Shrinkage Methods

(ISLR 6.2)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

Shrinkage Methods

Ridge regression and *Lasso*

- The subset selection methods use least squares to fit a linear model that contains a subset of the predictors.
- As an alternative, we can fit a model containing all p predictors using a technique that shrinks the coefficient estimates towards zero.
- Shrinking the coefficient estimates can significantly reduce variance.

Ridge regression

- Recall that OLS estimator $\hat{\boldsymbol{\beta}}$ minimizes the RSS.
- *Ridge regression* uses a slightly different objective function

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{RSS}}+\underbrace{\lambda\sum_{j=1}^{p}\beta_j^2}_{\text{shrinkage penalty}}$$

  - As with least squares, ridge regression seeks coefficient estimates that fit the data well, by making the RSS small.
  - The second term $\lambda\sum_{j=1}^{p}\beta_j^2$, is small when $\beta_1,\ldots,\beta_p$ are close to zero, and so it has the effect of shrinking the estimates of β_j's towards zero.
  - $\lambda\geq 0$ is a *tuning parameter*, is to be determined separately. It controls the relative impact of these two terms.

Selecting λ

Selecting a good value for λ is critical

$$\hat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

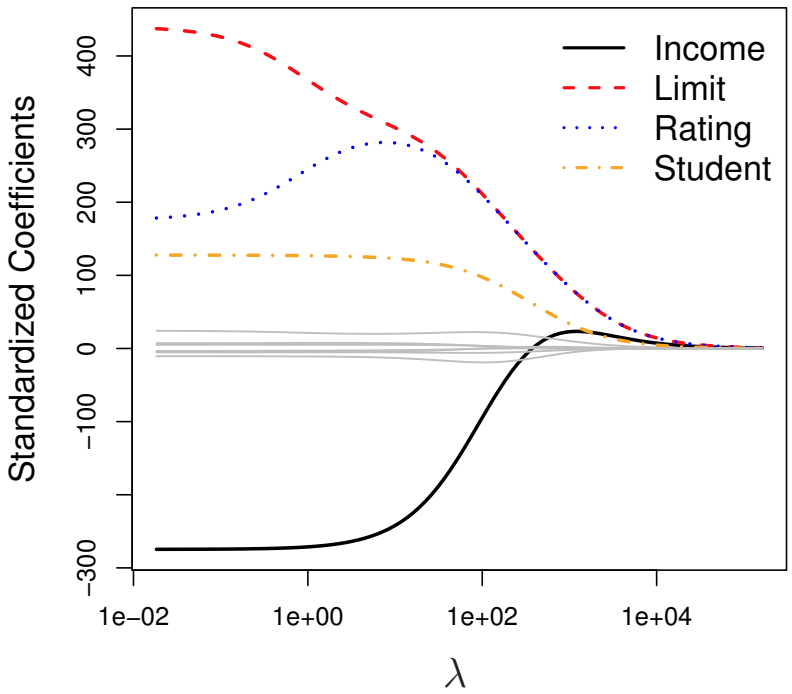
- When $\lambda = 0$, we get the OLS.
- When λ is very large, we get all zeros.
- Cross-validation is usually used to find an optimal λ.

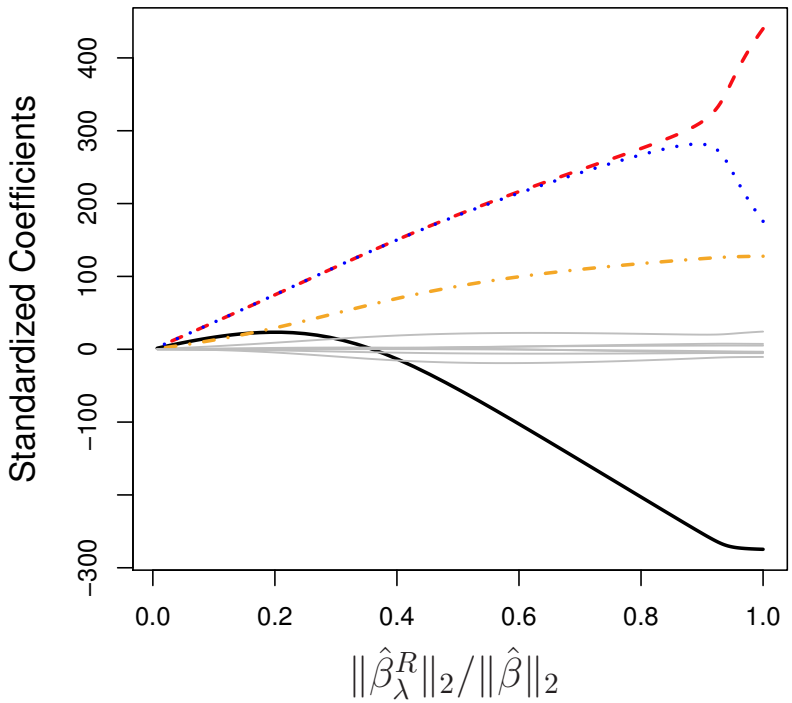
The Credit data example

When λ increases,

- the standardized coefficients shrink towards zero in general;
- individual coefficients may occasionally increase.

ℓ_2 norm: $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$





Standardizing predictors: centering

$$\hat{\boldsymbol{\beta}}_{\lambda}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- The intercept β_0 is not included in the penalty term.
- We do not want to shrink the intercept, which is simply a measure of the mean value of the response when all predictors take value at zero.
- Usually, all predictor are first centered to have mean zero,

$$x_{ij} \longrightarrow x_{ij} - \bar{x}_j, \quad \text{for } i = 1, 2, \ldots, n,$$

where $\bar{x}_j$ is the sample mean of X_j; then

$$\hat{\beta}_0 = \bar{y}$$

Standardizing predictors: scaling

- The OLS coefficient estimates are invariant of scaling:

$$\beta_j X_j = \underbrace{(\beta_j/c)}_{\gamma_j}\underbrace{(cX_j)}_{\tilde{X}_j} \Longrightarrow \hat{\gamma}_j^{\text{OLS}} = \hat{\beta}_j^{\text{OLS}}/c$$

- In contrast, the ridge regression coefficient estimates are NOT invariant of scaling, due to the penalty term

$$\sum_{j=1}^{p} \beta_j^2 \neq \sum_{k\neq j} \beta_k^2 + \gamma_j^2.$$

- Usually, after centering, all predictor are then re-scaled.

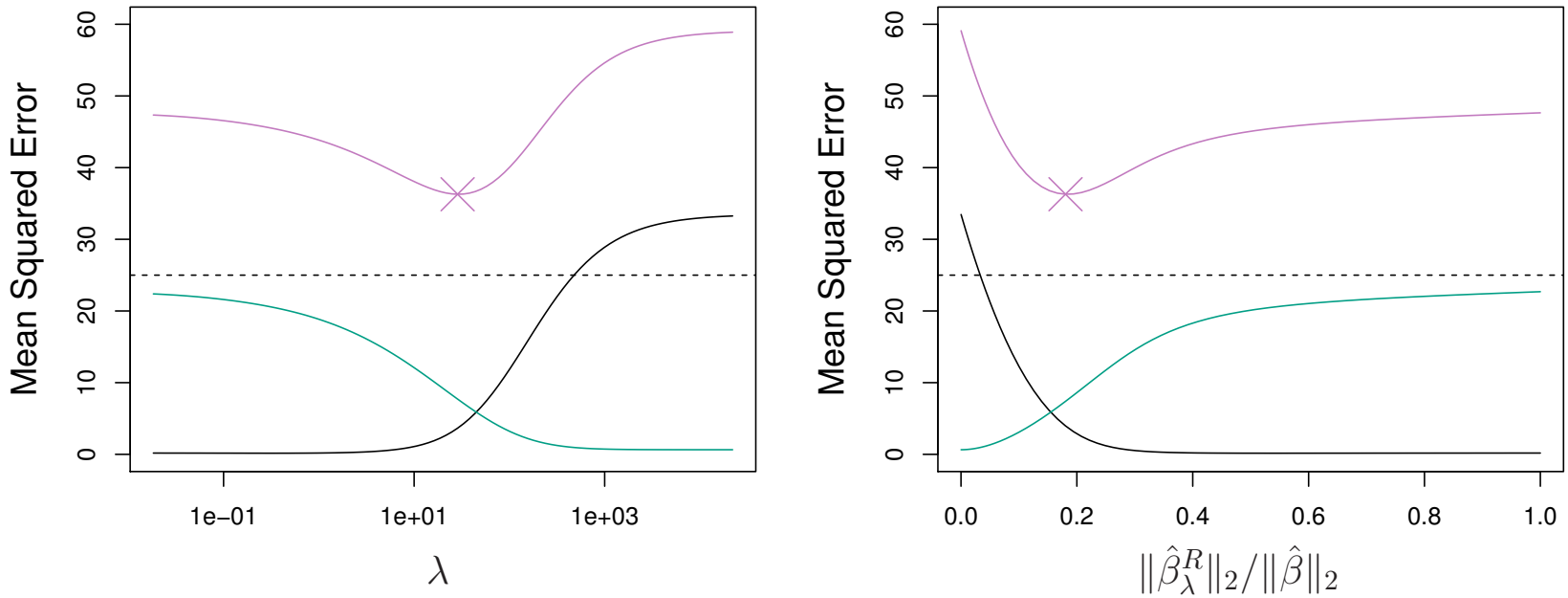
$$x_{ij} \longrightarrow \frac{x_{ij} - \bar{x}_j}{\hat{\sigma}_j}, \quad \text{for } i = 1, 2, \ldots, n,$$

where $\bar{x}_j$ and $\hat{\sigma}_j$ are the sample mean and standard deviation of X_j.

Why can shrinking towards zero be a good thing to do?

The Bias-Variance tradeoff.

Variance, squared bias, test MSE for a simulated data: $n = 50, p = 45$.



- OLS estimates generally have low bias but can be highly variable, in particular when n and p are of similar size or when $p > n$.
- The penalty term makes the ridge regression estimates biased but can also substantially reduce variance

Ridge regression has a closed form solution

After centering and scaling predictors, and centering the response, we can consider a linear regression model with no intercept:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \mathbf{X} = (X_1, X_2, \cdots, X_p) \in \mathbb{R}^{n \times p}$$

Then the ridge regression solution is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} \ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y} \end{aligned}$$

- Ridge regression can be used even when $p > n$.

The lasso (least absolute shrinkage and selection operator)

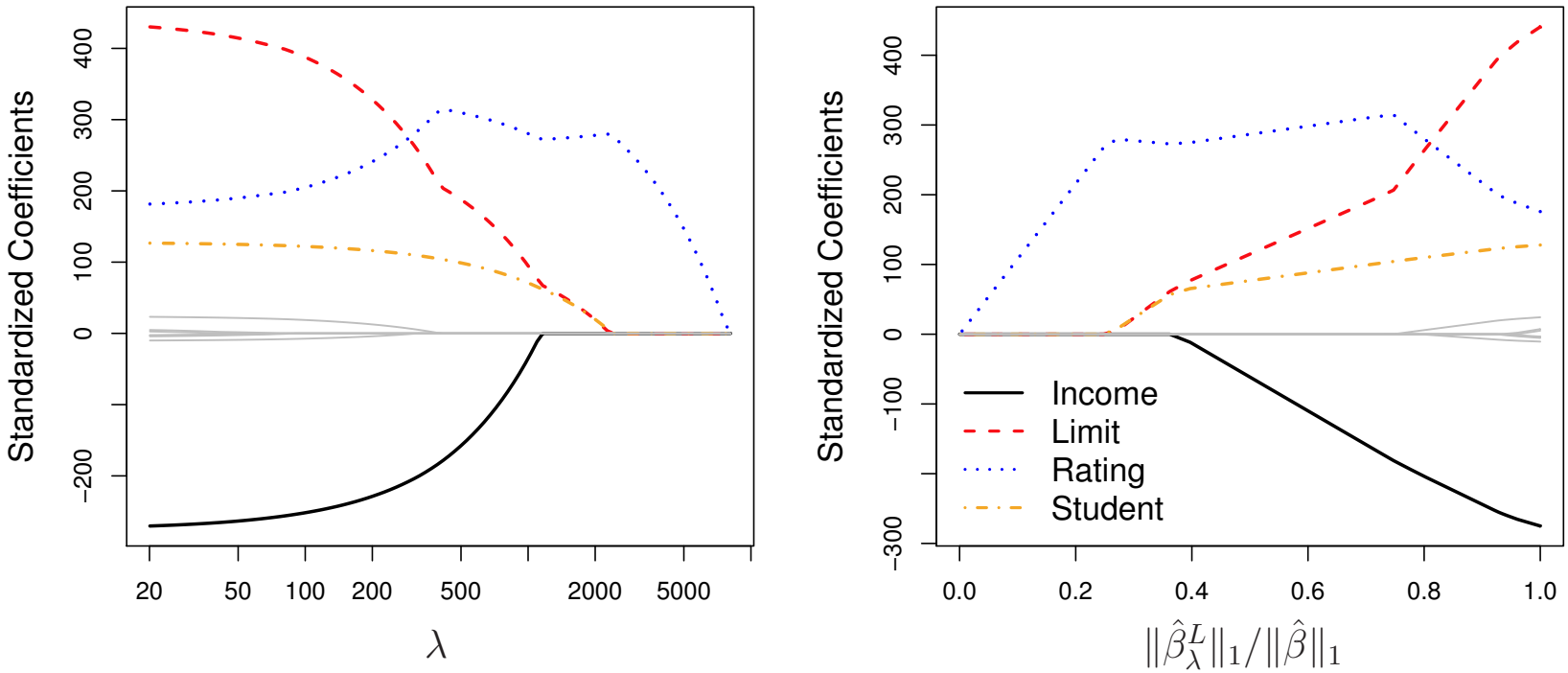
- Ridge regression is not perfect
- One significant problem is that the penalty term will never force any of the coefficients to be exactly zero. Thus, the final model will include all variables, which makes it harder to interpret
- A more modern alternative is the lasso
- The lasso works in a similar way to ridge regression, except it uses a different penalty term

The lasso has an ℓ_1 penalty

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The ℓ_1 norm of a coefficient vector $\boldsymbol{\beta}$ is given by $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.
  - The ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
  - Like best subset selection, the lasso performs *variable selection*.
  - The lasso yields *sparse models* – it is simple to interpret.
- Again β_0 is not included in the penalty term.
- Predictors are centered and scaled to one in the pre-processing step.
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice.

The Credit data example



The variable selection property of the lasso

One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| < s,$$

and

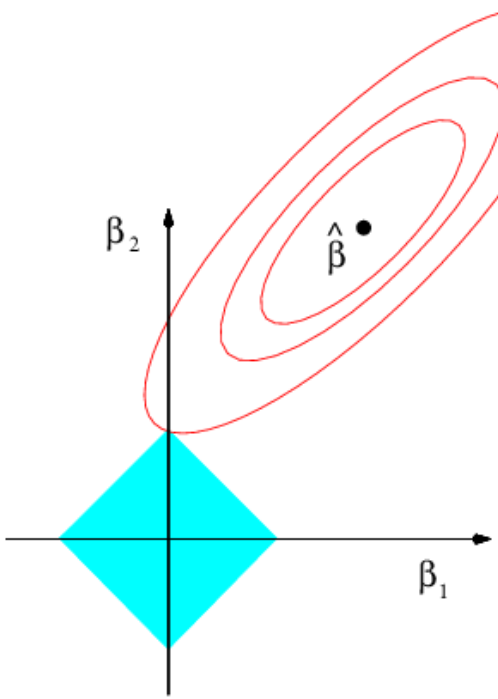
$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 < s,$$

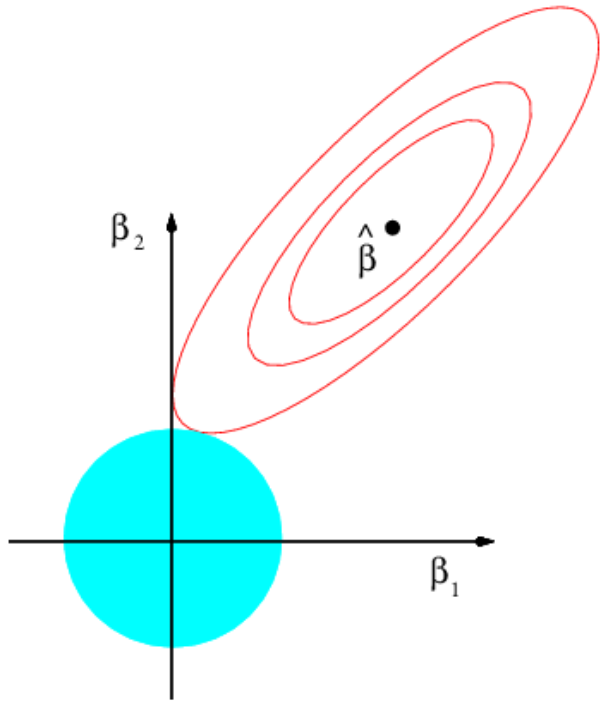
respectively.

- For every value of λ, there is some s such that the above formulas will give the same lasso and ridge regression coefficient estimates.

The lasso vs. ridge regression: sparsity

An illustration $p = 2$, for a given s.
Red ellipses that are centered around $\hat{\beta}^{\text{OLS}}$: regions of constant RSS.



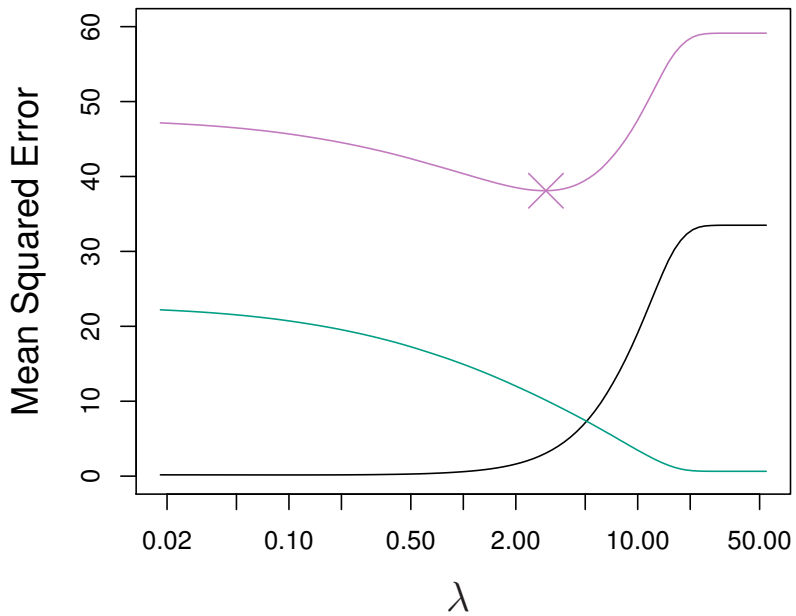


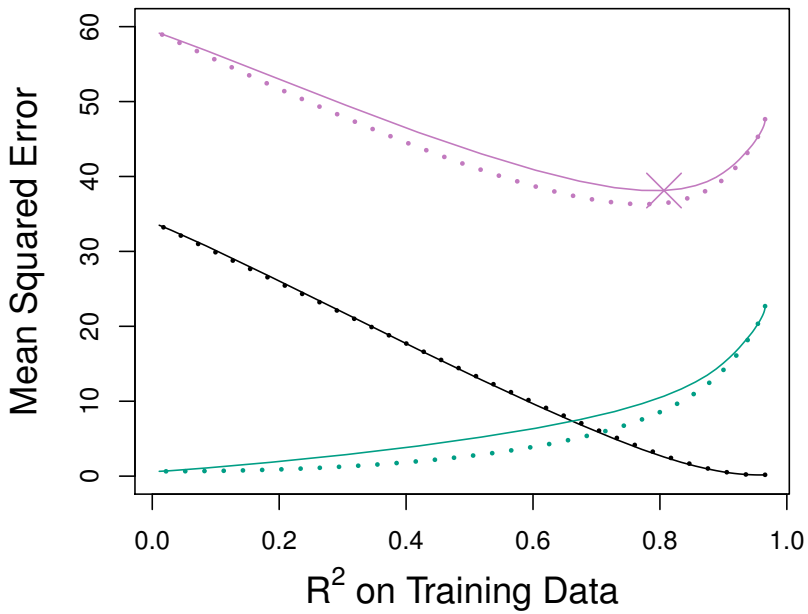
Blue area: $\sum_{j=1}^{p} |\beta_j| < s$ Blue area: $\sum_{j=1}^{p} \beta_j^2 < s$

The lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis.

The lasso vs. ridge regression: prediction accuracy

For a simulated data: $n = 50, p = 45$. All predictors are related to Y.



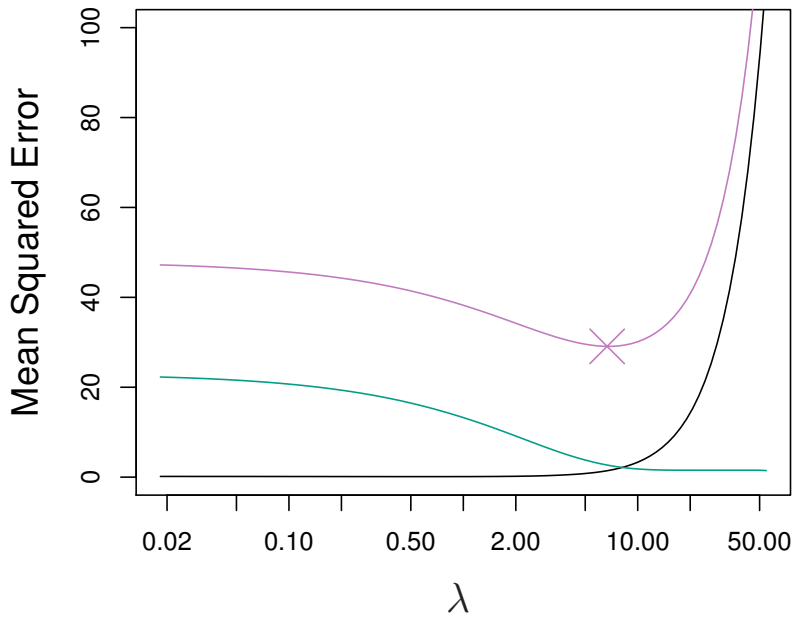


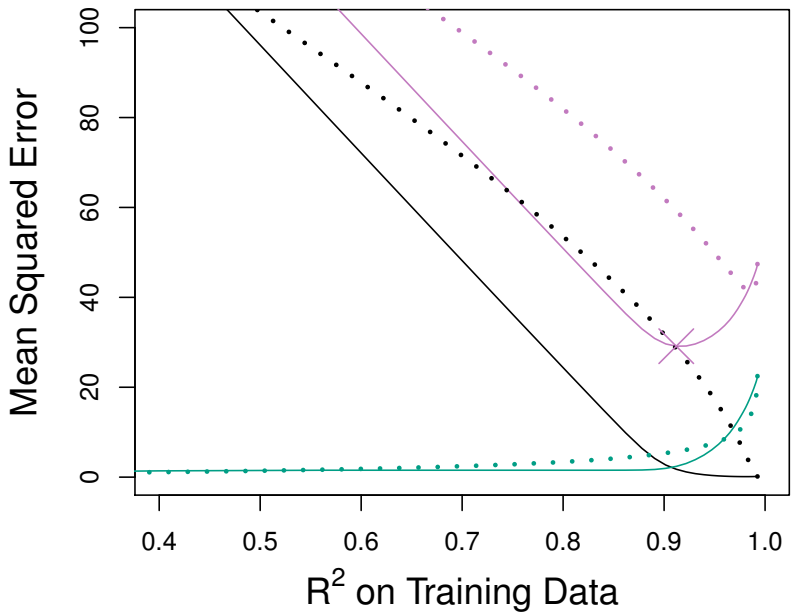
Lasso MSE:
Variance, squared bias, test MSE

Lasso (solid)
ridge regression (dotted)

The lasso vs. ridge regression: prediction accuracy

For a simulated data: $n = 50, p = 45$. Only 2 predictors are related to Y.





Lasso MSE:
Variance, squared bias, test MSE

Lasso (solid)
ridge regression (dotted)

The lasso vs. ridge regression: prediction accuracy

- These two examples illustrate that neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect the lasso to perform better when the response is a function of only a relatively small number of predictors.
- However, the number of predictors that is related to the response is never known for real data sets.
- A technique such as cross-validation can be used in order to determine which approach is better on a particular data set.

A special case: the normal means problem

Consider a linear regression, where $n = p$, $\mathbf{X} = \mathbf{I}_n$, no intercept, i.e.,

$$Y_i = \beta_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$$

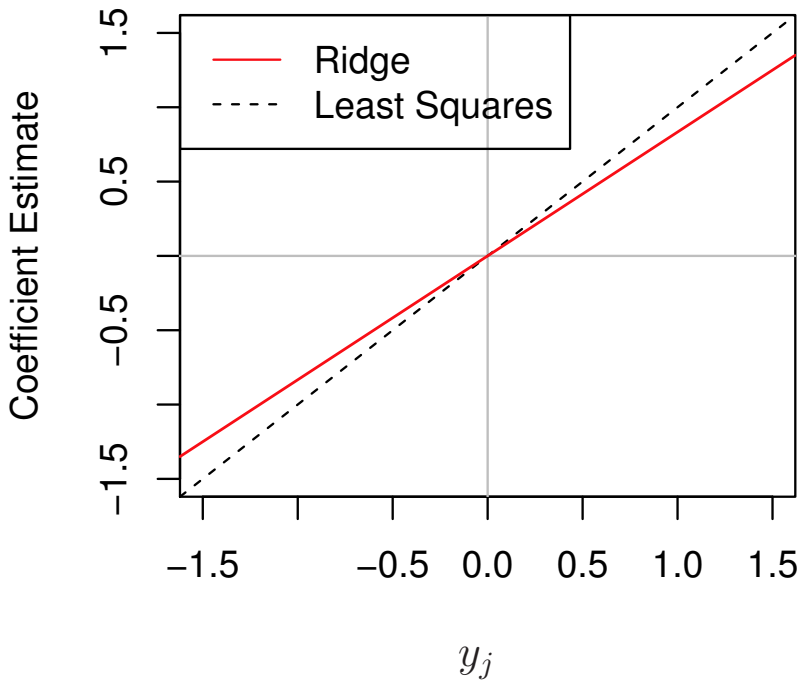
- RSS $= \sum_{i=1}^{n} (Y_i - \beta_i)^2$.
- OLS estimator: $\hat{\beta}_i = Y_i$
- Ridge estimator:

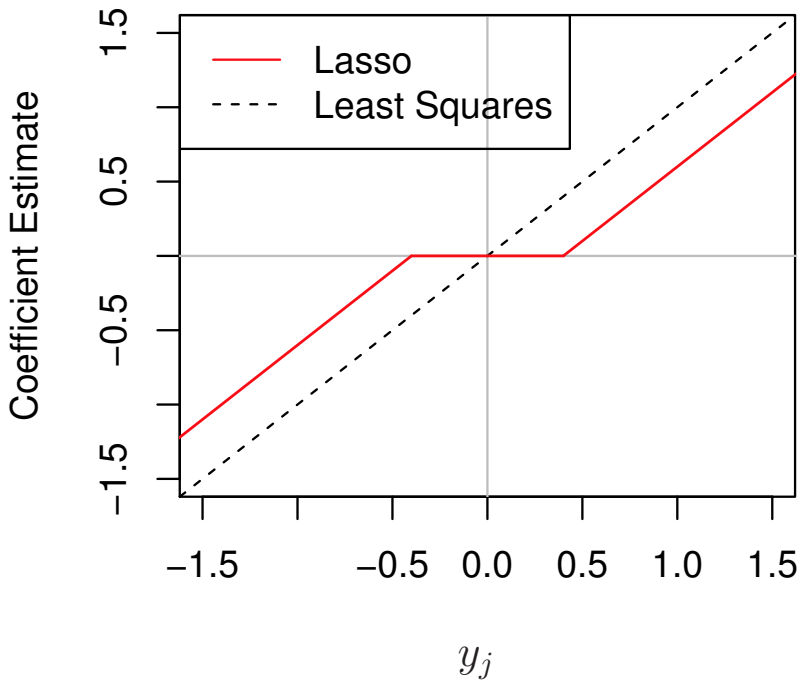
$$\hat{\beta}_i^{\text{ridge}} = \frac{Y_i}{1 + \lambda}$$

- Lasso estimator:

$$\hat{\beta}_i^{\text{lasso}} = \begin{cases} Y_i - \lambda/2 & \text{if } Y_i > \lambda/2 \\ 0 & \text{if } -\lambda/2 \le Y_i \le \lambda/2 \\ Y_i + \lambda/2 & \text{if } Y_i < -\lambda/2 \end{cases}$$

Proportional shrinking vs. soft thresholding





The elastic net: combining the ℓ_1 and ℓ_2 penalties

- Objective function is

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\left(\alpha\sum_{j=1}^{p}|\beta_j|+(1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$$

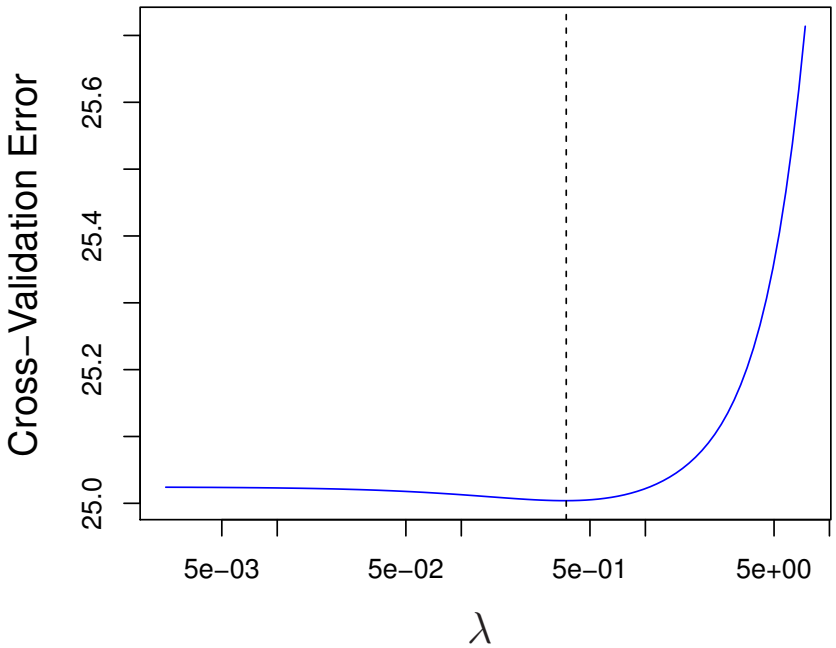
- The elastic net selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge.
- Parameter α:

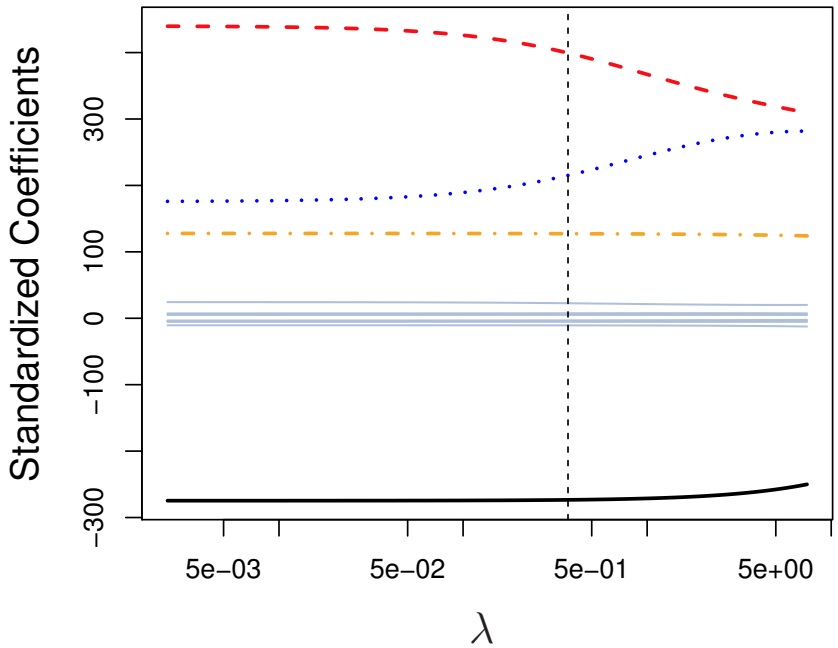
$$\alpha=0 \Longleftrightarrow \text{ridge regression}$$
$$\alpha=1 \Longleftrightarrow \text{lasso}$$

Selecting the tuning parameter λ: cross validation

- We choose a grid of λ values, and compute the cross validation error rate for each value of λ.
- We then select the tuning parameter value for which the cross validation error is smallest.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

The `Credit` data example: ridge regression





The simulation example: lasso

$n = 50, p = 45$. Only 2 predictors are related to Y.