Linear Methods for Regression

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Outline

Introduction

Linear Regression Models and Least Squares

Subset Selection

Shrinkage Methods

Methods Using Derived Input Directions

Discussions

Summary

Introduction

- Let $X = [X_1, \ldots, X_p]^T$ be a data point, a linear regression model assumes

$$E(Y|X)$$

is a linear function of $X_1, \ldots, X_p$

- Advantages
  - They are simple and often provide an adequate and interpretable description
  - They can sometimes outperform nonlinear models
    - Small numbers of training cases, low signal-to-noise ratio or sparse data
  - Linear methods can be applied to transformations of the inputs

Outline

- Introduction
- **Linear Regression Models and Least Squares**
- Subset Selection
- Shrinkage Methods
- Methods Using Derived Input Directions
- Discussions
- Summary

Linear Regression Models

The Linear Regression Model

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- β_j's are unknown coefficients

The variable X_j could be

- Quantitative inputs
- Transformations of quantitative inputs
  - Log, square-root or square
- Basis expansions ($X_2 = X_1^2, X_3 = X_1^3$)
- Numeric coding of qualitative inputs

Least Squares

Given a set of training data $(x_1, y_1) \cdots (x_N, y_N)$ where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$

Minimize the Residual Sum of Squares

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$\ldots_j\Big)^2 = \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta$$

Valid if the y_i's are conditionally independent given the inputs x_i

A Geometric Interpretation

$p = 2$

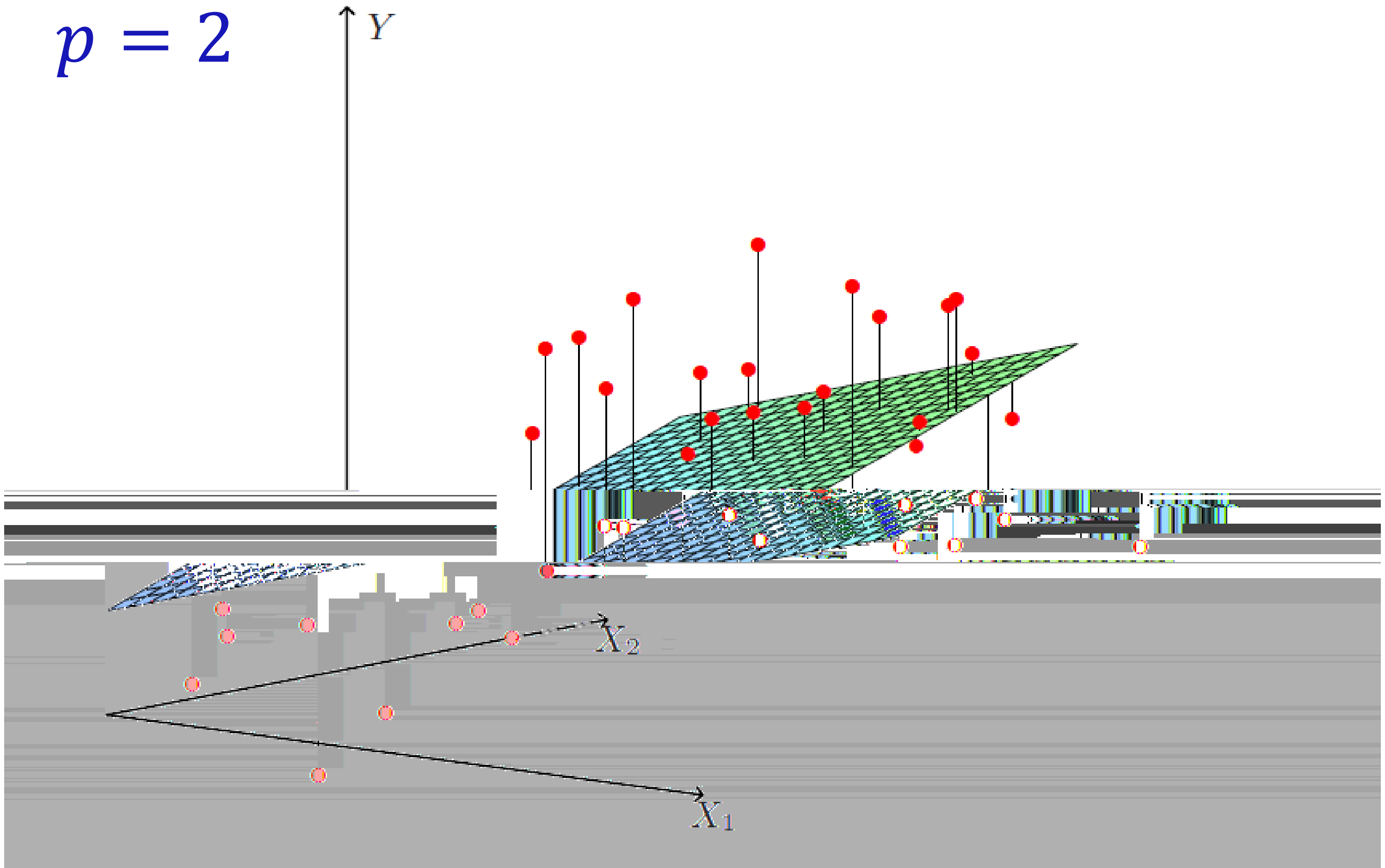


FIGURE 3.1. Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.

Optimization (2)

Differentiate with respect to β

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

Set the derivative to zero

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

Assume $X^T X$ is invertible

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Predictions

The Prediction of x_0

$$\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$$

The Predictions of Training Data

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Let $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p]$

$$\hat{\beta} = \operatorname{argmin} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

$\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the subspace spanned by $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p$

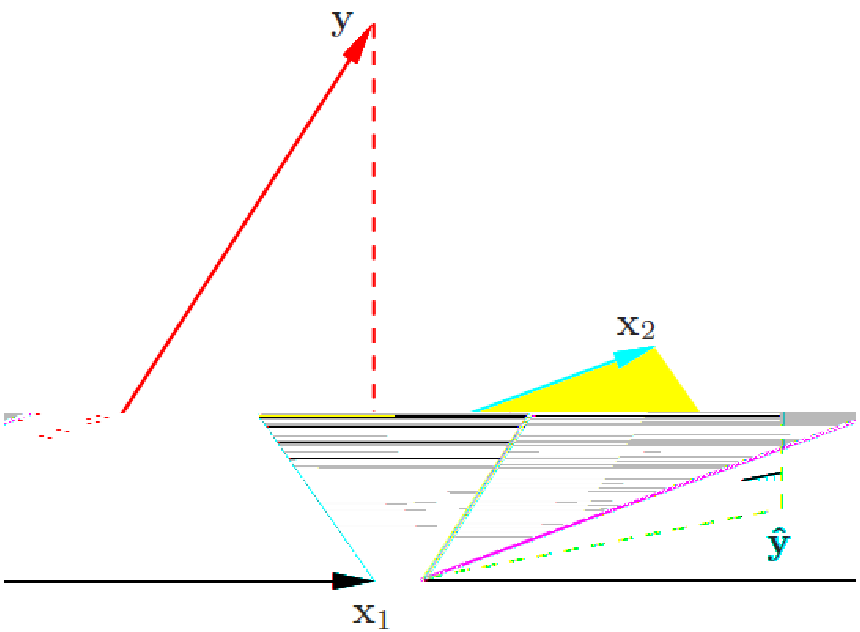
Predictions

The Prediction of x_0

$$\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$$

The Predictions of Training Data

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$



Understanding (1)

Assume the linear model is right, but the observation contains noise

$$\begin{aligned} Y &= \mathrm{E}(Y|X_1, \ldots, X_p) + \varepsilon \\ &= \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \end{aligned}$$

Where $\epsilon \sim N(0, \sigma^2)$

Then $\hat{\beta} = (X'X)^{-1}X'Y$

$= (X'X)^{-1}X'(X\beta + \varepsilon) \qquad \varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]'$

$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon$

$= \beta + (X'X)^{-1}X'\varepsilon$

Understanding (2)

Since $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ is a Gaussian random vector, thus

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$$

is also a Gaussian random vector

$$E(\hat{\beta}) = \beta + E\left((X^T X)^{-1} X^T \epsilon\right)$$
$$= \beta + (X^T X)^{-1} X^T E(\epsilon) = \beta$$

$$Cov(\hat{\beta}) = Cov\left((X^T X)^{-1} X^T \epsilon\right)$$
$$= (X^T X)^{-1} X^T Cov(\epsilon) X (X^T X)^{-1}$$
$$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = (X^T X)^{-1} \sigma^2$$

Thus $\hat{\beta} \sim N\left(\beta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2\right)$

Expected Prediction Error (EPE)

Given a test point x_0, assume

$$y_0 = f(x_0) + \varepsilon_0, \quad \varepsilon_0 \sim N(0, \sigma^2)$$

The EPE of $\tilde f(x_0) = x_0^T \tilde\beta$ is

$$\text{EPE}(x_0) = E_{y_0|x_0} E_{\mathcal{T}} (y_0 - \tilde f(x_0))^2 = \sigma^2 + \text{MSE}(\tilde f(x_0)).$$

The Mean Squared Error (MSE)

$$\text{MSE}\left(\tilde f(x_0)\right) = E\left(x_0^T \tilde\beta - f(x_0)\right)^2$$

$$= E\left(x_0^T \tilde\beta - E(x_0^T \tilde\beta)\right)^2 + \left(E(x_0^T \tilde\beta) - f(x_0)\right)^2$$

$$\text{Variance}(x_0^T \tilde\beta) \qquad \text{Bias}(x_0^T \tilde\beta)$$

EPE of Least Squares

Under the assumption that

$$y_0 = f(x_0) + \varepsilon_0, \quad f(x_0) = x_0^T \beta, \quad \varepsilon_0 \sim N(0, \sigma^2)$$

The EPE of $\hat{f}(x_0) = x_0^T \hat{\beta}$ is

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \sigma^2 + E\left(x_0^T\hat{\beta} - x_0^T\beta\right)^2$$

$$= \sigma^2 + \mathrm{MSE}\left(x_0^T\hat{\beta}\right)^2$$

The Mean Squared Error (MSE)

$$\mathrm{MSE}\left(x_0^T\hat{\beta}\right) = E\left(x_0^T\hat{\beta} - x_0^T\beta\right)^2$$

$$= E\left(x_0^T\hat{\beta} - E\left(x_0^T\hat{\beta}\right)\right)^2$$

$$= \mathrm{Var}\ x_0^T\hat{\beta}$$

The Gauss–Markov Theorem

$\hat{\beta}$ has the smallest variance among all linear unbiased estimates.

We aim to estimate $f(x_0) = x_0^T \beta$, the estimation of $\hat{f}(x_0) = x_0^T \hat{\beta}$ is

From precious discussions, we have

and for all $c^T \mathbf{y}$

Multiple Outputs (1)

Suppose we aim to predict K outputs

Multiple Outputs (2)

The Residual Sum of Squares

$$\begin{aligned}\mathrm{RSS}(\mathbf{B}) &= \sum_{k=1}^{K}\sum_{i=1}^{N}(y_{ik} - f_k(x_i))^2 \\ &= \mathrm{tr}[(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})]\end{aligned}$$

The Solution

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

It is equivalent to performing K independent least squares

Large-scale Setting

The Problem

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Sampling

Faster least squares approximation

Outline

- Introduction
- Linear Regression Models and Least Squares
- **Subset Selection**
- Shrinkage Methods
- Methods Using Derived Input Directions
- Discussions
- Summary

Subset Selection

Limitations of Least Squares

Prediction Accuracy: the least squares estimates often have low bias but large variance

Interpretation: We often would like to determine a smaller subset that exhibit the strongest effects

Shrink or Set Some Coefficients to Zero

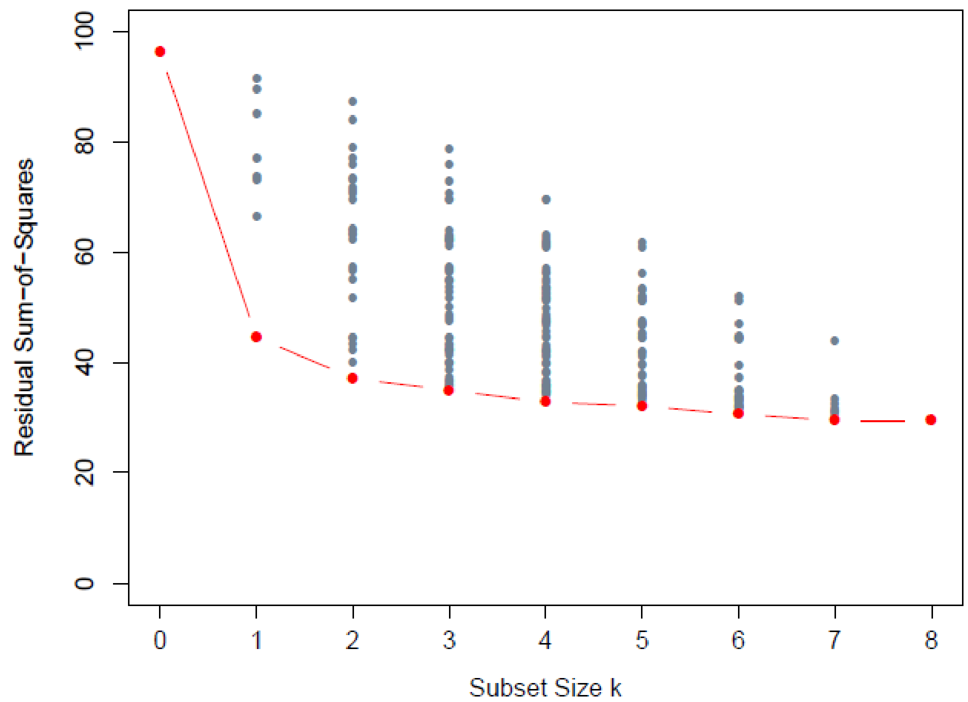
We sacrifice a little bit of bias to reduce the variance of the predicted values

Best-Subset Selection

Select the subset of variables (features) such that the RSS is minimized

$$\text{RSS}(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2$$

ὴ 8



Forward- and Backward-**Stepwise** Selection

Forward-stepwise Selection

1. Start with the intercept
2. Sequentially add into the model the predictor that most improves the fit

Backward-stepwise Selection

1. Start with the full model
2. Sequentially delete the predictor that has the least impact on the fit

Both are greedy algorithms

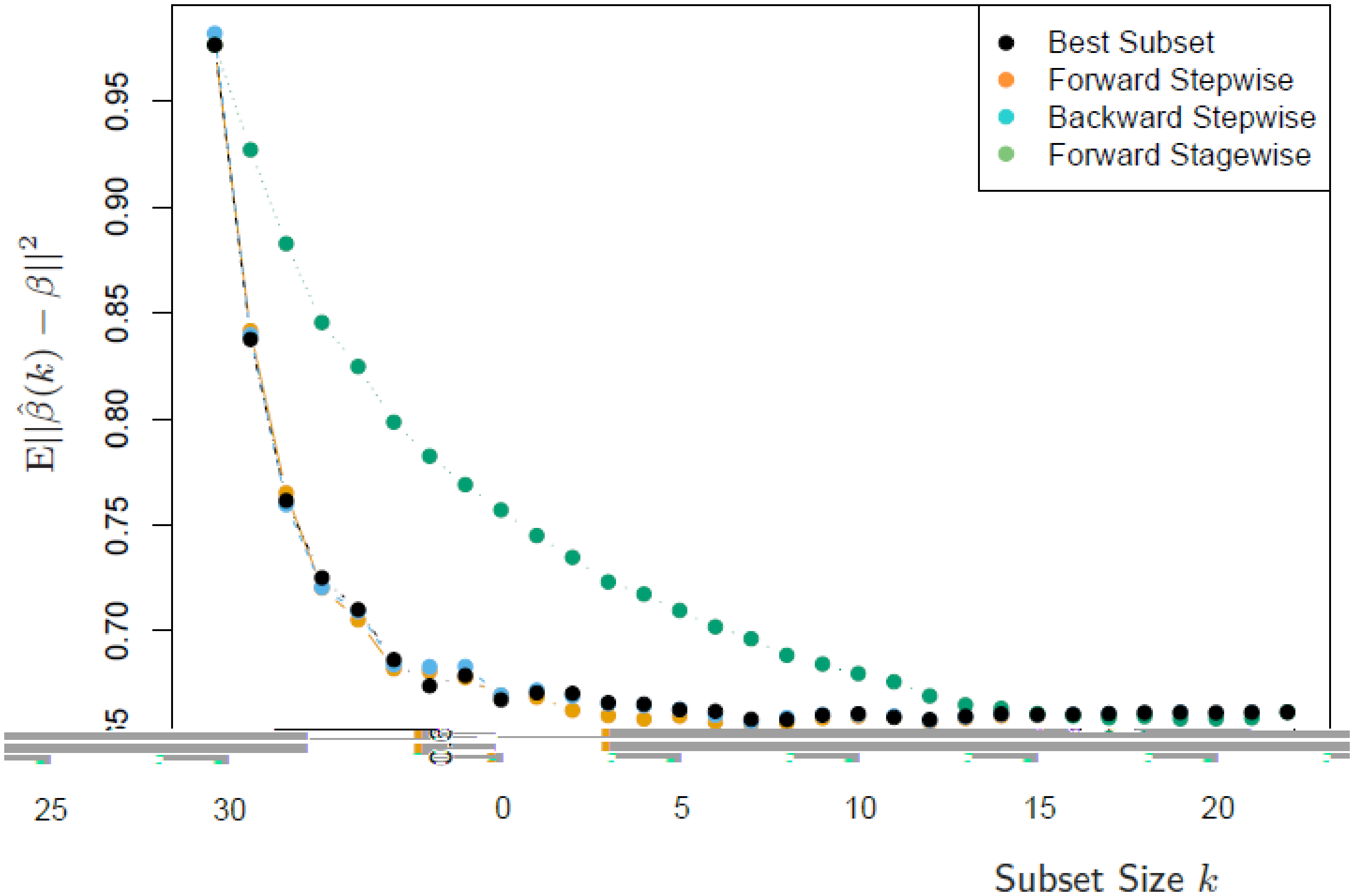
Both can be solved quite efficiently

Forward-Stagewise Regression

1. Start with an intercept equal to $\bar{y}$ and centered predictors with coefficients initially all 0
2. Identify the variable most correlated with the current residual
3. Compute the simple linear regression coefficient of the residual on this chosen variable

None of the other variables are adjusted when a term is added to the model

Comparisons



Outline

- Introduction
- Linear Regression Models and Least Squares
- Subset Selection
- **Shrinkage Methods**
- Methods Using Derived Input Directions
- Discussions
- Summary

Shrinkage Methods

Limitation of Subset Selection

A discrete process—variables are either retained or discarded

It often exhibits high variance, and so doesn't reduce the prediction error

Shrinkage Methods

More continuous, low variance

Ridge Regression

The Lasso

Least Angle Regression

Ridge Regression

Shrink the regression coefficients

By imposing a penalty on their size

The Objective

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

$\lambda \geq 0$ is a complexity parameter

An Equivalent Form

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

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

Coefficients cannot be too large even when variables are correlated

Optimization (1)

Let $\mathbf{X}$ be a matrix with each row an input vector

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

$\beta = [\beta_1, \dots, \beta_d]^T$ and $\mathbf{y} = [y_1, \dots, y_n]^T$

The Objective Becomes

$$\min_{\beta \in \mathbb{R}^d,\, b \in \mathbb{R}} \|\mathbf{y} - \mathbf{X}\beta - b\mathbf{1}_n\|_2^2 + \lambda \|\beta\|_2^2$$

Where $\mathbf{1}_n = [1, \dots, 1]^T \in \mathbb{R}^n$

Optimization (2)

Differentiate with respect to β_0 and set it to zero

$$2\mathbf{1}^T(y - X\beta - \mathbf{1}\beta_0) = 0$$

$$\beta_0 = \frac{1}{n}\mathbf{1}^T(y - X\beta)$$

Differentiate with respect to β and set it to zero

$$2X^T(X\beta - y + \mathbf{1}\beta_0) + 2\lambda\beta = 0$$

$$X^T\left(X\beta - y - \frac{1}{n}\mathbf{1}\mathbf{1}^T(X\beta - y)\right) + \lambda\beta = 0$$

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

Optimization (3)

The Final Solution

Let $H = \mathbf{I} - \frac{[illegible]}{[illegible]}$ [illegible]

Understanding (1)

Assume **X** is centered, then

Let the SVD of **X** be

- ή ʺI, … , ʺI contains the left singular vectors
- ᾬ is a diagonal matrix with diagonal entries ὨὨὨ 0

Then, we examine the prediction of training data

Understanding (2)

Least Squares

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

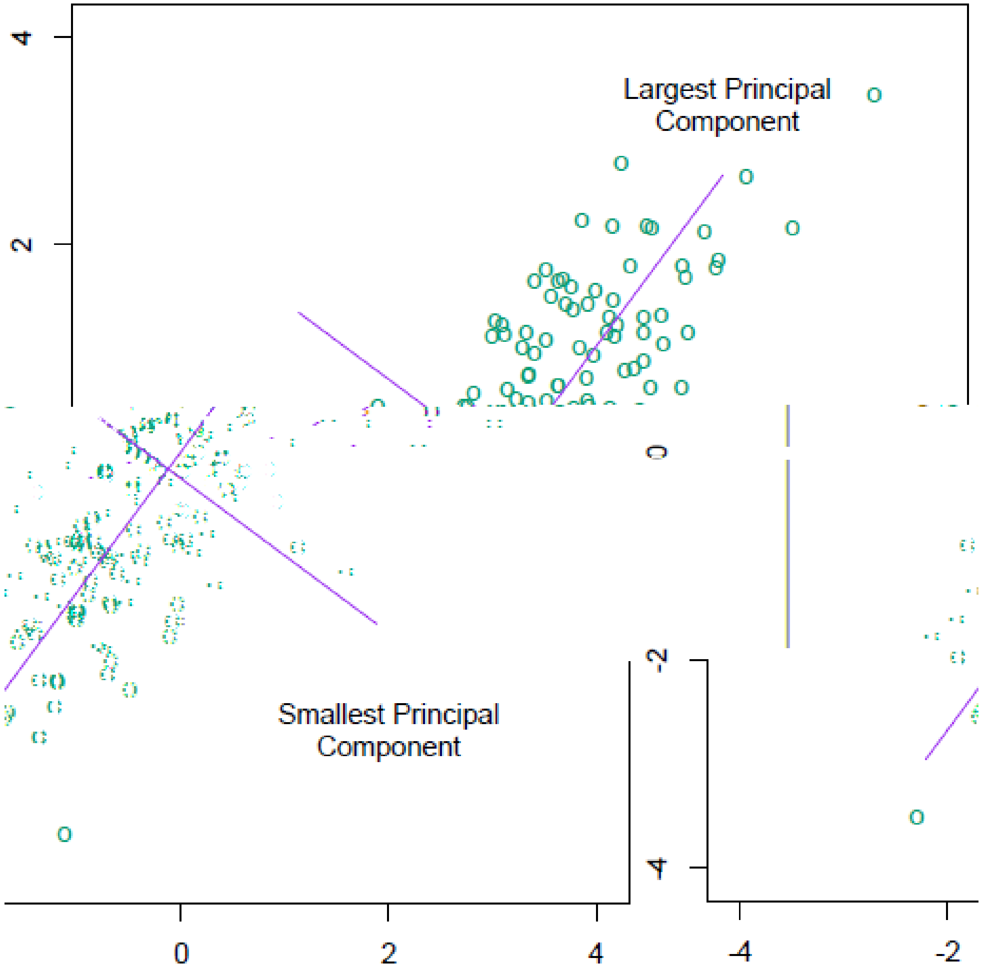
Ridge Regression

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

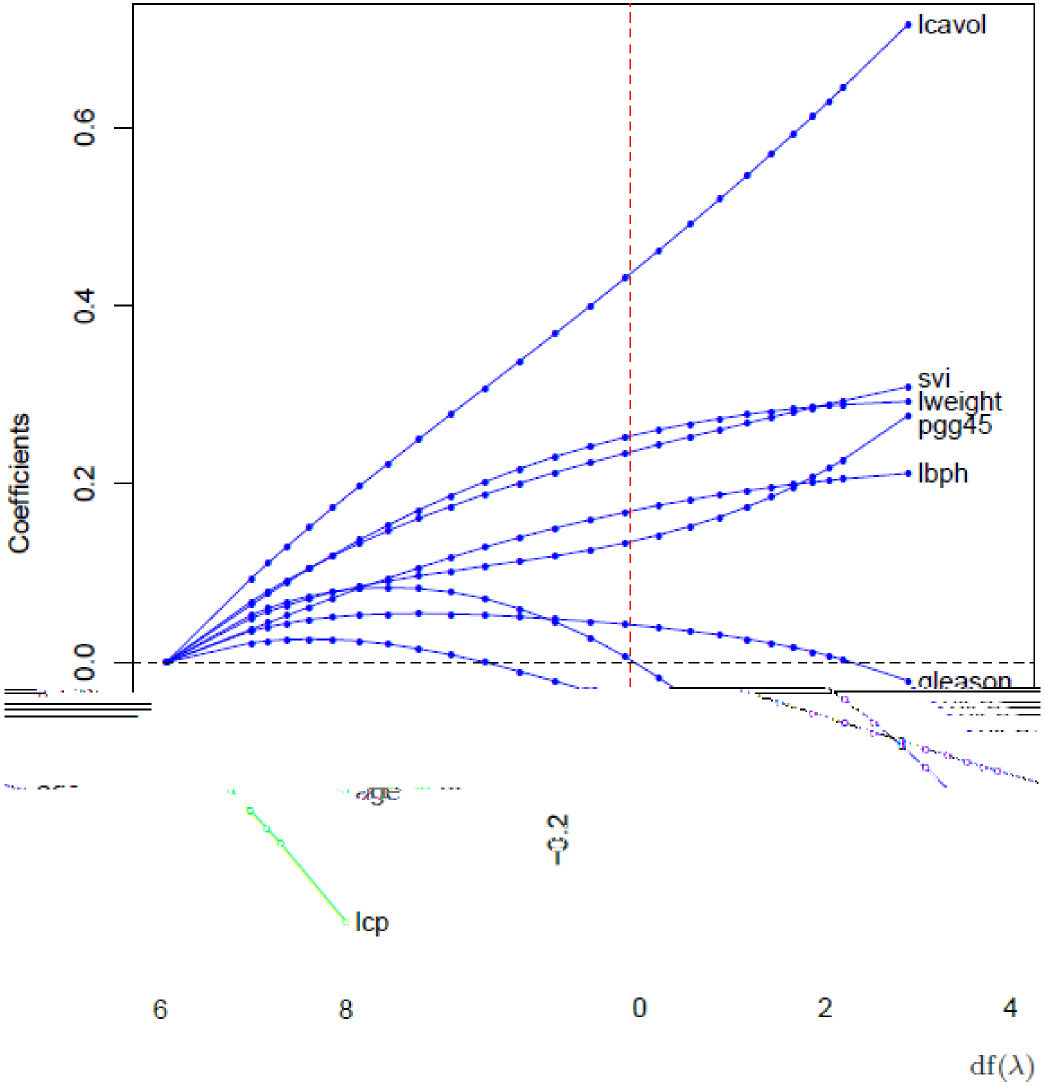
Shrink the coordinates by $\frac{d_j^2}{d_j^2+\lambda} \leq 1$

Understanding (3)

Connection with PCA



An Example



lcavol
svi
lweight
pgg45
lbph
gleason
age
lcp
Coefficients
0.0
0.2
0.4
0.6
−0.2
6
8
0
2
4
df(λ)

Optimization

The First Formulation

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

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

Gradient descent followed by Projection [1]

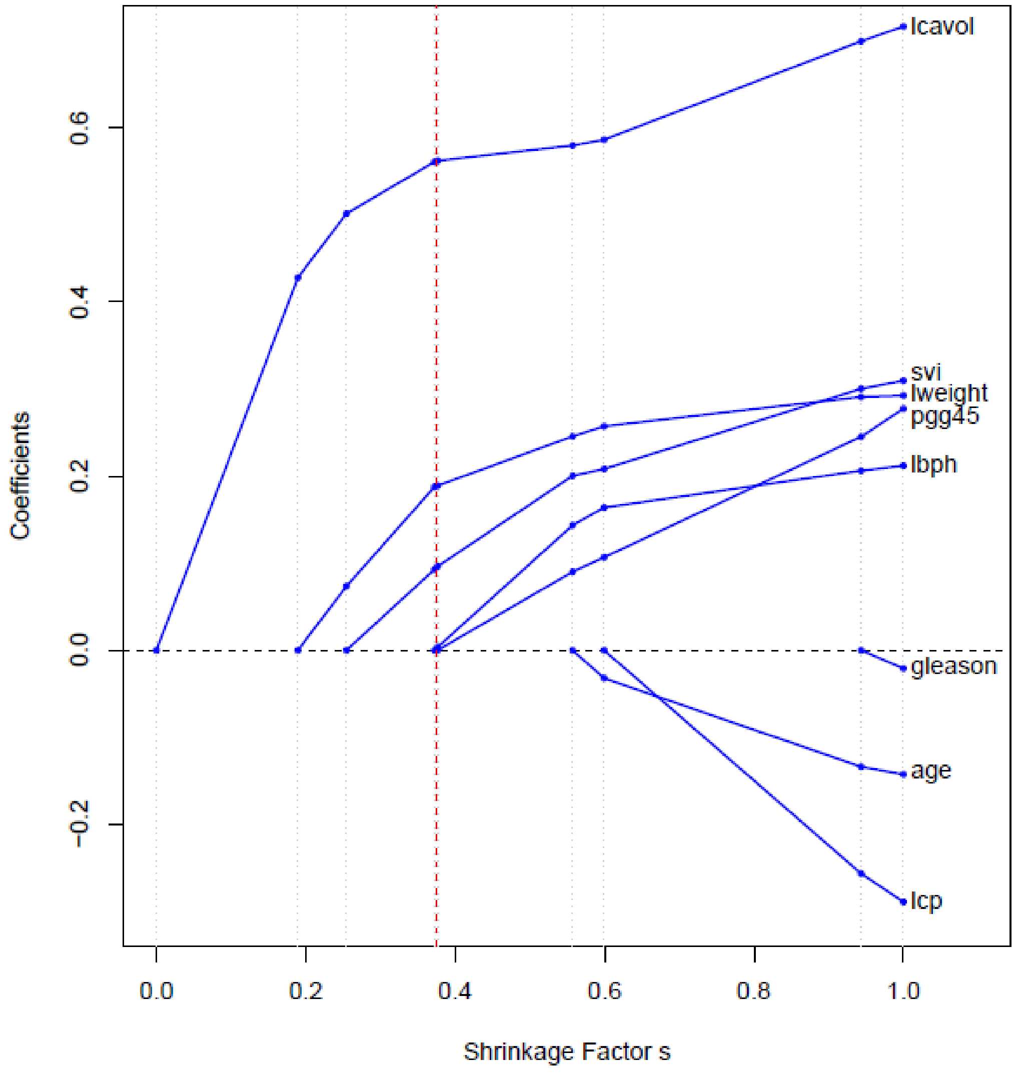
The Second Formulation

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

Convex Composite Optimization [2]

An Example

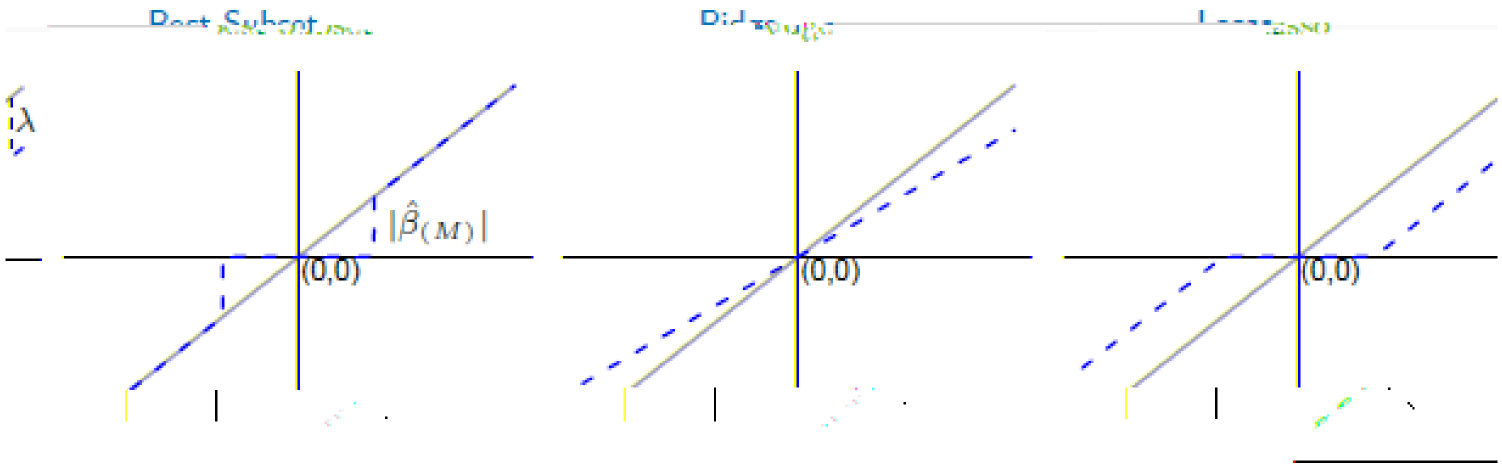
Hit 0
Piece-wise linear



Subset Selection, Ridge, Lasso

Columns of $\mathbf{X}$ are orthonormal

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



Hard-thresholding

Scaling

Soft-thresholding

Ridge v.s. Lasso (1)

Ridge Regression

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

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

ℓ_2-norm appears in the constraint

Lasso

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

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

ℓ_1-norm appears in the constraint

Ridge v.s. Lasso (2)

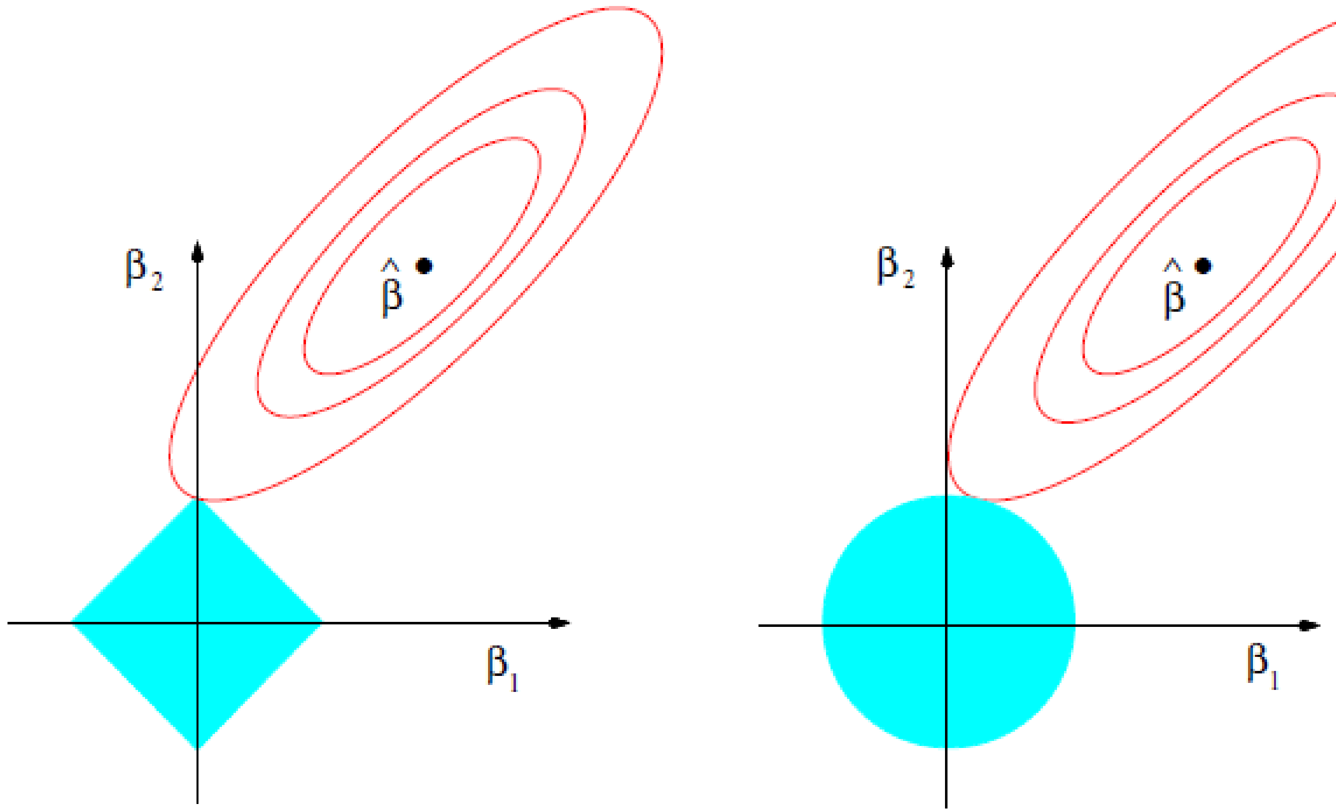


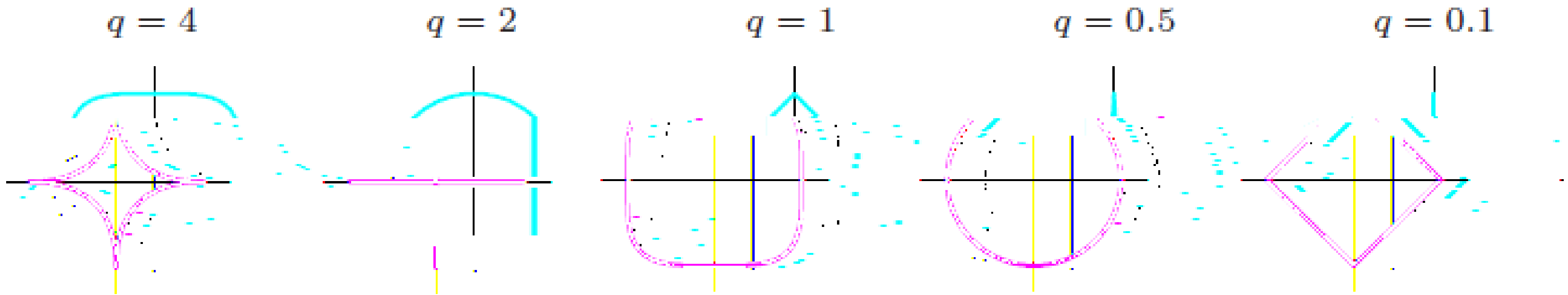
FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Generalization (1)

A General Formulation

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

Contours of Constant Value of $\sum_j |\beta_j|^q$



Generalization (2)

A Mixed Formulation

The *elastic-net* penalty

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

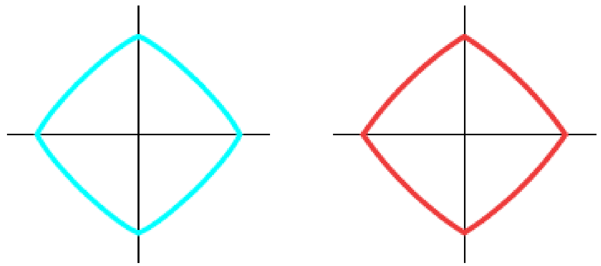


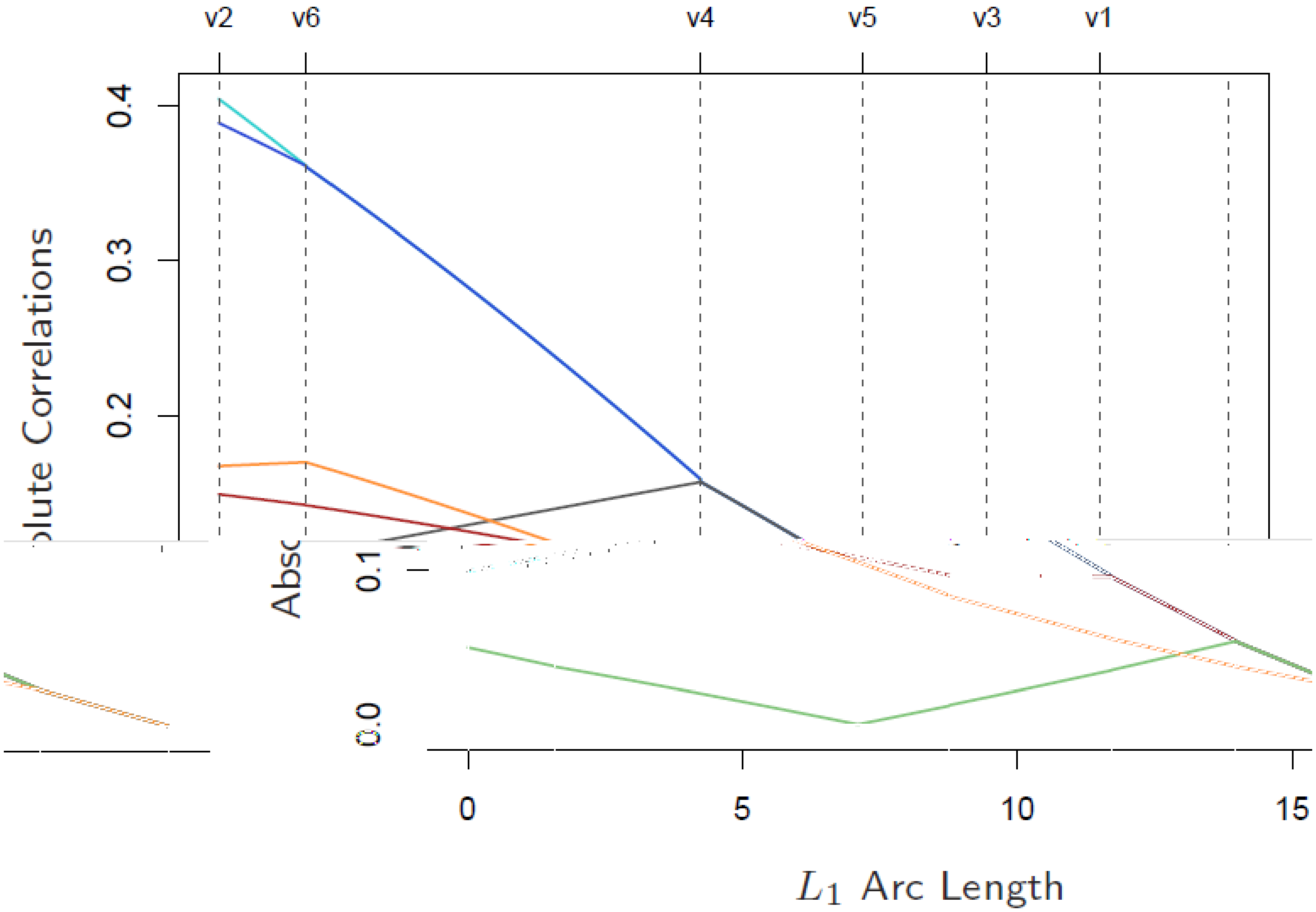
FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j (\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

Least Angle Regression (LAR)

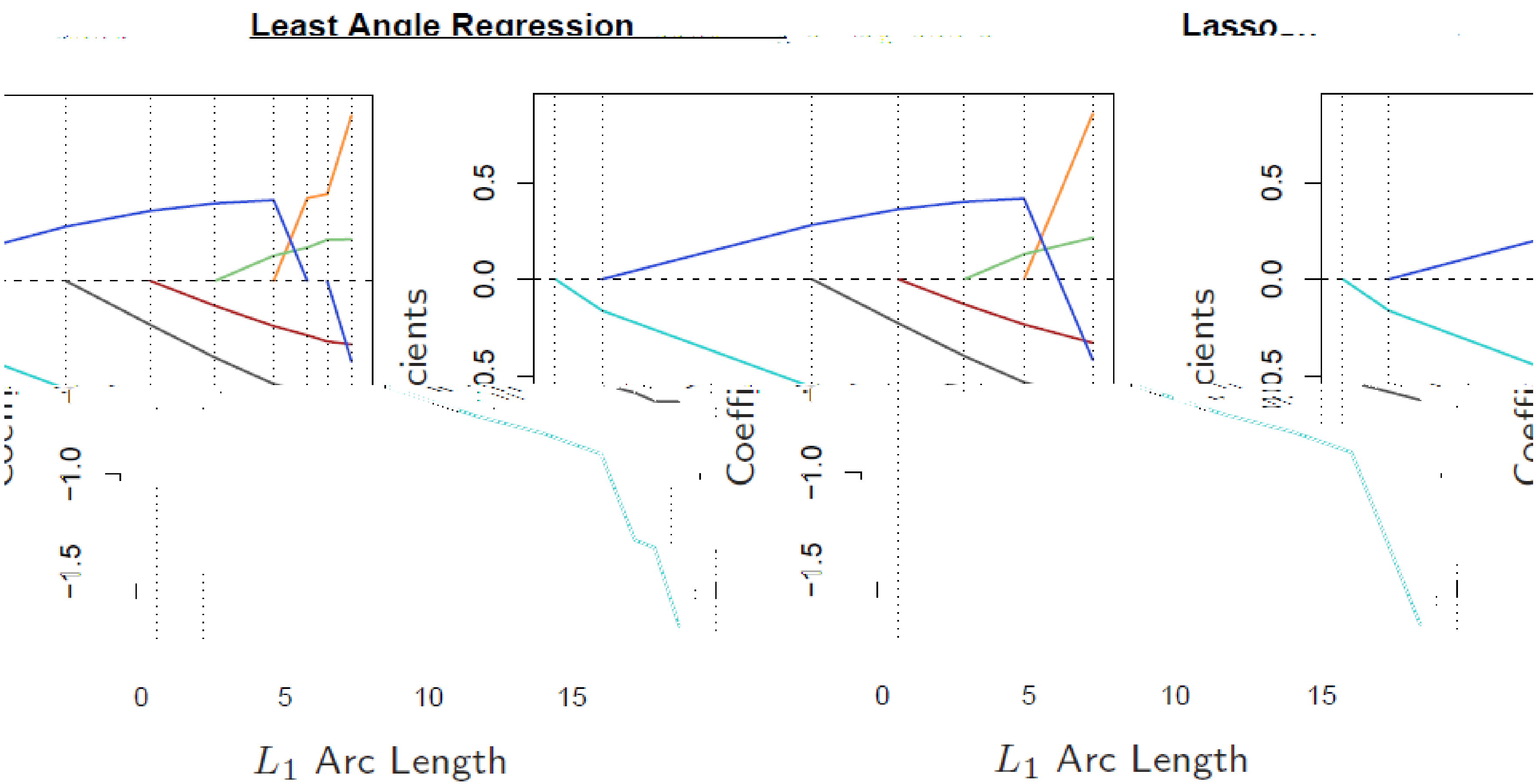
The Procedure

1. Identify the variable most correlated with the response
2. Move the coefficient of this variable continuously toward its least squares value
3. As soon as another variable "catches up" in terms of correlation with the residual, the process is paused
4. The second variable then joins the active set, and their coefficients are moved together in a way that keeps their correlations tied and decreasing

An Example


v2
v6
v4
v5
v3
v1
0.4
0.3
0.2
0.1
0.0
lute Correlations
Absc
0
5
10
15
L_1 Arc Length

LAS v.s. Lasso



Least Angle Regression
Lasso
0.5
0.0
−1.0
−1.5
0
5
10
15
L_1 Arc Length
0
5
10
15
L_1 Arc Length

Outline

- Introduction
- Linear Regression Models and Least Squares
- Subset Selection
- Shrinkage Methods
- **Methods Using Derived Input Directions**
- Discussions
- Summary

Methods Using Derived Input Directions

We have a large number of inputs

Often very correlated

1. Generate a small number of linear combinations

$$Z_m, \; m = 1, \ldots, M$$

of the original inputs X_j

2. Use Z_m in place of X_j as inputs in the regression

Linear Dimensionality Reduction + Regression

Principal Components Regression (PCR)

The linear combinations Z_m are generated by PCA

$$\mathbf{z}_m = \mathbf{X}v_m$$

$\mathbf{X}$ is centered, and v_m is the m-th right singular vector

Since $\mathbf{z}_m$'s are orthogonal

$$\hat{\mathbf{y}}^{\text{pcr}}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m \qquad \sum_{m=1}^{M} \hat{\theta}_m v_m = \hat{\beta}^{\text{pcr}}(M)$$

where $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$

PCR v.s. Ridge

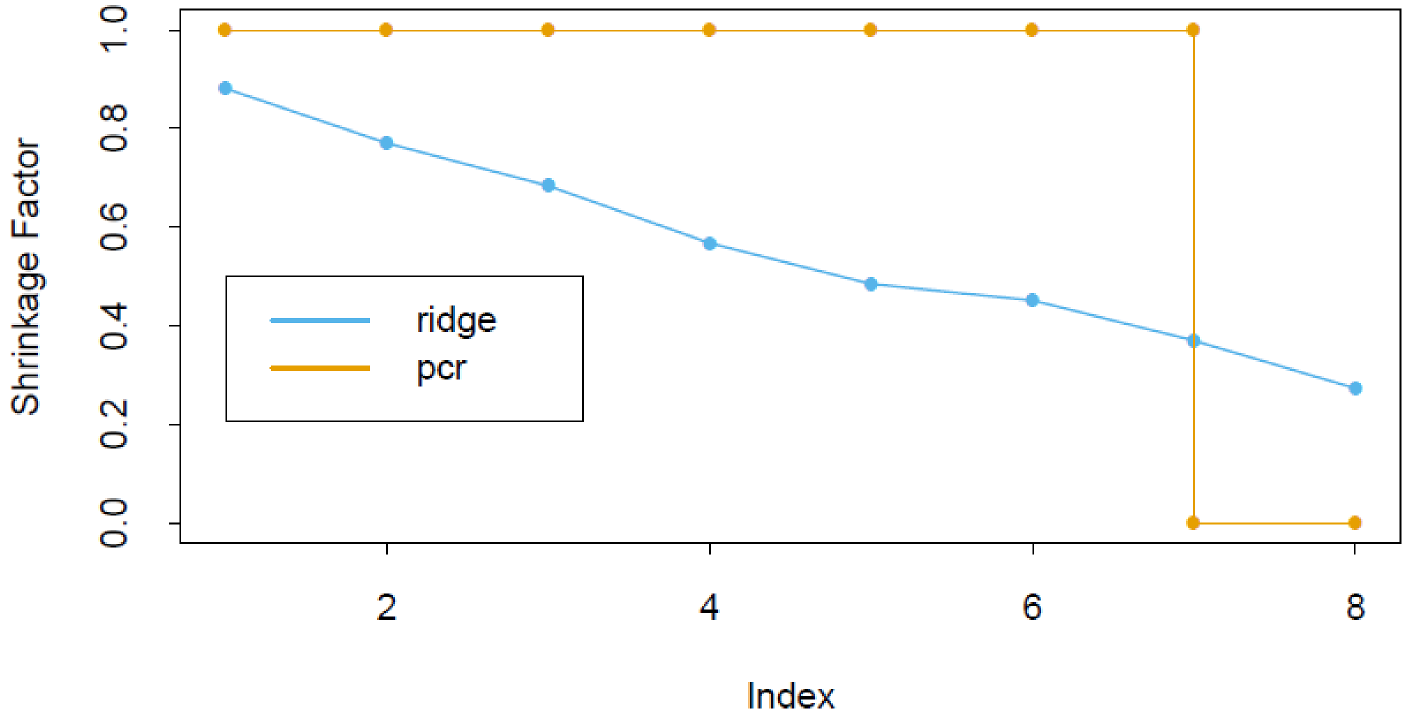


FIGURE 3.17. *Ridge regression shrinks the regression coefficients of the principal components, using shrinkage factors $d_j^2/(d_j^2+\lambda)$ as in (3.47). Principal component regression truncates them. Shown are the shrinkage and truncation patterns corresponding to Figure 3.7, as a function of the principal component index.*

Partial Least Squares (PLS)

The Procedure

1. Compute $\hat{\varphi}_{1j} = \langle \mathbf{x}_j, \mathbf{y} \rangle$ for each feature $\mathbf{x}_j$
2. Construct the 1st derived input $\mathbf{z}_1 = \sum_j \hat{\varphi}_{1j} \mathbf{x}_j$
3. $\mathbf{y}$ is regressed on $\mathbf{z}_1$ giving coefficient $\hat{\theta}_1$
4. Orthogonalize $\mathbf{x}_1, \ldots, \mathbf{x}_p$ with respect to $\mathbf{z}_1$
5. Repeat the above process

Outline

Introduction

Linear Regression Models and Least Squares

Subset Selection

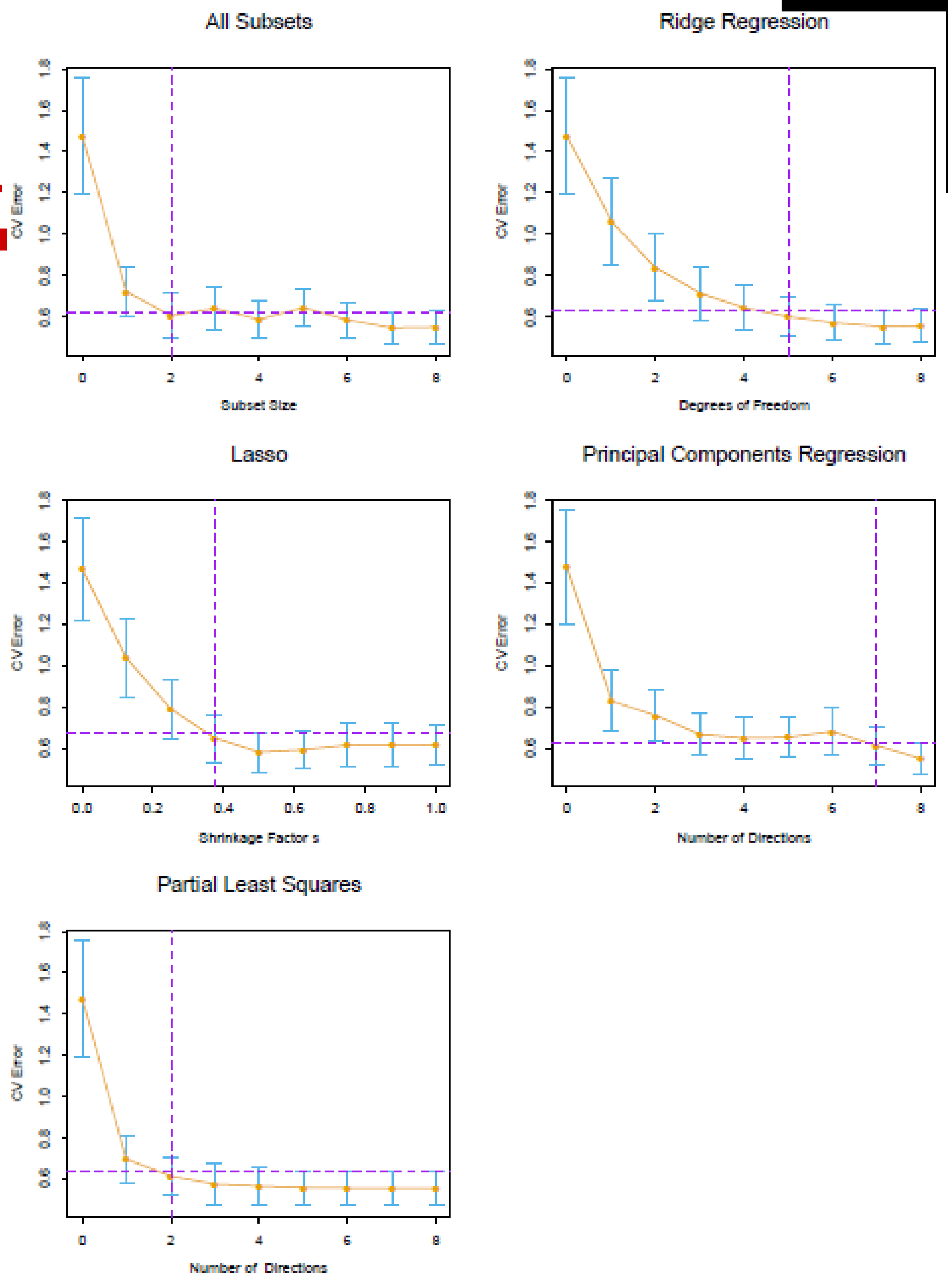
† Square7

†

†

Discussions (1

Model complexity increases as we move from left to right.

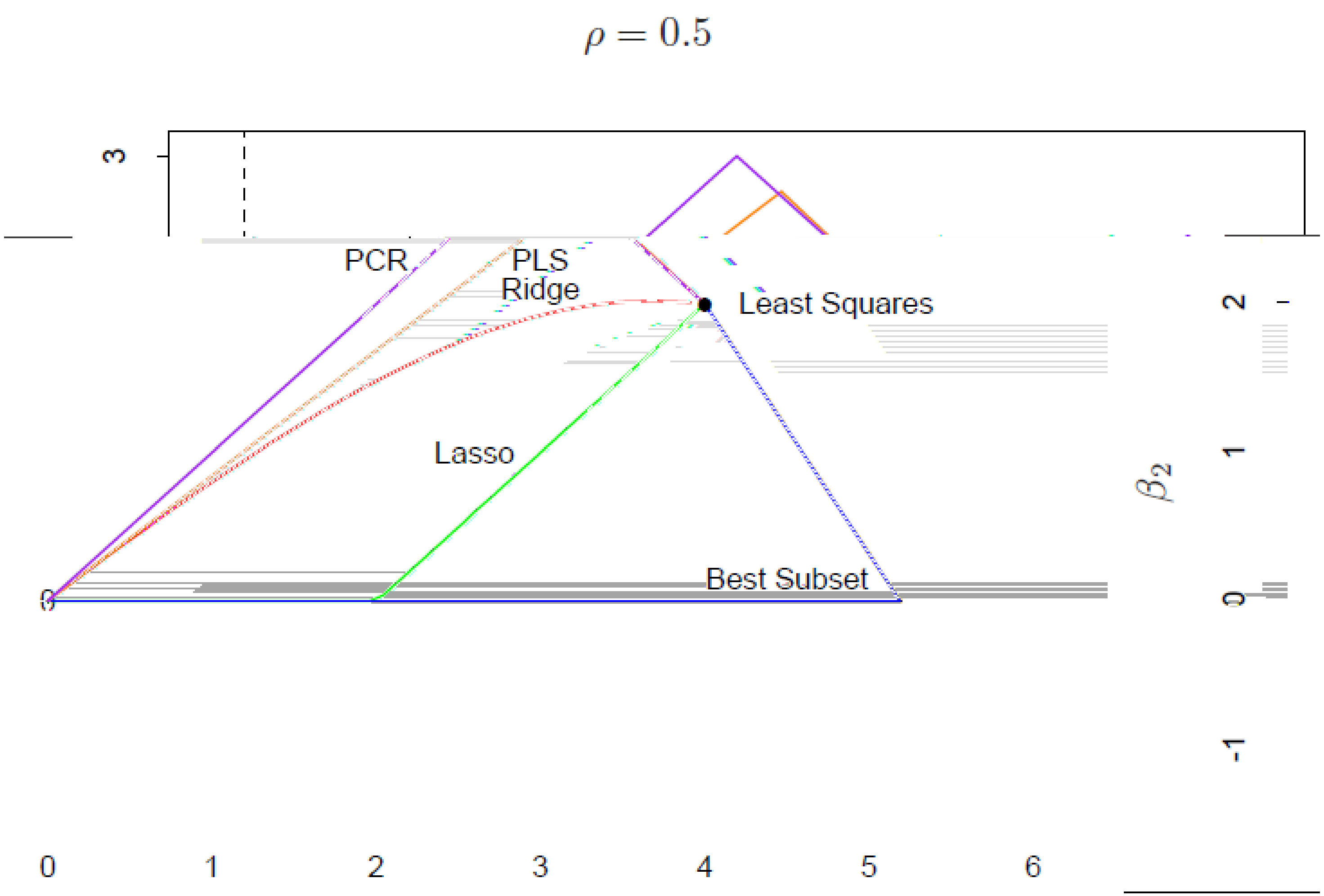


Discussions (2)

TABLE 3.3. Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152 [illegible]

Discussions (3)



Outline

- Introduction
- Linear Regression Models and Least Squares
- Subset Selection
- Shrinkage Methods
- Methods Using Derived Input Directions
- Discussions
- **Summary**

Summary

- Linear Regression Models
- Least Squares
- Shrinkage Methods
  - Ridge Regression
  - Lasso
  - Least Angle Regression (LAR)
- Methods Using Derived Input Directions
  - Principal Components Regression (PCR)
  - Partial Least Squares (PLS)

Reference

[1] Duchi et al. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In Proceedings of the 25th international conference on Machine learning, pp. 272-279, 2008.

[2] Nesterov. Gradient methods for minimizing composite functions. Mathematical Programming, 140(1): 125-161, 2013.