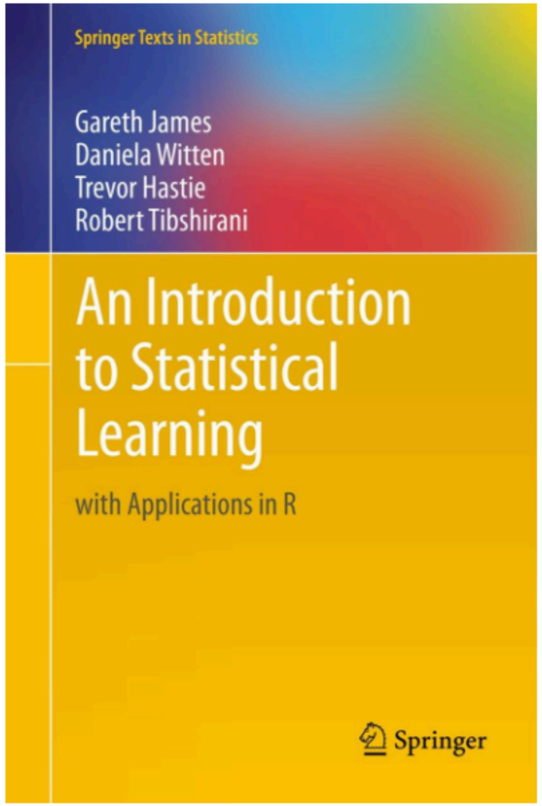
6. Regularization

ESC Spring 2018 – Data Mining and Analysis

SeoHyeong Jeong



Textbook:

An Introduction to Statistical Learning

Lecture Slides:

Stanford Stats 202: Data Mining and Analysis

Spring 17’ ESC Statistical Data Analysis

REFERENCE

Reading:

An Introduction to Statistical Learning

chapter 6.2 Shrinkage Methods

Table of Contents

Shrinkage Methods

The idea is to perform a linear regression, while regularizing or shrinking the coefficients $\hat{\beta}$ towards 0.

Why would shrunk coefficients be better?

- This introduces bias, but may significantly decrease the variance of the estimates. If the latter effect is larger, this would decrease the test error.
- There are Bayesian motivations to do this: the prior tends to shrink the parameters. (we don't go in to depth)

Ridge Regression

Ridge regression solves the following optimization:

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

In blue, we have the RSS of the model (the loss).

In red, we have the squared ℓ_2 norm of β, or $||\beta||_2^2$ (the **penalty** or **regularization**).

- Note that the intercept is not penalized, just the slopes!
- The parameter λ is a tuning parameter. It modulates the importance of fit vs. shrinkage.
- We find an estimate $\hat{\beta}_\lambda^R$ for many values of λ and then choose λ by cross-validation.

Bias-Variance Tradeoff

In a simulation study, we compute bias, variance, and test error as a function of λ .

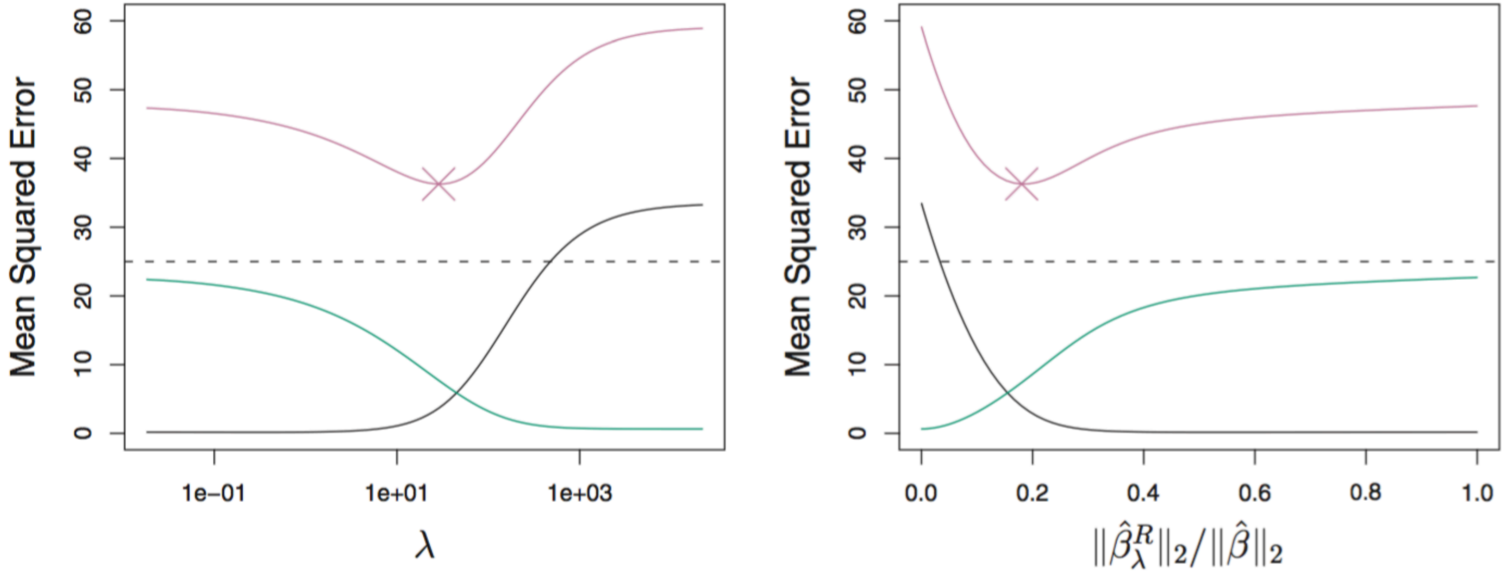


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of λ and $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Ridge Regression

- In least-squares linear regression, scaling the variables has no effect on the fit of the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

  Multiplying X_1 by c can be compensated by dividing $\hat{\beta}_1$ by c, i.e. after doing this we have the same RSS.
- In ridge regression, this is not true as the penalty term discourages large coefficients.
- In practice, what do we do?
  - Scale each variable such that it has sample variance 1 before running the regression.
  - This prevents penalizing some coefficients more than others.

Example. Ridge Regression

Ridge regression of `default` in the `Credit` dataset.

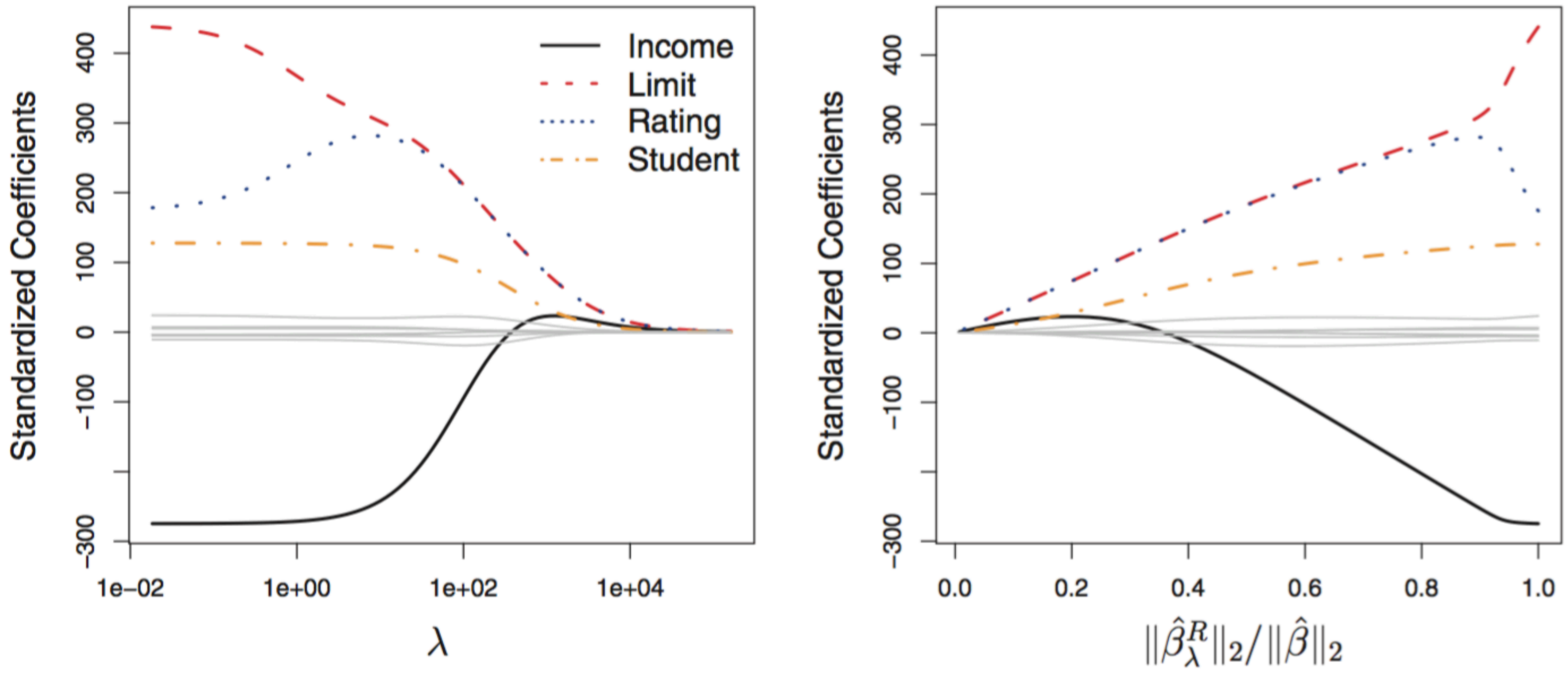


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Selecting λ by Cross-Validation

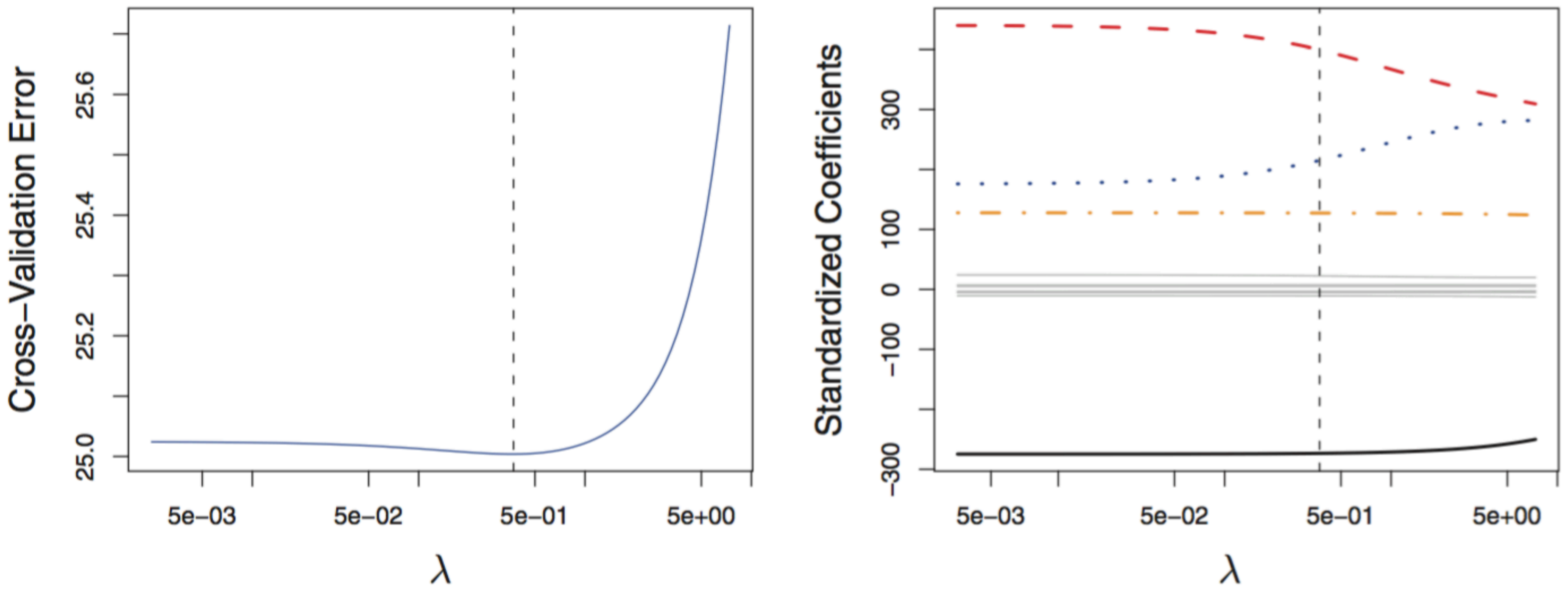


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

Why Does Ridge Regression Improve Over Least Squares?

Ridge regression works best in situations where the least squares estimates have *high variance*. (meaning that a small change in the training data can cause a large change in the coefficient estimates.)

When do we have high variance in a model?

- When the relationship between the response and the predictors is close to linear
- When the number of variables p is almost as large as the number of observations n.
- When $p > n$, then the least squares estimates do not have a unique solution.

The Lasso

Lasso solves the following optimization:

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

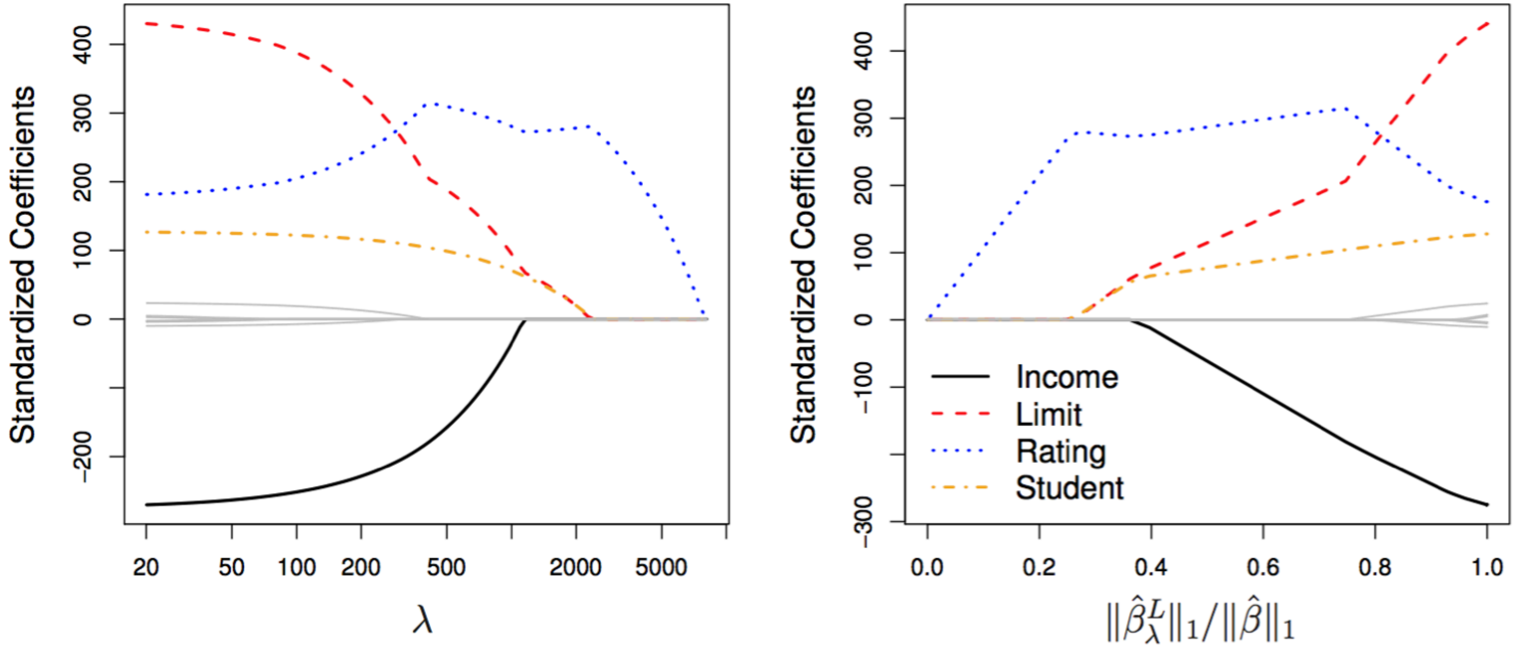
In blue, we have the RSS of the model (the loss).
In red, we have the squared ℓ_1 norm of β, or $||\beta||_1$ (the **penalty** or **regularization**).

Why would we use the Lasso instead of Ridge regression?

- Ridge regression shrinks all the coefficients to a non-zero value.
- The Lasso shrinks some of the coefficients all the way to zero. Alternative to best subset selection or stepwise selection.

Example. The Lasso

Lasso regression of `default` in the `Credit` dataset.



Comparing to pg. 9 (Example. ridge regression), Lasso regression shrinks coefficients to zero, whereas ridge regression shows a lot of pesky small coefficients throughout the regularization.

An Alternative Formulation for Regularization

- **Ridge:** for every λ, there is an s such that $\hat{\beta}^R_\lambda$ solves:

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

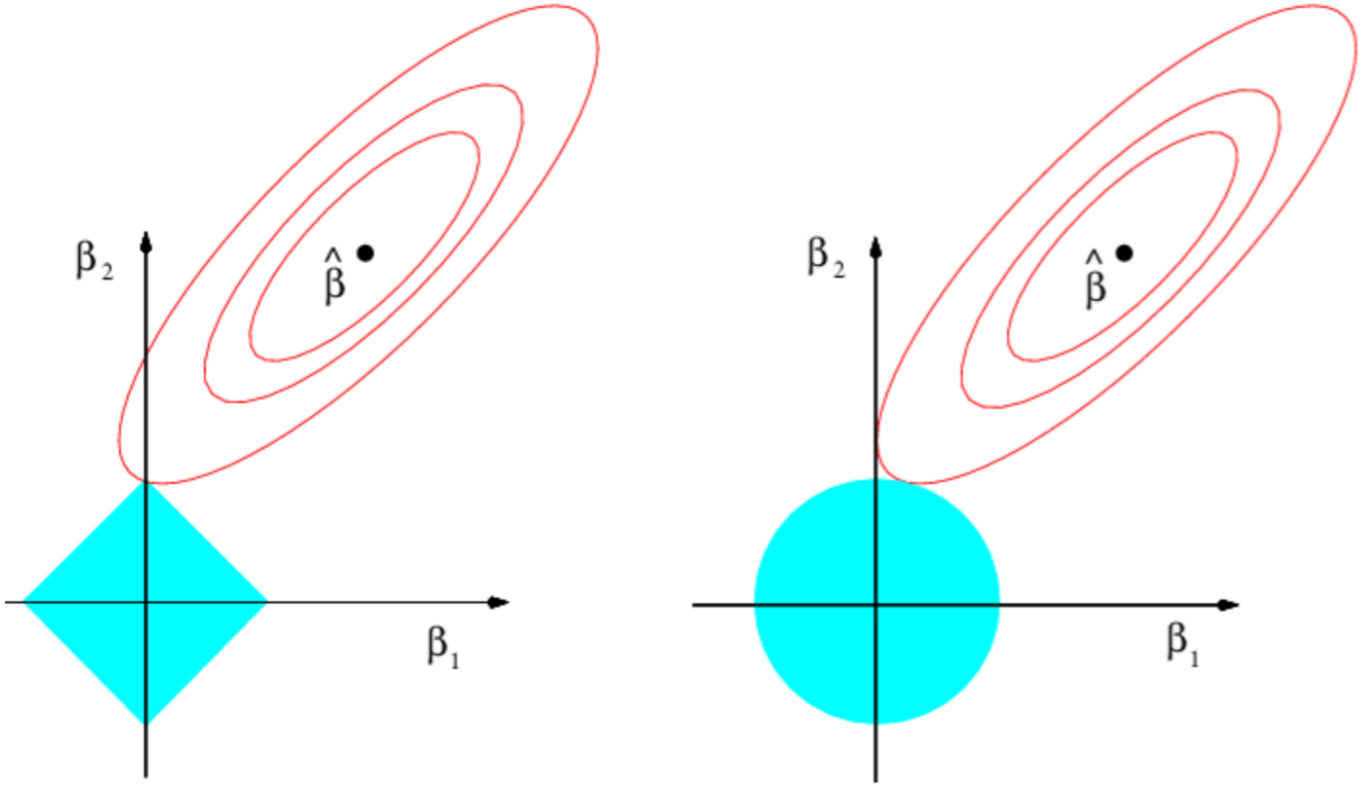
- **Lasso:** for every λ, there is an s such that $\hat{\beta}^L_\lambda$ solves:

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

- **Best subset:**

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \right\} \quad \text{s.t.} \quad \sum_{j=1}^{p} \mathbf{1}(\beta_j \neq 0) < s.$$

Visualizing Ridge and the Lasso with 2 predictors



The Lasso

◆ : $\sum_{j=1}^{p} |\beta_j| < s$

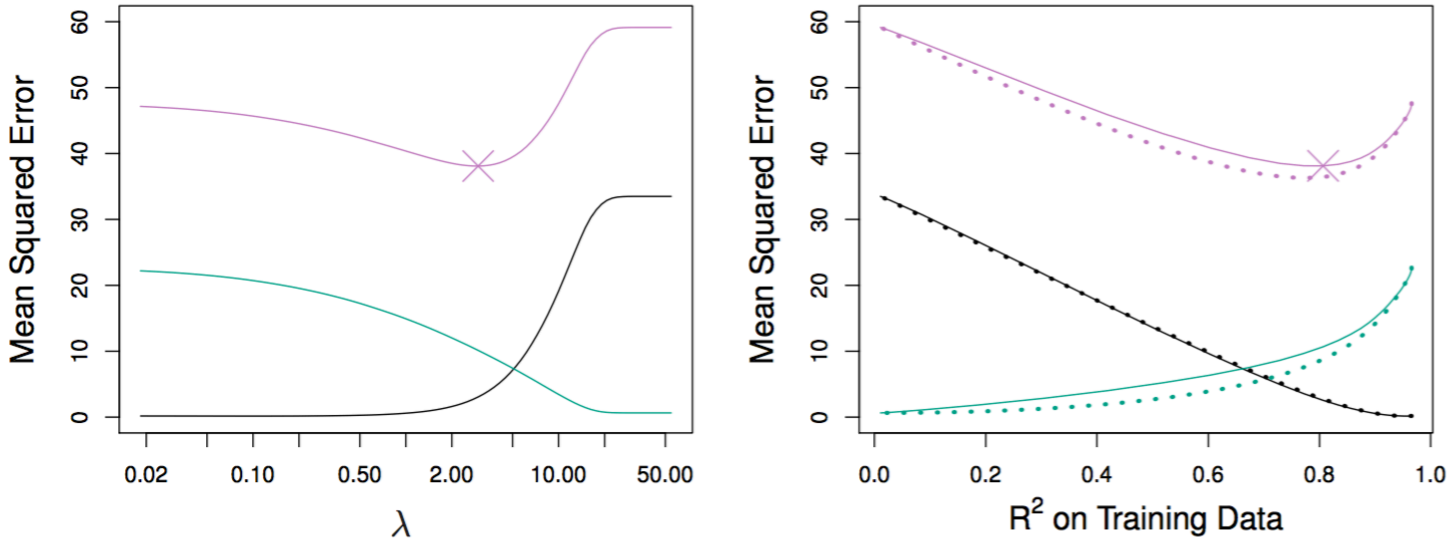
Ridge Regression

● : $\sum_{j=1}^{p} \beta_j^2 < s$

(Red ellipses are equal RSS contours)

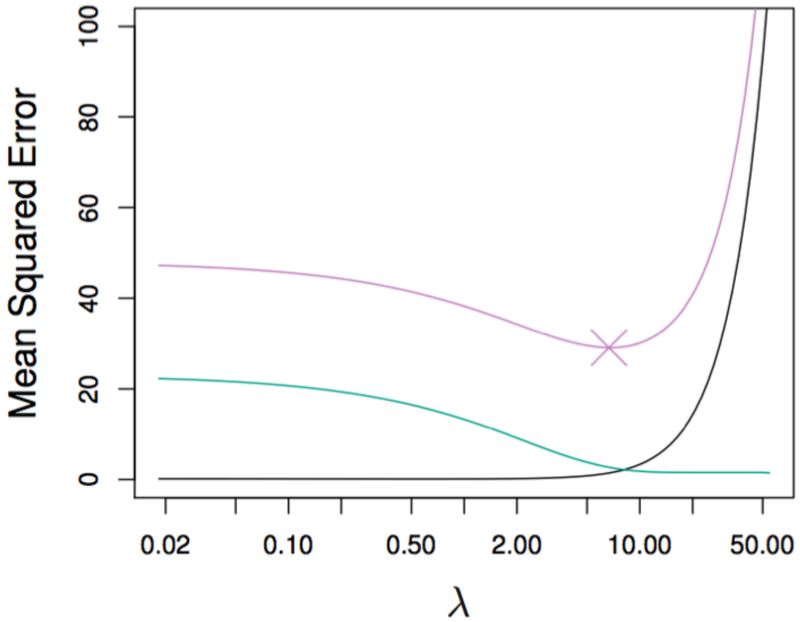
When is the Lasso better than Ridge?

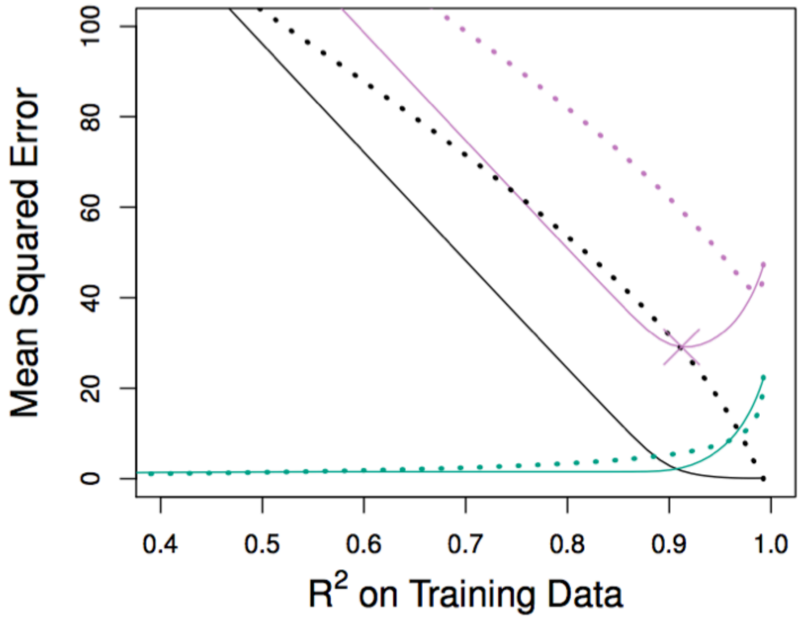
Example 1. All coefficients β_j are non-zero.



- Bias, Variance, MSE. The Lasso(—), Ridge(---).
- The bias is about the same for both methods.
- The variance of Ridge regression is smaller, so is the MSE.
- Rule: Lasso is typically no better than ridge for prediction when most variables are useful for prediction.

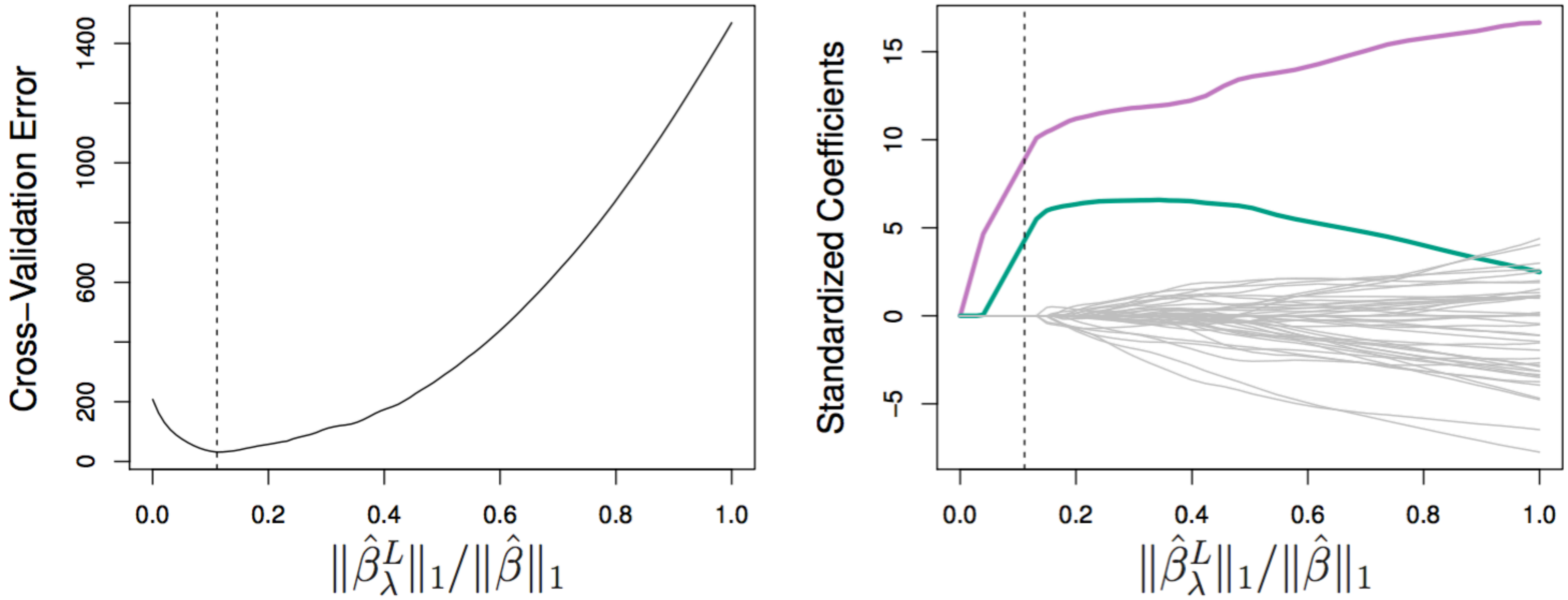
Example 2. Only 2 of 45 coefficients β_j are non-zero.





- Bias, Variance, MSE. The Lasso(—), Ridge(---).
- The bias, variance, and MSE are lower for the Lasso.
- Rule: Lasso is especially effective when most variables are not useful for prediction.

Choosing λ by Cross-Validation



A Special Case in Ridge Regression

Suppose $n = p$ and our matrix of predictors is $\mathbb{X} = I$.
Then, the objective function in Ridge regression can be simplified:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda\beta_j^2.$$

It is easy to show that

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda}.$$

A Special Case in The Lasso

The objective function is:

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

and we can minimize the terms that involve each β_j separately:

$$(y_j - \beta_j)^2 + \lambda|\beta_j|.$$

It is easy to show that

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| < \lambda/2. \end{cases}$$

Lasso and Ridge as a function of λ