Shrinkage Methods of Linear Regession
: Ridge and LASSO

연세대학교 통계데이터사이언스 석사과정 이청파 (leechungpa@naver.com)

Review

Test (=generalization) **error** : predictoin error over an independent test sample (X, Y)

$$Err_\tau = E_{X,Y}[L(Y, \hat{f}_\tau(X))|\tau]$$

Expected prediction error : The randomness in the training set τ is averaged over.

$$Err = E_{X,Y}[Err_\tau] = E_{X,Y}[L(Y, \hat{f}_\tau(X))]$$

Training error : optimistic estimate of the test error Err_τ

$$e\bar{r}r = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}_\tau(x_i))$$

In-sample error

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}[L(y_i^0, \hat{f}_\tau(x_i))|\tau]$$

Linear model case : $\hat{f}_\tau(x) = x^T\hat{\beta}$ **where** $Y = f(X) + \epsilon$

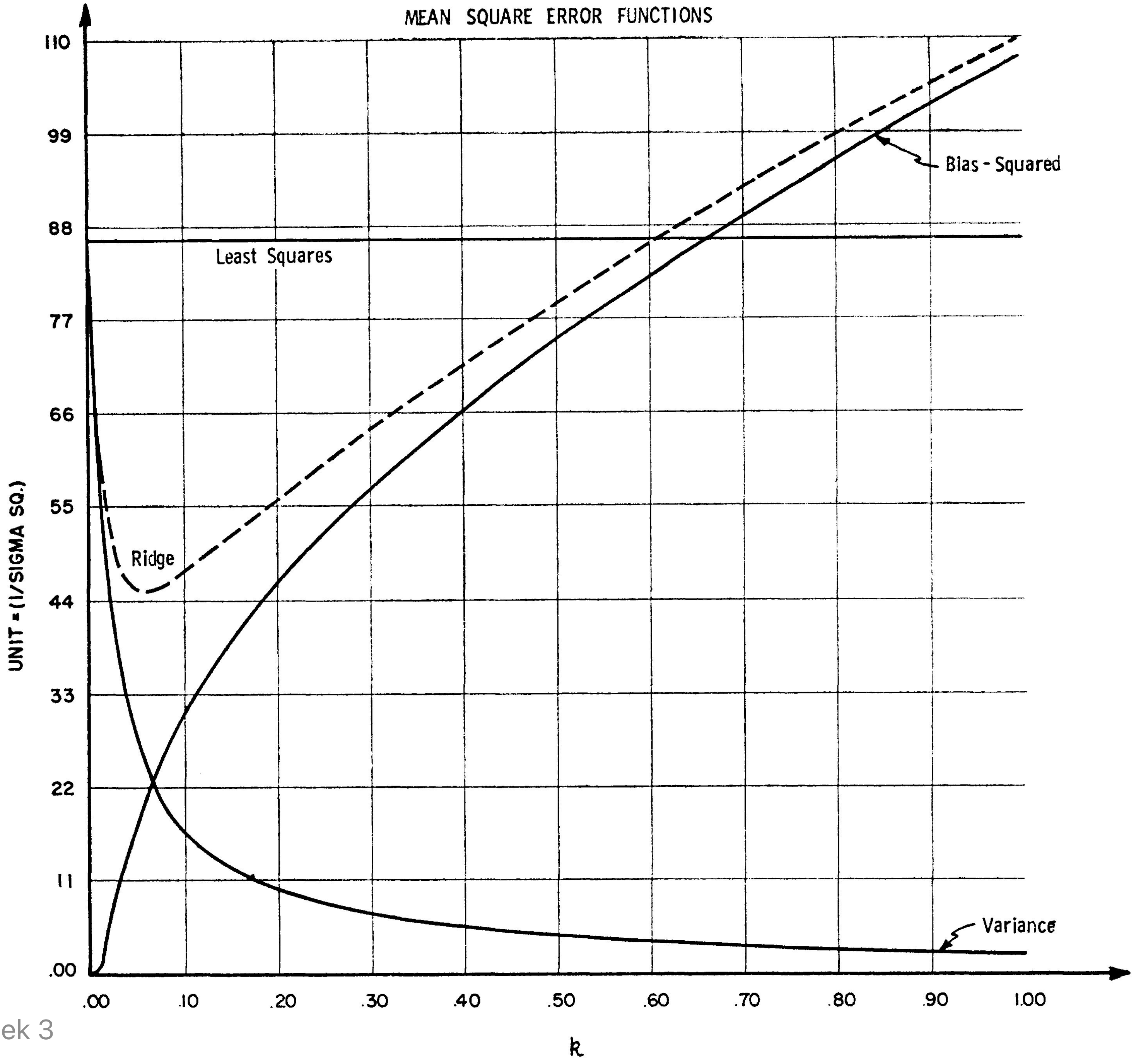
Expected prediction error of a linear model with L^2 loss function at $X = x_0$:

$$
\begin{aligned}
Err(x_0) &= E_Y[L(Y, \hat{f}_\tau(X))|X = x_0] \\
&= \sigma_\epsilon^2 + \left\{f(x_0) - E[x_0^T\hat{\beta}]\right\}^2 + \left\|X(X^TX)^{-1}x_0\right\|_2^2\sigma_\epsilon^2
\end{aligned}
$$

In above linear regeression case, we can decompose the average squared bias.

$$
E_{x_0}\left\{f(x_0) - E[x_0^T\hat{\beta}]\right\}^2 = E_{x_0}\underbrace{\left\{f(x_0) - x_0^T\hat{\beta}^{\text{BLUE}}\right\}^2}_{\text{model bias}} + E_{x_0}\underbrace{\left\{x_0^T\hat{\beta}^{\text{BLUE}} - E[x_0^T\hat{\beta}]\right\}^2}_{\text{estimation bias}}
$$

- $\hat{\beta}^{\text{BLUE}}$: BLUE of linear regression ($\hat{\beta} = \arg\min_\beta \left\|y - X^T\beta\right\|_2^2$)
- When the linear model is ordinary least squres method, the estimation basis is 0.


MEAN SQUARE ERROR FUNCTIONS
UNIT = (1/SIGMA SQ.)
110
99
88
77
66
55
44
33
22
11
.00
.00
.10
.20
.30
.40
.50
.60
.70
.80
.90
1.00
k
Least Squares
Ridge
Bias - Squared
Variance

Pros and cons of subset selection

Subset selection : best-subset, foward-stepwise, backward-stepwise selection, etc

$$\hat{\beta}^{\text{best-subset}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_0 \right\} \qquad \text{where } \|u\|_p = \sum_{i=1}^{N} |u_i|^0 \ (= \sum_{i=1}^{N} I(u_i \neq 0))$$

- Produces an interpretable model
- (Possibly) lower prediction error than full model

However,

- Subset selection is discrete process.
    - Variables are either retainded or discarded.
- Correlated variables often exhibits high variance.
    - When many there are many correlated vaiables, a widely large positive coefficient on one variable can be canceled by a similaraly large negative coefficient on its correlated.
    - High variance results to high prediction error than full model.

$\Longrightarrow$ More continuous and lower variability shrinkage methods are needed.

Ridge

$$(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}^{\text{ridge}}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 \right\} \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t$$

$$\iff \quad (\hat{\beta}_0^{\text{ridge}}, \hat{\beta}^{\text{ridge}}) = \underset{(\beta_0,\beta)\in\mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$\overset{\text{standardized}}{\iff} \quad \hat{\beta}^{\text{ridge}} = \underset{\beta\in\mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 \right\} \quad \text{subject to } \|\beta\|_2^2 \leq t$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (=0)$$

$$\overset{\text{standardized}}{\iff} \quad \hat{\beta}^{\text{ridge}} = \underset{\beta\in\mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\}$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (=0)$$

Cf. standard l^p-norm : $\|u\|_p = \left(\sum_{i=1}^{N} |u_i|^p \right)^{1/p}$

Ridge shrinks the regression coefficients by imposing a penalty on their size.

$$(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}^{\text{ridge}}) = \underset{(\beta_0, \beta) \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- $\lambda \geq 0$: complexity parameter
    - controls the amount of shirnkage
    - large λ shrinks coefficients toward 0
    - λ and t have a one-to-one correspondence
- β_0 is not included in the penalty term.
    - Penalization of the intercept would make the regression depend on the origin. (Consider simple regression with a negative slope.)

If we standarize X and y,

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right\} \\ \hat{\beta}_0^{\text{ridge}} &= \bar{y} \;\; (=0)\end{aligned} \implies \hat{\beta}^{\text{ridge}} = (X^TX + \lambda I)^{-1}X^Ty$$

- Standarizeds the inputs before solving.
  - The solutions are not equivariant under scaling of the inputs.
  - So X is $N \times p$ matrix, not $N \times (p+1)$.
  - $\hat{\beta}_0^{\text{ridge}}$ must be $\bar{y}\ (=0)$. (DIY)
- Ridge was first introduced in statisitcs to make X^TX nonsingular (to solve linear regression), even if X^TX is not of full rank.

Interpretations of Ridge : Singular Value Decomposition

X is decomposed by orthogonal matrices U, V and a diagonal matirx D.

$$X = UDV^T \qquad \Longrightarrow \qquad \begin{aligned} X\hat{\beta}^{\text{ridge}} &= X(X^TX + \lambda I)^{-1}X^Ty \\ &= UD(DD + \lambda I)^{-1}DU^Ty \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^T y \end{aligned}$$

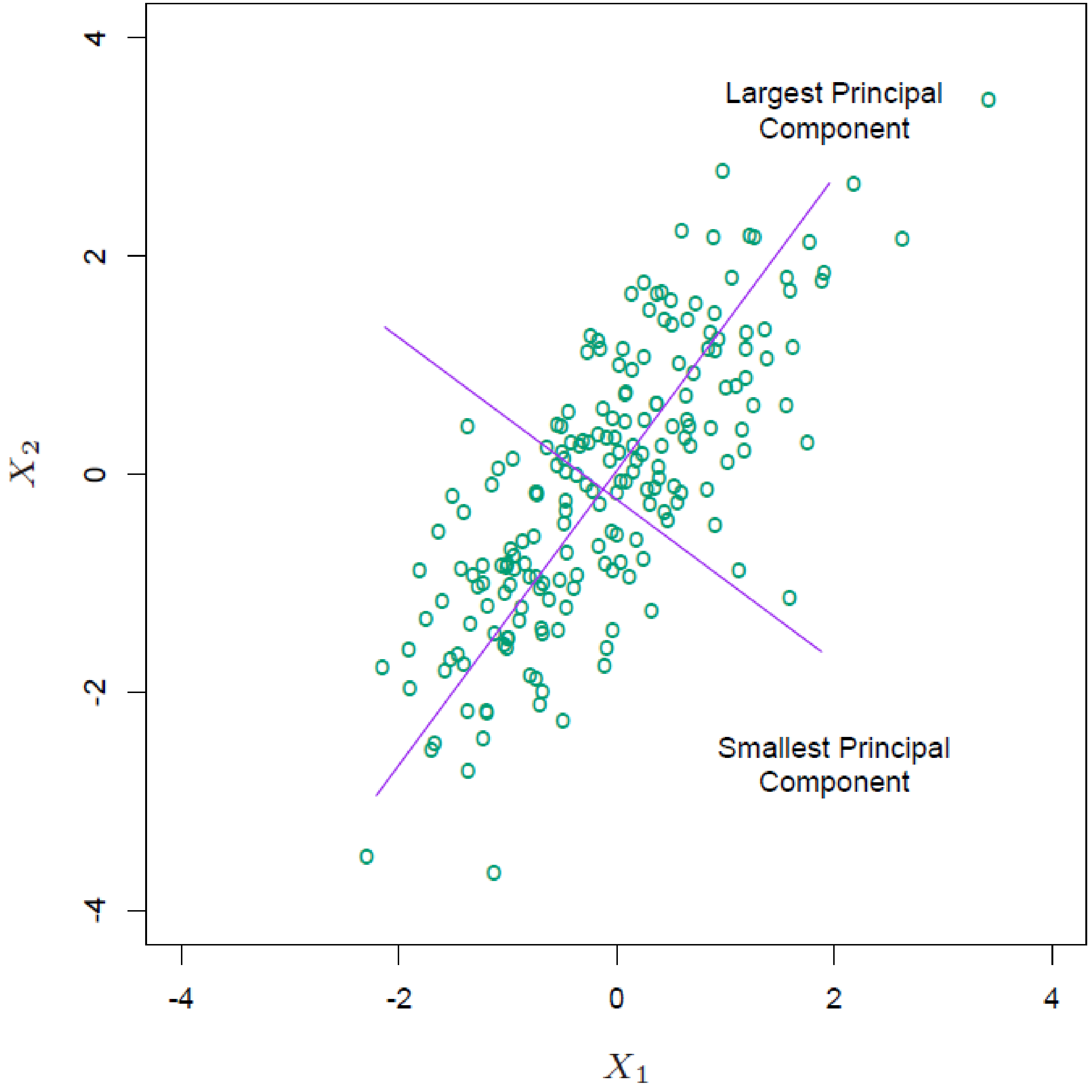
- $\mathbf{u}_j$ are the columns of U, which spans the column space of X.
- Comparing to linear regression ($X\hat{\beta} = UU^Ty$), $\frac{d_j^2}{d_j^2+\lambda}$ shrinkages j^{th} coordinate.

$$X^T X = VDDV^T$$

By using the pricipal componets of the variables X, there are eigen- vectors $\mathbf{v}_i$ also called the i^{th} principal component direction of X.

$$Var(\mathbf{z}_i) = Var(X\mathbf{v}_i) = \frac{d_i^2}{N}$$

The small singular values d_j has small variance, and ridge regression shrinks these directions the most.



The configuration of the data allow us to determine its gradient more accurately in the long direction than the short.

Ridge regression protects against the potentially high variance of gradients estimated in the short directions.

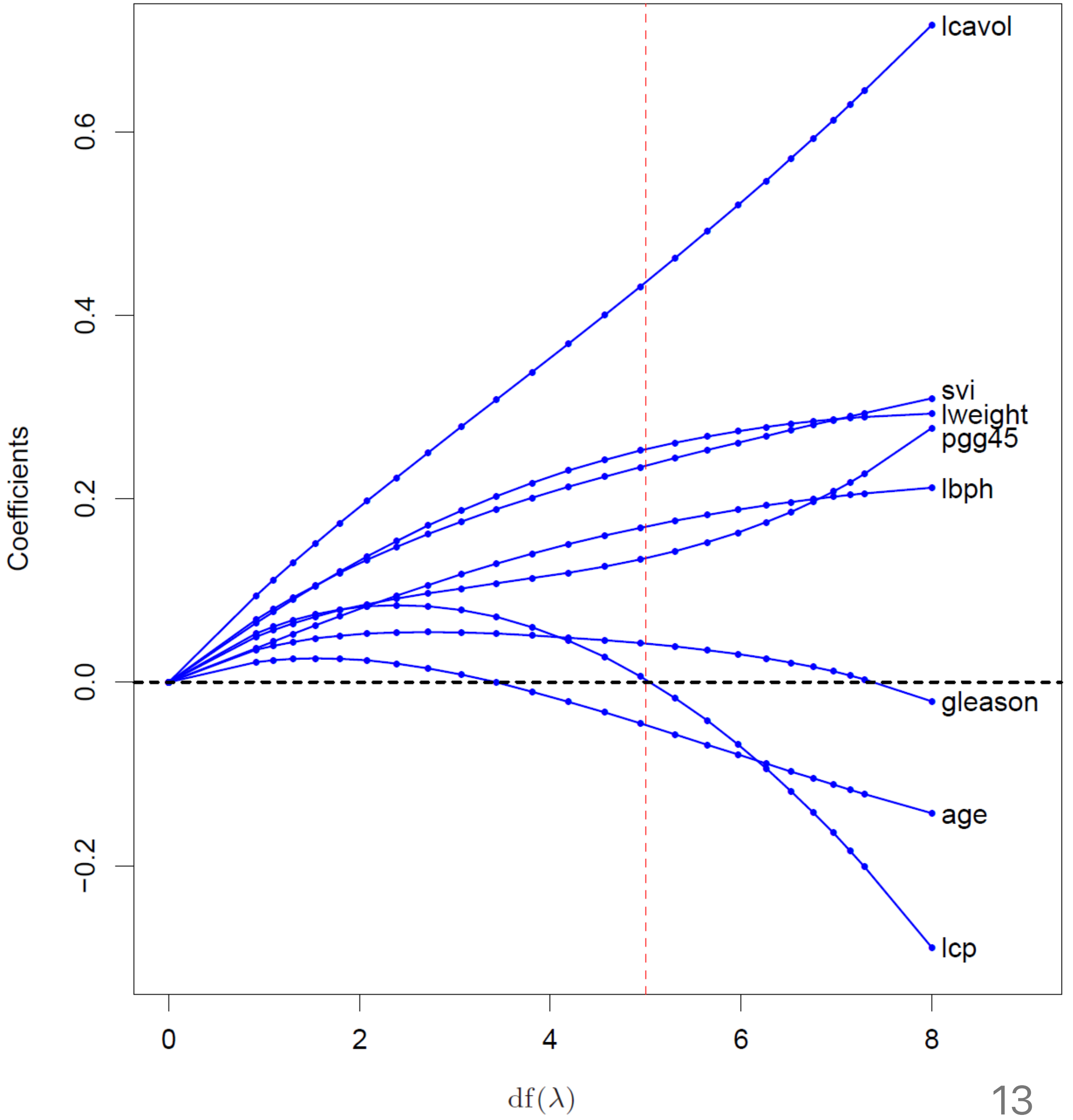
> **Implicit assumption of ridge :**
> **The response tend to vary most in the directions of high variance of inputs.**

More details about PCA will be on next week.

Effective degrees of freedom

$$
\begin{aligned}
df(\lambda) &\overset{\text{def}}{=} trace[H_\lambda^{\text{ridge}}] \\
&= trace[X(X^T X + \lambda I)^{-1} X^T] \\
&= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}
\end{aligned}
$$

- $df(\lambda)$ of linear regression is the number of variables p.
  - $df(\lambda) = p$ when no regularization $\lambda = 0$.



LASSO

Least Absolute Shrinkage and Selection Operator

$$(\hat{\beta}_0^{\text{lasso}}, \hat{\beta}^{\text{lasso}}) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \quad \textbf{subject to} \sum_{j=1}^{p} |\beta_j| \le t$$

$$\overset{\text{standardized}}{\Longleftrightarrow} \hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 \right\} \quad \textbf{subject to} \ \|\beta\|_1 \le t$$

$$\hat{\beta}_0^{\text{lasso}} = \bar{y} \ \ (= 0)$$

$$\overset{\text{standardized}}{\Longleftrightarrow} \hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \ \ (= 0)$$

If we standarize X and y,

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$
$$\hat{\beta}_0^{\text{ridge}} = \bar{y} \;\; (= 0)$$

- Standarizeds the inputs before solving, likewise Ridge.
- LASSO use L^1 penalty instead of L^2
    - The solution of LASSO is nonlinear in the y. (No closed form)
    - The shrinkage method of LASSO makes the coefficient exactly 0, unlike Ridge.
    - The lasso does a kind of continuous subset selection.

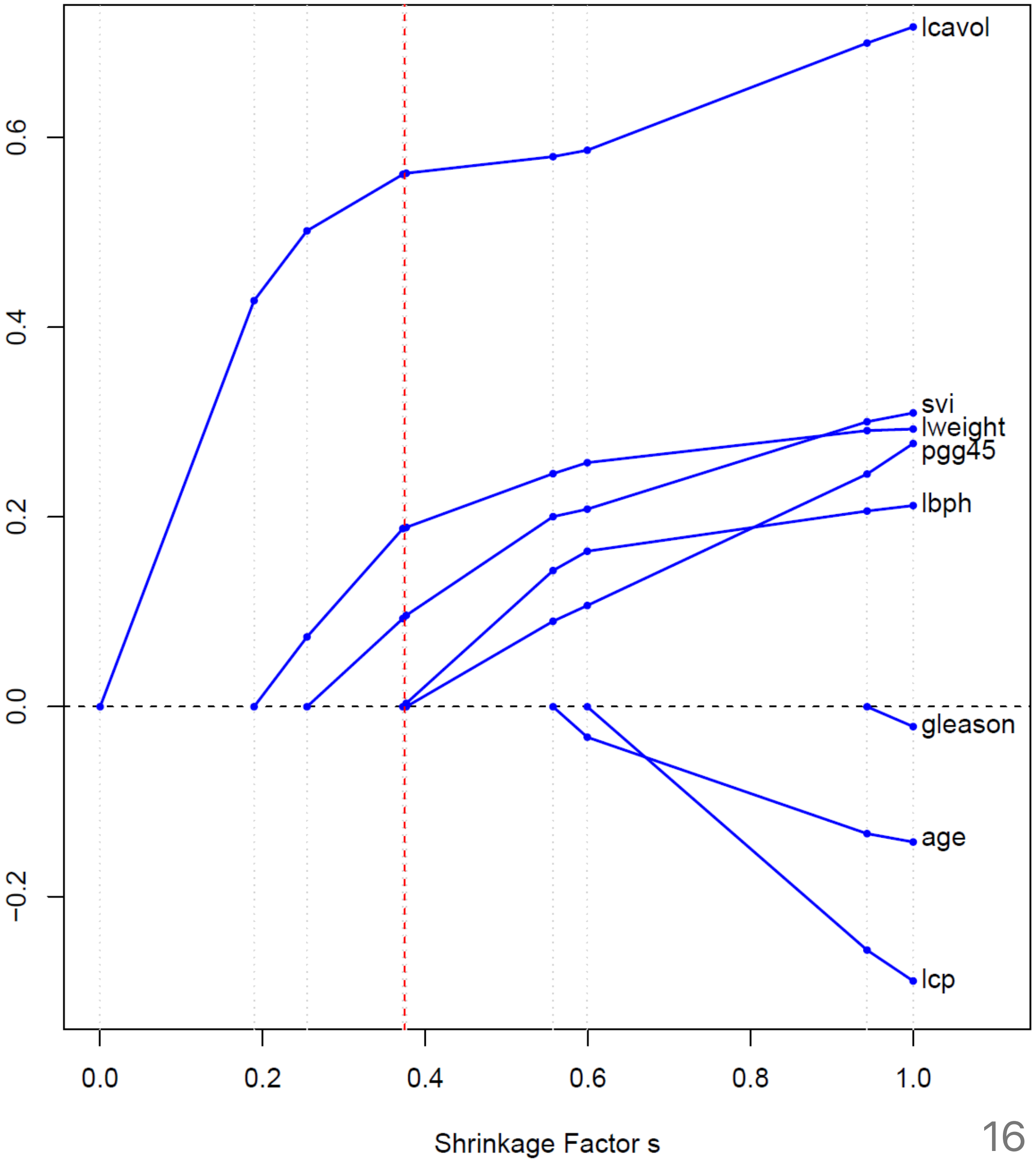
Interpretations of LASSO
: Standardized shrinkage factor

Recall LASSO subject to $||\beta||_1 \leq t$.

If we choose t larger than $||\hat{\beta}||_1$ where $\hat{\beta}$ is LSE, then $\hat{\beta}^{\text{lasso}} = \hat{\beta}$.

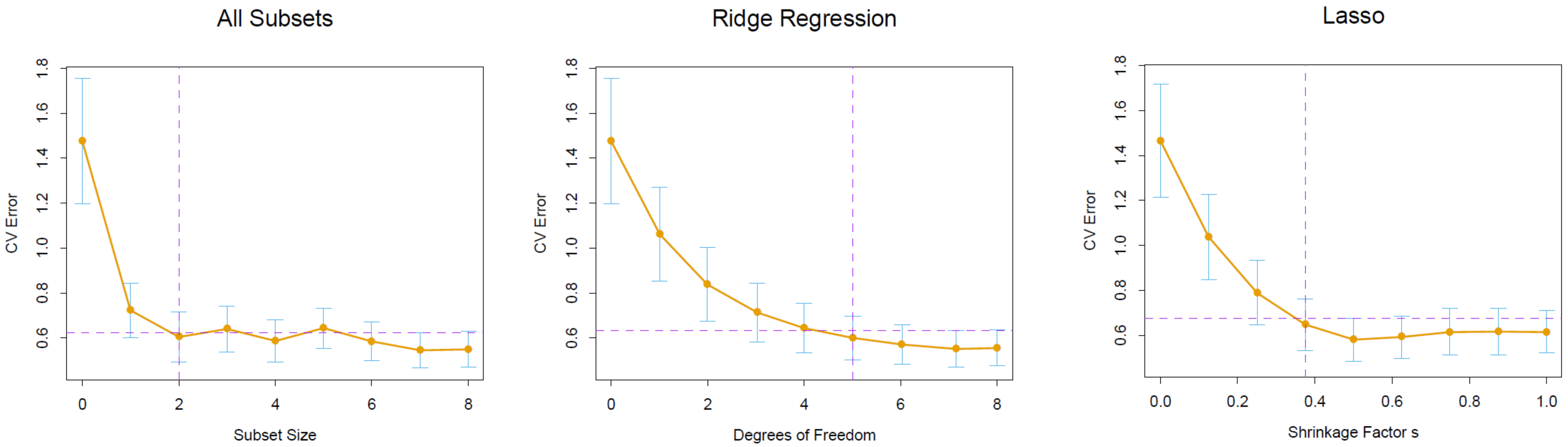
$$s \overset{\text{def}}{=} \frac{t}{||\hat{\beta}||_1}$$

Using maximum, standardize the shrinkage factor like above.



Compare 3 methods : subset selection, ridge, lasso

Above 3 methods should adaptively choose λ (or p, t) to minimize an estimate of expected prediction error.



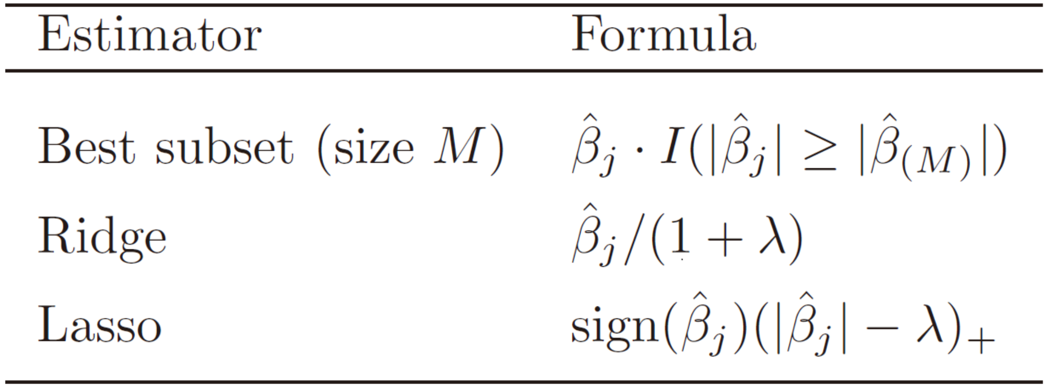
All Subsets

Ridge Regression

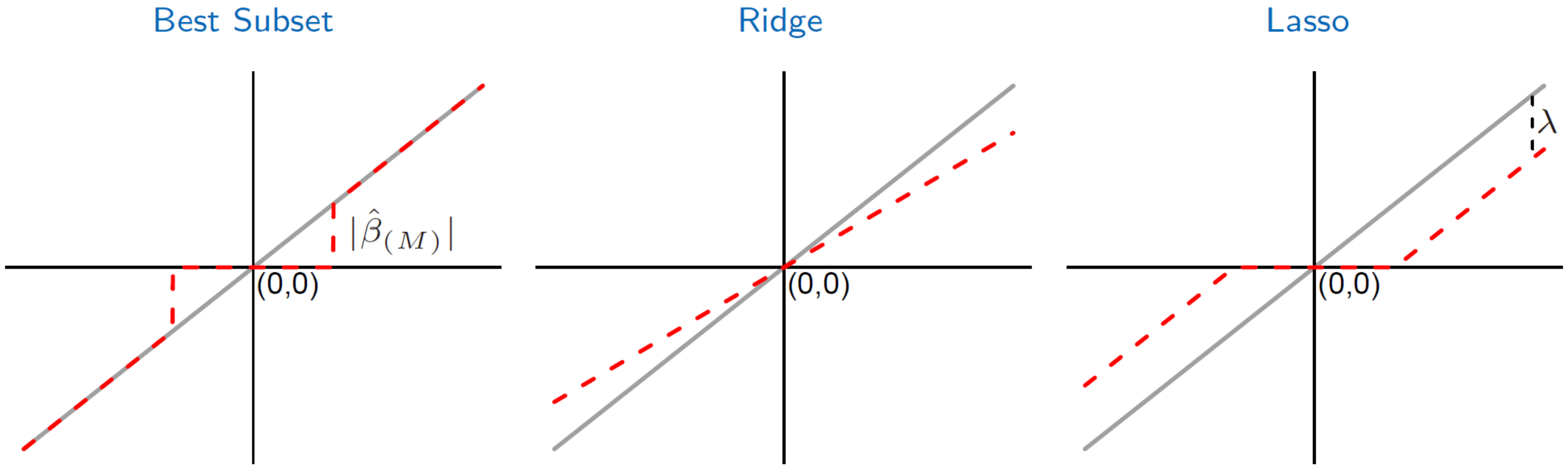
Lasso

Case of an orthonormal input matrix X :

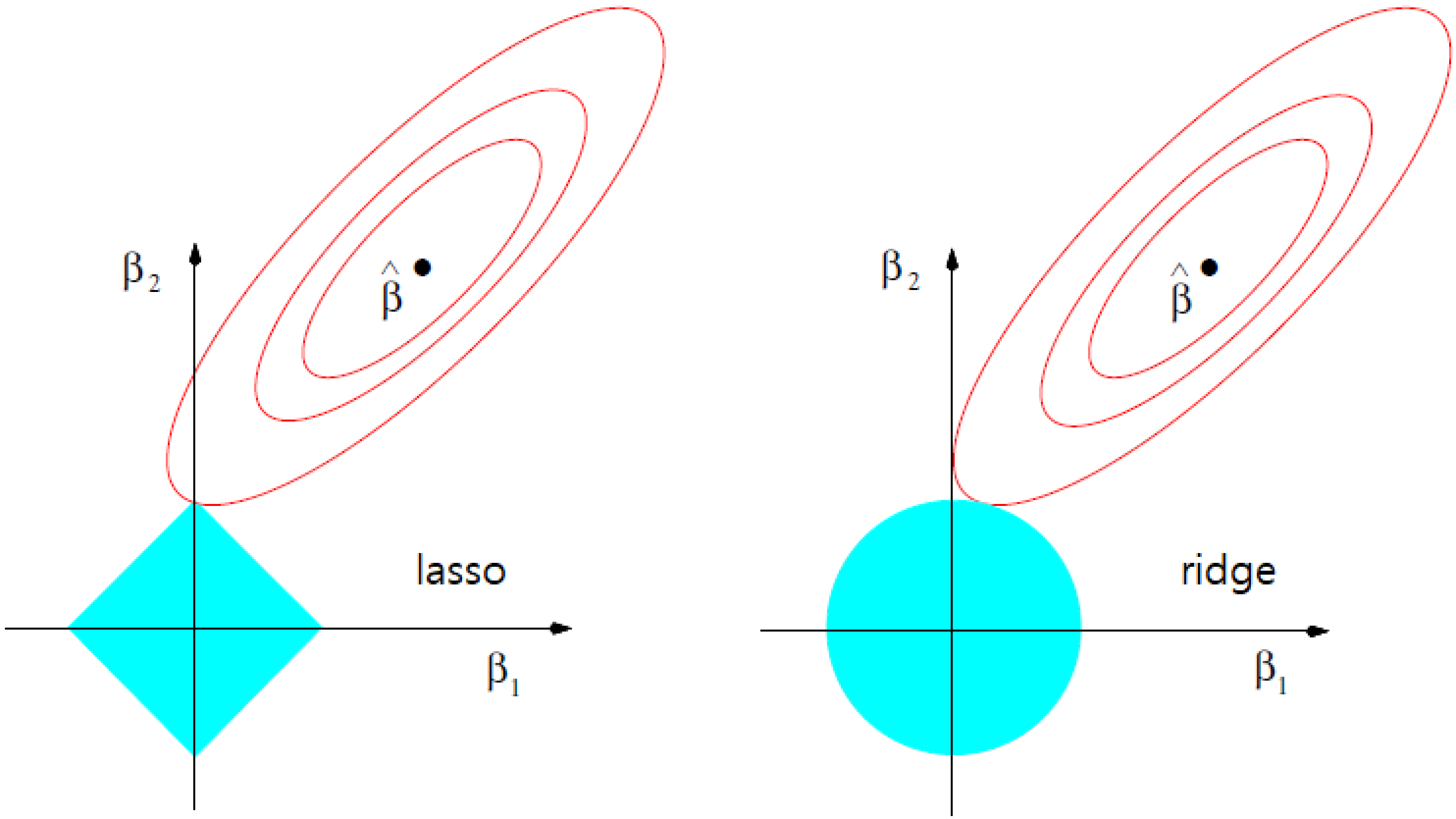
In an orthonomral case, there are explicit solutions.



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



Case of a non-orthogonal input matrix X :



The solid blue areas are the constraint regions, while the red ellipses are the contours of the least squares error function.

Generalize version of regularization regression

For given $q \geq 0$,

$$(\hat{\beta}_0^{q-\text{norm}}, \hat{\beta}^{q-\text{norm}}) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \quad \text{subject to } \sum_{j=1}^{p} |\beta_j|^q \leq t$$

$$\overset{\text{standardized}}{\iff} \quad \hat{\beta}^{\text{q-norm}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \right\}$$

$$\hat{\beta}_0^{\text{q-norm}} = \bar{y} \ \ (= 0)$$

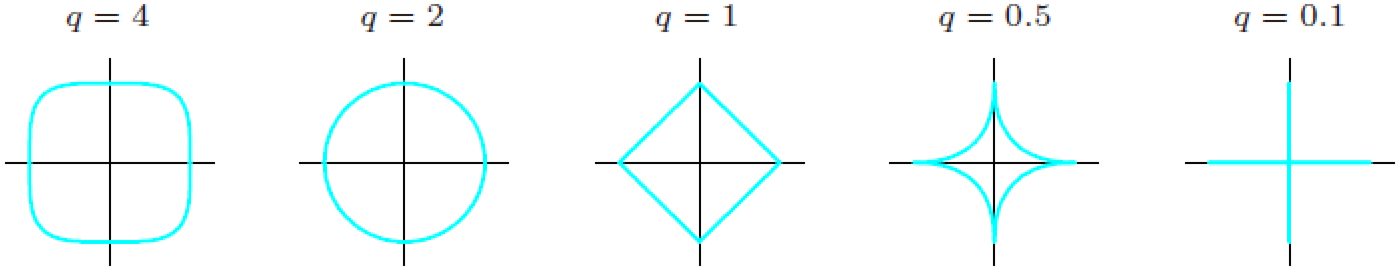
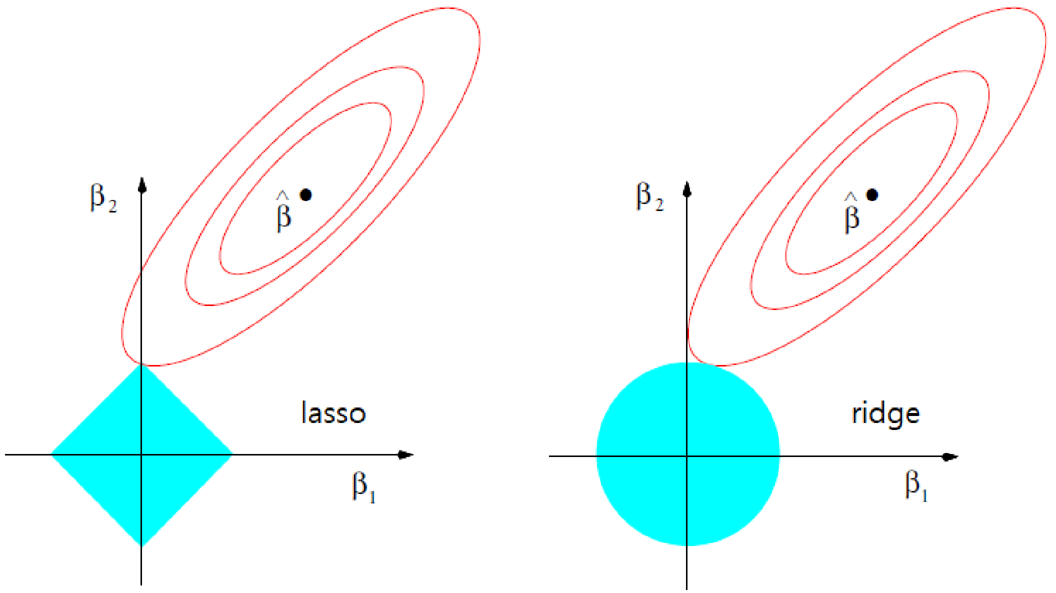


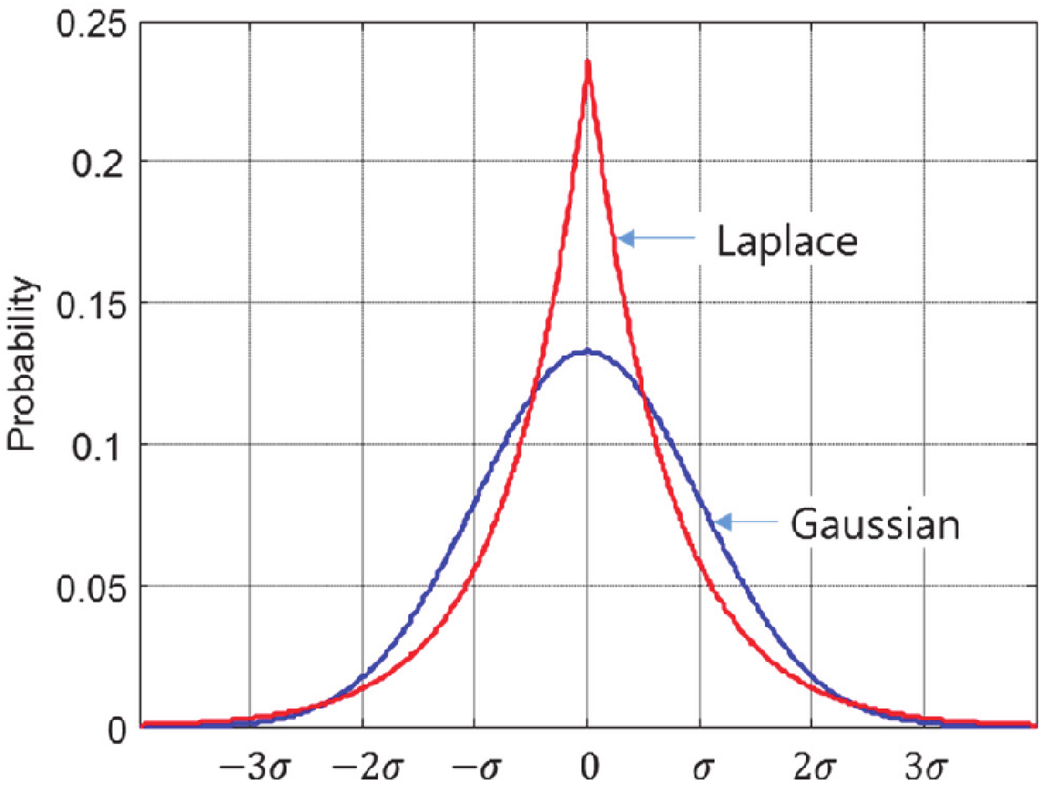
FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Bayesian approaches of Generalize version of ridge and lasso

- log-prior : $\log \beta^{\text{q-norm}} \sim \lambda \|\beta\|_q^q$
  - If $q = 0$ (best-subset selection), $\beta^{\text{q-norm}} \sim exp\{\lambda\|\beta\|_0\}$
  - If $q = 1$ (LASSO), $\beta^{\text{q-norm}} \sim exp\{\lambda\|\beta\|_1\}$ (Laplace)
    ($q = 1$ is the smallest q such that the constraint region is convex.)
  - If $q = 2$ (ridge), $\beta^{\text{q-norm}} \sim exp\left\{\lambda\|\beta\|_2^2\right\}$ (Gaussian)
- likelihood : $(X, y)|\beta \sim exp\left\{\|y - X\beta\|_2^2\right\}$ (Gaussian)
- log-posterior : $\log \beta^{\text{q-norm}}|(X, y) \sim \left\{\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q\right\}$

$$\implies \log \hat{\beta}^{\text{q-norm}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\log \beta^{\text{q-norm}}|(X, y)\}$$





- If $q = 1$ (LASSO), $\beta^{\text{q-norm}} \sim exp\left\{\lambda\|\beta\|_1\right\}$ (Laplace)
- If $q = 2$ (ridge), $\beta^{\text{q-norm}} \sim exp\left\{\lambda\|\beta\|_2^2\right\}$ (Gaussian)

$\Longrightarrow$ 3 methods are Bayes estimates with different priors.

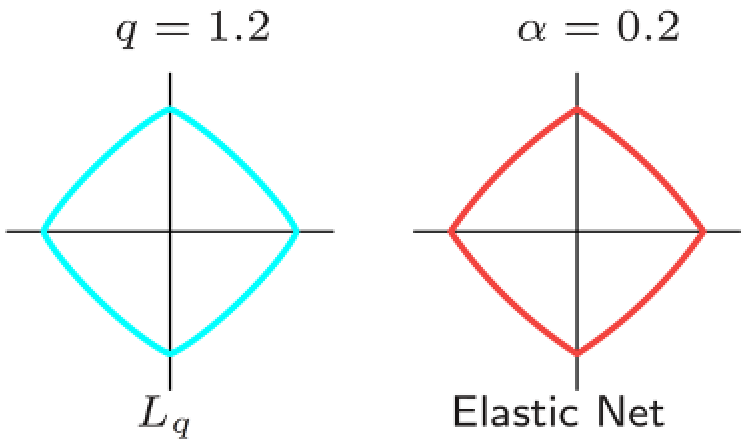
Elastic net

$$(\hat{\beta}_0^{\text{elastic}}, \hat{\beta}^{\text{elastic}}) = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \quad \text{subject to } \sum_{j=1}^{p} (\alpha \beta_j + (1-\alpha)|\beta_j|) \le t$$

$$\overset{\text{standardized}}{\iff} \quad \hat{\beta}^{\text{elastic}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \|y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_2^2 + (1-\alpha)\|\beta\|_1 \right) \right\}$$

$$\hat{\beta}_0^{\text{elastic}} = \bar{y} \;\; (= 0)$$

- Values of $q \in (1, 2)$ suggest a compromise between the lasso and ridge regression.
  - Although $||\beta||_q^q$ for $q \in (1, 2)$ is differentiable at 0, the regularization does not share the ability of lasso, setting coefficients exactly to zero.
- The elastic-net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.



- Above contours are the constant value of $q = 1.2$ norm regularization (left), and the elastic-net for $\alpha = 0.2$ (right).
- The elastic-net has sharp (non-differentiable) corners, while the $q = 1.2$ penalty does not.
- It also has considerable computational advantages over the L^q penalties.

Homework

Exercise. 3.29

Suppose we fit a ridge regression with a given shrinkage parameter $\lambda \in \mathbb{R}^+$ on a single variable x_1. (Notice that x_1 is a $N \times 1$ vector.)

1. (Essential) Show that the coefficient must be $\frac{X^T y}{X^T X + \lambda}$ where $X = x_1$.
2. (Essential) We now include an exact copy $x_2 = x_1$, so our new design matrix would be $X = [x_1 | x_2]$. Using this matrix, re-fit our ridge regression. Show that both coefficients are identical, and derive their value.
3. (Extra) Show in general that if m copies of a variable x_j, are included in a ridge regression, so X would be $[x_1 | x_2 | \cdots | x_m]$, their coefficients are all the same.

References

- The Elements of Statistical Learning (2008) Trevor Hastie, Robert Tibshirani, Jerome Friedman
- Ridge Regression: Biased Estimation for Nonorthogonal Problems (1970) Arthur E. Hoerl, Robert W. Kennard