

Lecture 14: Shrinkage

Reading: Section 6.2

STATS 202: Data mining and analysis

October 23, 2019

Shrinkage methods

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

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**).

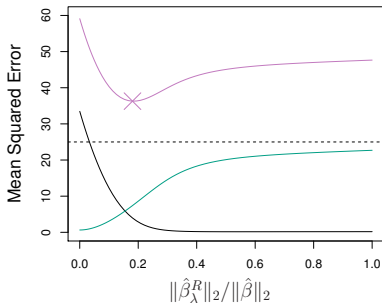
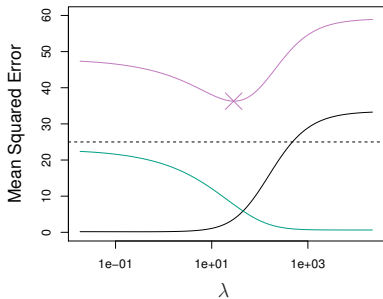
Note that the intercept is not penalized, just the slopes!

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 .



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 , ie. after doing this we have the same RSS.

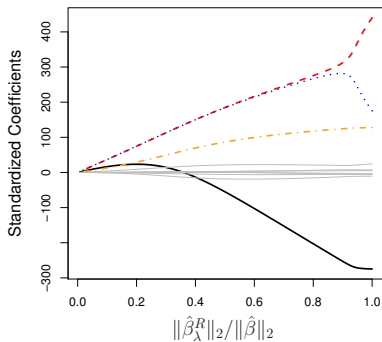
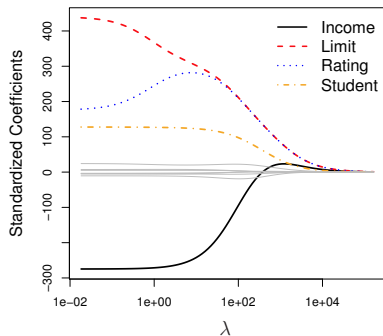
In ridge regression, this is not true as the penalty term discourages large coefficients.

In practice, what do we do?

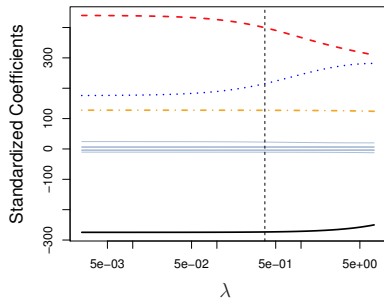
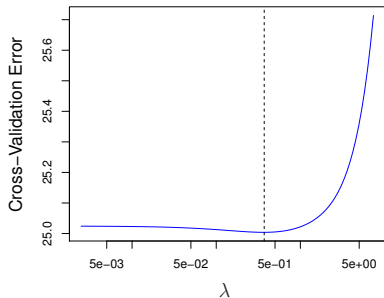
- ▶ Scale each variable such that it has sample variance 1 before running the regression.
- ▶ This prevents penalizing some coefficients more than others.

Example. Ridge regression

Ridge regression of default in the Credit dataset.



Selecting λ by cross-validation



The Lasso

Lasso regression solves the following optimization:

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

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

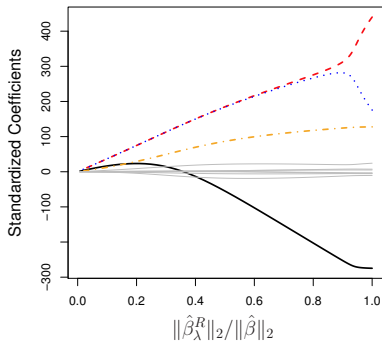
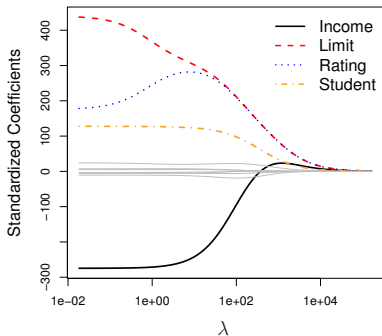
In red, we have the ℓ_1 norm of β , or $\|\beta\|_1$ (the **penalty** or **regularization**).

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. Ridge regression

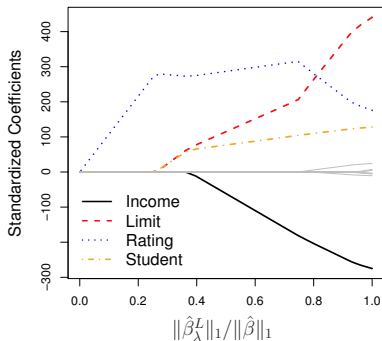
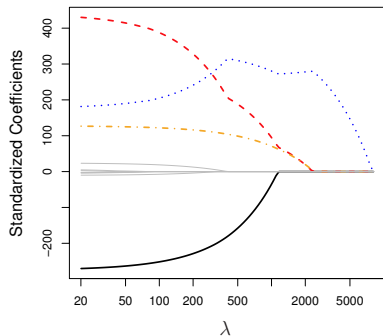
Ridge regression of default in the Credit dataset.



A lot of pesky small coefficients throughout the regularization path.

Example. The Lasso

Lasso regression of default in the Credit dataset.



Those coefficients are shrunk to zero.

An alternative formulation for regularization

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

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

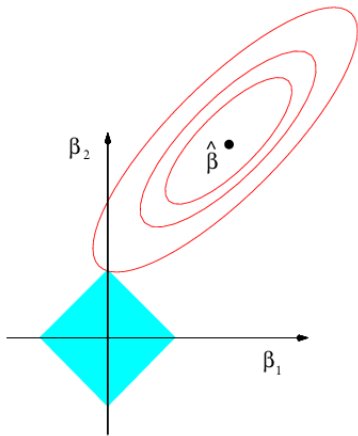
- **Lasso:** for every λ , there is an s such that $\hat{\beta}_\lambda^L$ solves:

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

- **Best subset:**

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

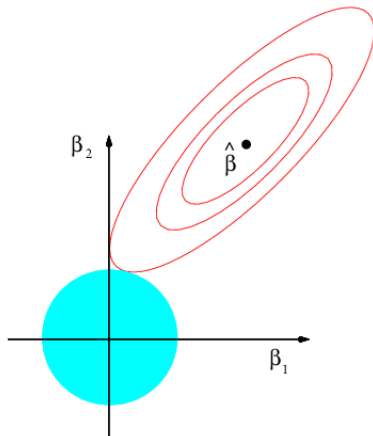
Visualizing Ridge and the Lasso with 2 predictors



The Lasso

◆ : $\sum_{j=1}^p |\beta_j| < s$

(Red ellipses are equal RSS contours)

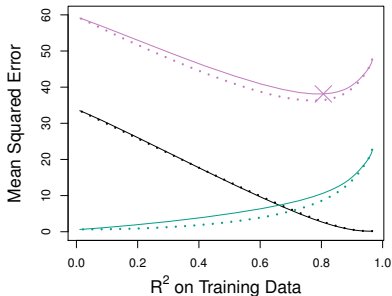
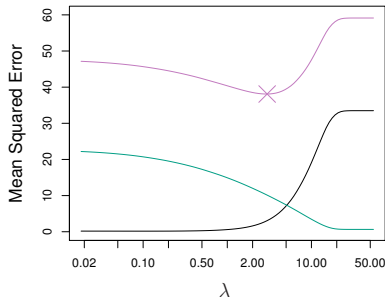


Ridge Regression

● : $\sum_{j=1}^p \beta_j^2 < s$

When is the Lasso better than Ridge?

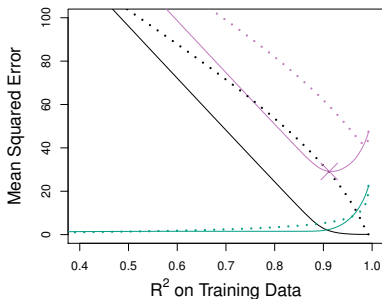
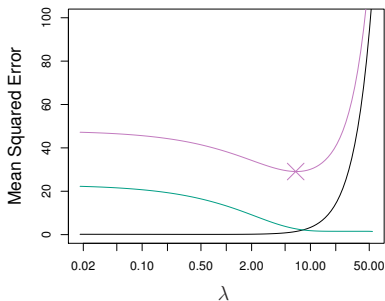
Example 1. All coefficients β_j are non-zero.



- ▶ 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 of thumb: Lasso typically no better than ridge for prediction when most variables are useful for prediction.

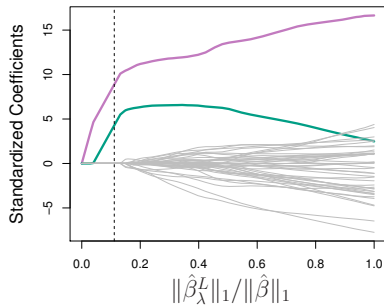
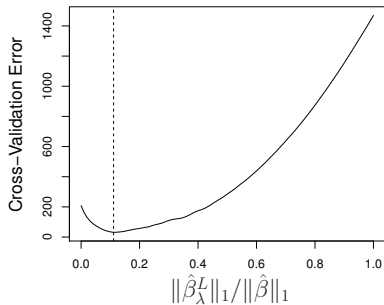
When is the Lasso better than Ridge?

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 of thumb: Lasso especially effective when most variables are not useful for prediction.

Choosing λ by cross-validation



A very special case

Suppose $n = p$ and our matrix of predictors is $\mathbf{X} = I$.

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 very special case

Similar story for the Lasso; the objective function is:

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

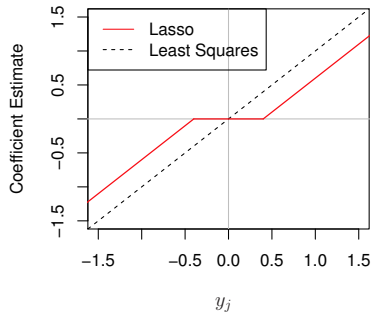
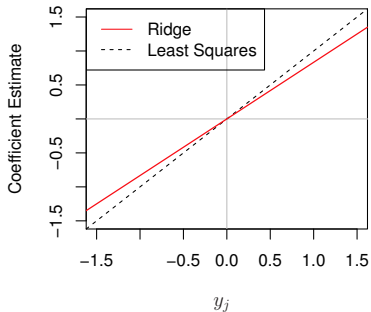
and we can minimize the terms that involve each β_j separately:

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

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 coefficients as a function of λ



Bayesian interpretations

Ridge: $\hat{\beta}^R$ is the posterior mean, with a Normal prior on β .

Lasso: $\hat{\beta}^L$ is the posterior mode, with a Laplace prior on β .

