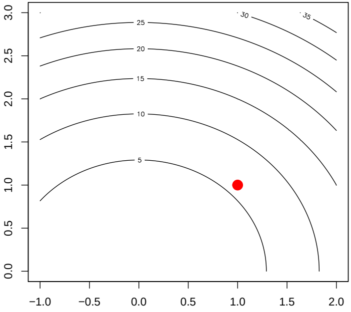
Optimization Point-of-view

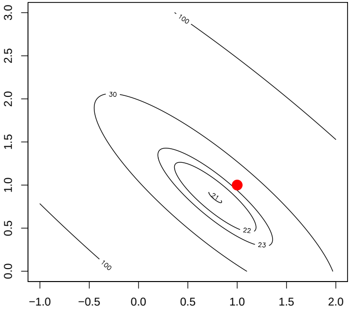
- The instability of having highly correlated variables can also be explained by the lack of convexity of the objective function
- The objective function of the OLS estimator is almost flat alone certain combinations of the $\boldsymbol{\beta}$ parameters
- The optimal solution is greatly affected by the random errors
- The Ridge penalty $\lambda\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}$ forces some convexity



OLS loss function $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

Linear Regression





Ridge penalty: $\lambda \boldsymbol{\beta}^\mathsf{T} \boldsymbol{\beta}$

Ridge objective function

Understanding the Shrinkage

- Suppose we have an orthonormal design matrix ($\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}$), then $\widehat{\boldsymbol{\beta}}^{\text{ols}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\mathbf{y}$ and

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{ridge}} =& (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} \\ =& (\mathbf{I} + \lambda\mathbf{I})^{-1}\widehat{\boldsymbol{\beta}}^{\text{ols}} \\ =& (1+\lambda)^{-1}\widehat{\boldsymbol{\beta}}^{\text{ols}},\end{aligned}$$

- This means that we just need to shrink each element of $\widehat{\boldsymbol{\beta}}^{\text{ols}}$ by a factor of $(1+\lambda)^{-1}$, i.e.,

$$\widehat{\boldsymbol{\beta}}_j^{\text{ridge}} = \frac{1}{1+\lambda}\widehat{\boldsymbol{\beta}}_j^{\text{ols}}, \quad \text{for all } j$$

Understanding the Shrinkage

- How about bias and variance under the orthonormal design
- $\text{Var}\big(\widehat{\boldsymbol{\beta}}_j^{\text{ridge}}\big) = \frac{1}{(1+\lambda)^2}\text{Var}\big(\widehat{\boldsymbol{\beta}}_j^{\text{ols}}\big)$ (reduced from OLS!)
- $\text{Bias}\big(\widehat{\boldsymbol{\beta}}_j^{\text{ridge}}\big) = \frac{-\lambda}{1+\lambda}\boldsymbol{\beta}_j$ (biased!)
- There always exists a λ such that the prediction error of $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ is smaller than $\widehat{\boldsymbol{\beta}}^{\text{ols}}$

Understanding the Shrinkage

- When the columns of $\mathbf{X}$ are not orthogonal, we can utilize PCA
- The relationship between Ridge and PCA can be understood by (assuming $\mathbf{X}$ centered) decomposing the covariance matrix

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

- This means $(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^\mathsf{T}$
- The Ridge fitted value $\widehat{\mathbf{y}}$ can be calculated as (since $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$)

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

Understanding the Shrinkage

- Hence, Ridge regression can be understood as
  (1) Perform principle component analysis of $\mathbf{X}$
  (2) Treat the principle components $\mathbf{u}_j$'s as new independent variables and project $\mathbf{y}$ onto the them: $\mathbf{u}_j^\mathsf{T}\mathbf{y}$ for each j
  (3) Shrink the projections using the factor $d_j^2/(d_j^2+\lambda)$
- Directions with smaller eigenvalues d_j get more relative shrinkage.
- The ridge fitted value of $\widehat{\mathbf{y}}$ is the sum of p shrunk projections.

Notes

- The Ridge regression solution is not invariant with respect to the scale of the predictors!
- The scale of variables determines d_j's, hence affect the shrinkage.
- A standard procedure: perform centering and scaling on $\mathbf{X}$, perform centering on $\mathbf{y}$, and fit linear regression on the normalized data without intercept. The parameters on the original scale can be reversely solved.
- The intercept term is not penalized.
- Some packages (e.g. "glmnet" package, and lm.ridge function in MASS package) handles the centering and scaling automatically.

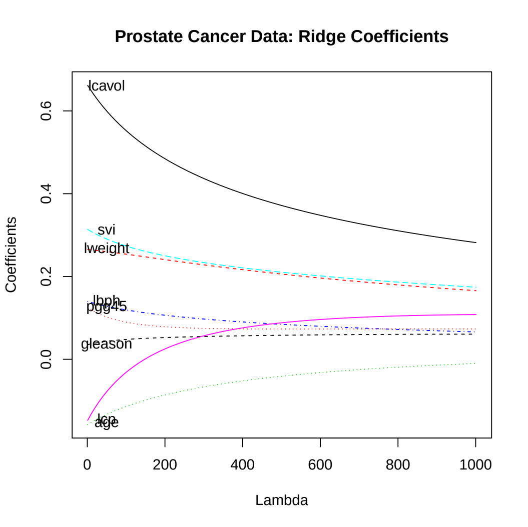
Tuning Parameter

- We need to tune the penalty term λ in a Ridge regression
- Cross-validation is possible, however, we also have some easier approach because Ridge regression, similar to linear regression, has some nice properties.
- The procedure is called GCV (generalized cross-validation)

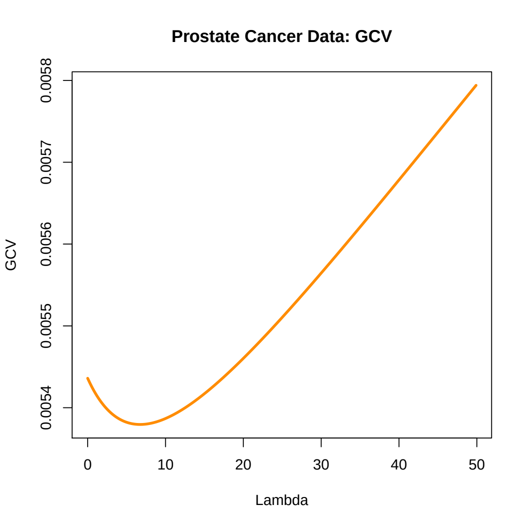
$$\text{GCV}(\lambda) = \frac{n^{-1}\|(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}\|^2}{\big(n^{-1}\text{Trace}(\mathbf{I} - \mathbf{S}_\lambda)\big)^2}$$

- GCV is motivated from the leave-one-out cross-validation. This is implemented in `lm.ridge`.

Prostate Cancer Example



Prostate Cancer Example



An Example

```
> library(ElemStatLearn)
> fit <- lm.ridge(lpsa~., prostate[, -10], lambda=seq(0,100,by=0.1))

> fit$lambda[which.min(fit$GCV)]

[1] 6.7

> round(fit$coef[, which.min(fit$GCV)], 4)

 lcavol lweight     age    lbph     svi     lcp gleason   pgg45
 0.5812  0.2580 -0.1255  0.1247  0.2839 -0.0593  0.0454  0.0968
```

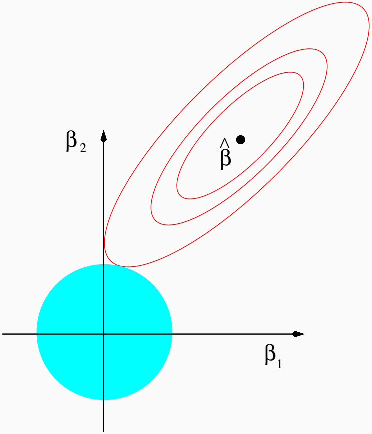
Alternative View

- An equivalent formulation is given by

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \\ \text{subject to} \quad & \sum_{j=1}^{p}\beta_j^2 \leq s \end{aligned}$$

- There is a one-to-one correspondence between the parameters λ and s, but we can't find the explicit formula.

Ridge Regression



Ridge constrained solution

Degrees of Freedom

- Although $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ is p-dimensional, it does not use the full potential of all p covariates due to the shrinkage.
- For example, if λ is very large, all the parameter estimates are 0. Then intuitively, the df should be close to 0. If λ is 0, then we reduce to the OLS with p df.
- The df of a Ridge regression is given by

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

which is always between 0 and p.

Lasso: Least Absolute Shrinkage and Selection Operator

Motivation

- The Ridge regression shrinks the coefficients towards 0, however, they are not exactly zero. Hence, we haven't achieve any "selection" of variables.
- Parsimony: we would like to select a small subset of predictions. Stepwise regression does not guarantee the global solution.
- Lasso provides a continuous process. We will discuss:
  - The formulation and convexity
  - The solution when $\mathbf{X}$ is orthogonal
  - Some examples

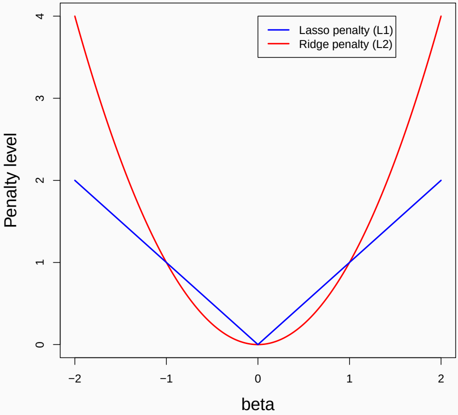
Lasso

Least absolute shrinkage and selection operator (Tibshirani 1996)

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- Shrinkage of the ℓ_1 norm of the parameters
- Property: some will be exactly 0, hence achieves selection of parameters

Lasso


Lasso penalty (L1)
Ridge penalty (L2)
Penalty level
beta
0
1
2
3
4
−2
−1
0
1
2

Lasso

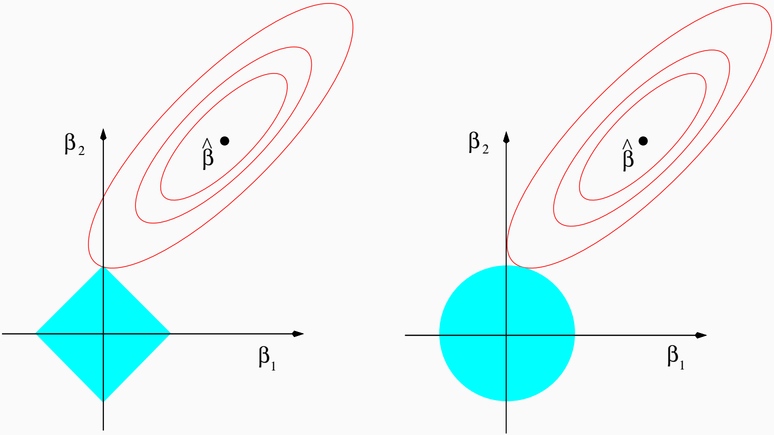
Equivalent Formulation

- The Lasso optimization problem is equivalent to

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} |\beta_j| \leq s \end{aligned}$$

- Each value of λ corresponds to an unique value of s.
- Compare Ridge and Lasso?

Ridge and Lasso



Comparing Lasso and Ridge solutions

Lasso Under Orthogonal Design

- Again, it will be helpful to view Lasso assuming orthogonal design, i.e., $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_{p\times p}$.
- We first analyze the loss part:

$$
\begin{aligned}
\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 =& \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}} + \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}} - \mathbf{X}\boldsymbol{\beta}\|^2 \\
=& \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}}\|^2 + \|\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}} - \mathbf{X}\boldsymbol{\beta}\|^2
\end{aligned}
$$

- The cross-product term is

$$
2(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}})^\mathsf{T}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}} - \mathbf{X}\boldsymbol{\beta}) = 2\mathbf{r}^\mathsf{T}(\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ols}} - \mathbf{X}\boldsymbol{\beta}) = 0,
$$

  since the second term is in the column space of $\mathbf{X}$, while $\mathbf{r}$ is orthogonal to that space.

Lasso Under Orthogonal Design

- Our Lasso problem can be rewritten as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\textsf{lasso}} &= \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ &= \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\textsf{ols}}\|^2 + \|\mathbf{X}\widehat{\boldsymbol{\beta}}^{\textsf{ols}} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1\end{aligned}$$

- Since $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\textsf{ols}}\|^2$ is not a function of $\boldsymbol{\beta}$, this problem is reduced to

$$\widehat{\boldsymbol{\beta}}^{\textsf{lasso}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\widehat{\boldsymbol{\beta}}^{\textsf{ols}} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Lasso Under Orthogonal Design

- Then, since $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_{p\times p}$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\,\text{lasso}} &= \arg\min_{\boldsymbol{\beta}} \ \|\mathbf{X}\widehat{\boldsymbol{\beta}}^{\,\text{ols}} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}} \ (\widehat{\boldsymbol{\beta}}^{\,\text{ols}} - \boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\widehat{\boldsymbol{\beta}}^{\,\text{ols}} - \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}} \ (\widehat{\boldsymbol{\beta}}^{\,\text{ols}} - \boldsymbol{\beta})^\mathsf{T}(\widehat{\boldsymbol{\beta}}^{\,\text{ols}} - \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 \\
&= \arg\min_{\boldsymbol{\beta}} \ \sum_{j=1}^{p}(\widehat{\beta}_j^{\,\text{ols}} - \beta_j)^2 + \lambda|\beta_j|.
\end{aligned}$$

- Note that each β_j is involved in a separate term, we can solve the lasso estimators individually from the OLS estimators.

Lasso Under Orthogonal Design

- Each of the β_j's is essentially solving for an optimization problem

$$\arg\min_{\beta} \, (\beta - a)^2 + \lambda|\beta|, \quad \lambda > 0$$

- The solution is simply

$$\begin{aligned}
\widehat{\beta}_j^{\text{lasso}} &= \begin{cases} \widehat{\beta}_j^{\text{ols}} - \lambda/2 & \text{if} \quad \widehat{\beta}_j^{\text{ols}} > \lambda/2 \\ 0 & \text{if} \quad |\widehat{\beta}_j^{\text{ols}}| \leq \lambda/2 \\ \widehat{\beta}_j^{\text{ols}} + \lambda/2 & \text{if} \quad \widehat{\beta}_j^{\text{ols}} < -\lambda/2 \end{cases} \\
&= \text{sign}\big(\widehat{\beta}_j^{\text{ols}}\big)\Big(|\widehat{\beta}_j^{\text{ols}}| - \lambda/2\Big)_+
\end{aligned}$$

- A large λ will shrink some of the coefficients to exactly zero, which achieves "variable selection".

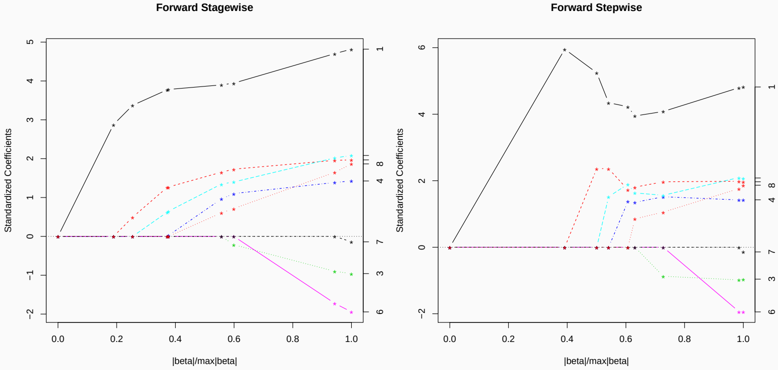
Computation of Lasso Solution

- When the covariates are not orthogonal, we will not be able to write down the explicit solution
- The Lasso problem is convex, although it may not be strictly convex in $\boldsymbol{\beta}$ when p is large
- The solution is a global minimum, but may not be unique

Computation of Lasso Solution

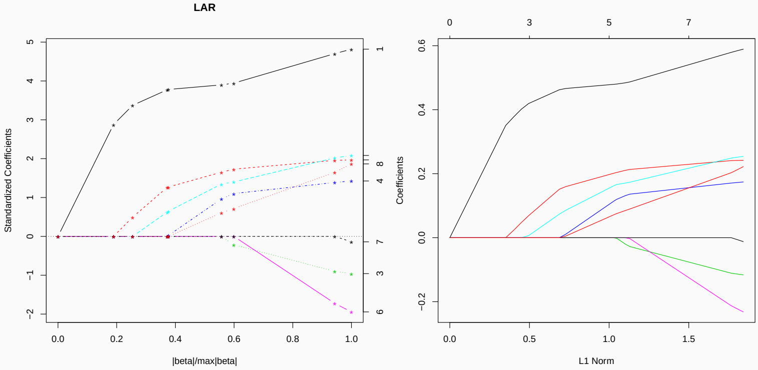
- There are algorithms that will produce equivalent solutions, although their computational costs are not the same
- Stage-wise regression (what is this?) Read ESL 3.3.3.
- Least angle regression (Efron et al. 2004) Read ESL 3.4.4.
- Coordinate descent (Friedman et al 2010): The most popular and fastest implementation, glmnet package
    - Also provides the solution path for an entire sequence of λ values
    - Start with the largest λ, use the previous estimation of β as a warm start for the solution of smaller λ

Lasso



Comparing stagewise regression with stepwise regression

Lasso



Comparing least angle regression with coordinate descent

ℓ_q Penalties

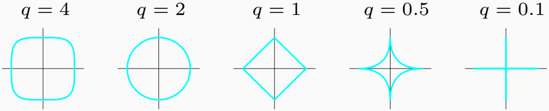


FIGURE 3.12. *Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q.*

- Ridge is ℓ_2 penalty
- Lasso is ℓ_1 penalty
- Best subset is ℓ_0 penalty
- Elastic-net is a combination of Lasso and Ridge:

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$$

R Functions

- Use R help and R manuals
- Linear models: function lm
- Ridge regression:
  - package MASS ; function lm.ridge
  - package glmnet ; function glmnet and cv.glmnet with alpha = 0
- Lasso:
  - package lars ; function lars
  - package glmnet ; function glmnet and cv.glmnet with alpha = 1
- Read more in `ISL` Ch 6.2.1. Check [ESL Video].