Ch 6.2: Shrinkage - The Lasso

Lecture 17 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Mon, Oct 19, 2022

Last time:

- Ridge Regression

This time:

- The Lasso

Announcements:

- HW # 6 posted, due next friday
- No class this Friday

Section 1

Last time

Goal

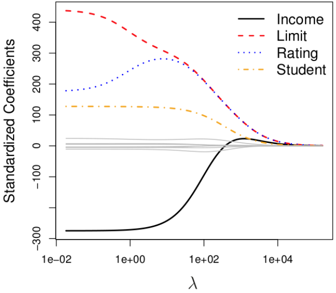
- Fit model using all p predictors
- Aim to constrain (regularize) coefficient estimates
- Shrink the coefficient estimates towards 0

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Ridge regression

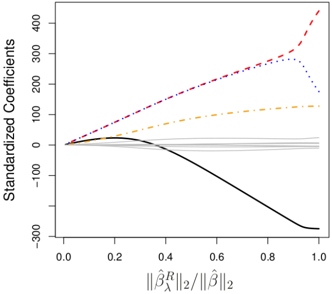
Before:

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



After:

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



Scale equivariance (or lack thereof)

Scale equivariant: Multiplying a variable by c (cX_i) just returns a coefficient multiplied by $1/c$ ($1/c\beta_i$)

Solution: standardize predictors

$$\widetilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

- Least squares is scale equivariant
- Ridge regression is not

Section 2

The Lasso

Same goal as before

- Fit model using all p predictors
- Aim to constrain (regularize) coefficient estimates
- Shrink the coefficient estimates towards 0

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

The lasso

Least Squares:

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

Ridge:

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

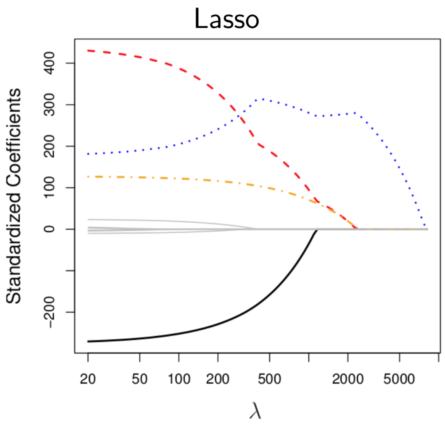
The Lasso:

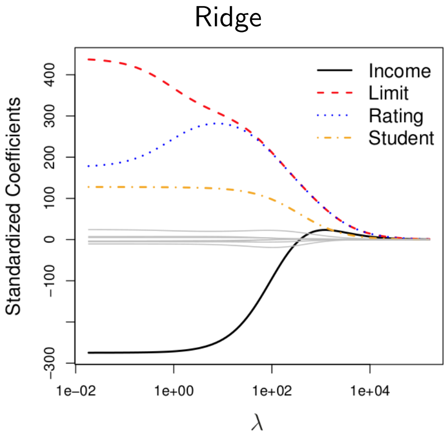
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

Subsets with lasso

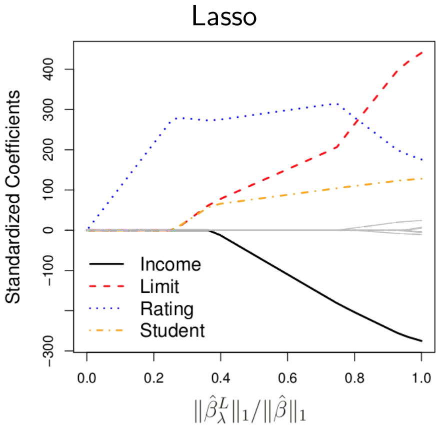
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

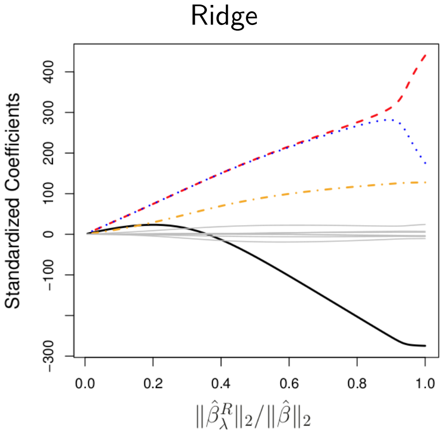
An example on `Credit` data set





More example on `Credit` data set





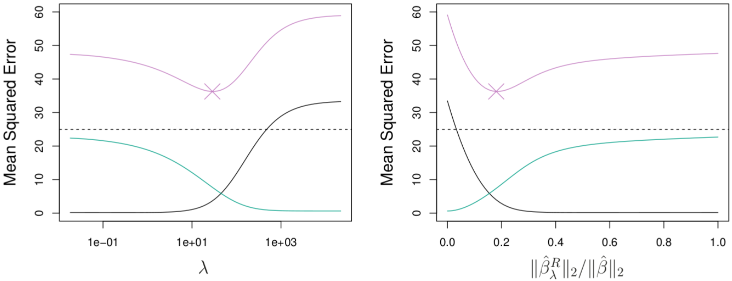
Scale equivariant: Multiplying a variable by c just returns a coefficient multiplied by $1/c$

Least squares **is** scale equivariant.
Ridge/Lasso **are very much not**.

Solution: standardize predictors

$$\widetilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

Bias-Variance tradeoff

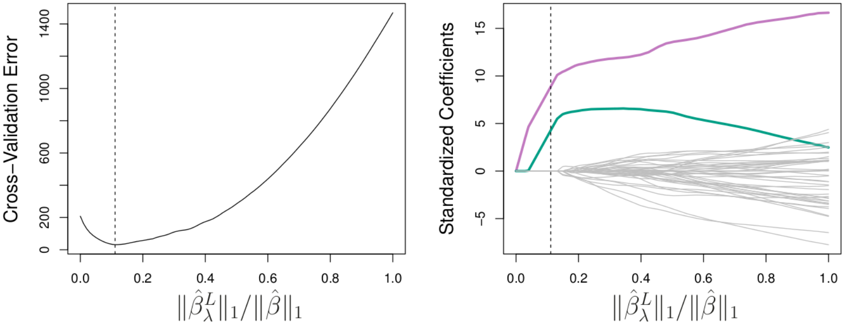


Squared bias (black), variance (green), and test mean squared error (purple) for simulated data.

Using Cross-Validation to find λ

- Choose a grid of λ values
- Compute the (k-fold) cross-validation error for each value of λ
- Select the tuning parameter value λ for which the CV error is smallest.
- The model is re-fit using all of the available observations and the selected value of the tuning parameter.

10-fold CV choice of λ for lasso and simulated data



Coding example

Section 3

Optimization Formulation

Another formulation for Ridge Regression

Find β to minimize:

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

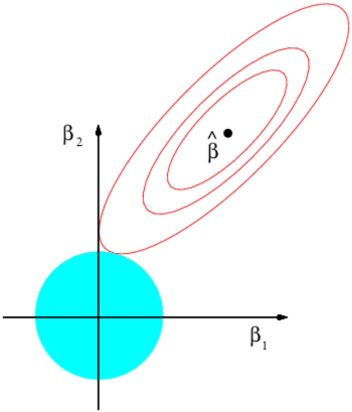
Find β to minimize

$$RSS$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq s$$

Visualization using disks



Find β to minimize

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

subject to

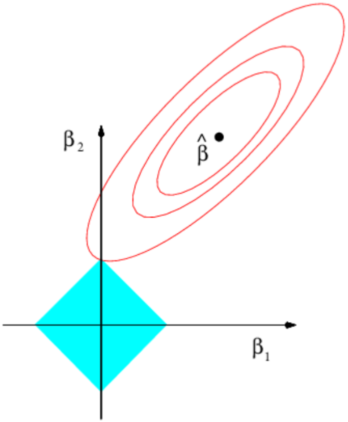
$$\sum_{j=1}^{p} \beta_j^2 \leq s$$

What about ℓ_1?

$$\|\beta\|_1 = \sum |\beta_i|$$

What does the set of points (β_1, β_2) for which $\|(\beta_1, \beta_2)\|_1 \leq s$ look like?

Same game for Lasso



Find β to minimize

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

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

Same game for subset selection

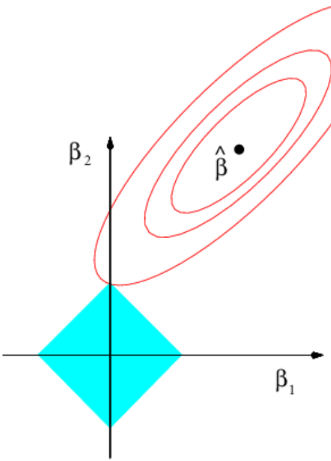
Find β to minimize

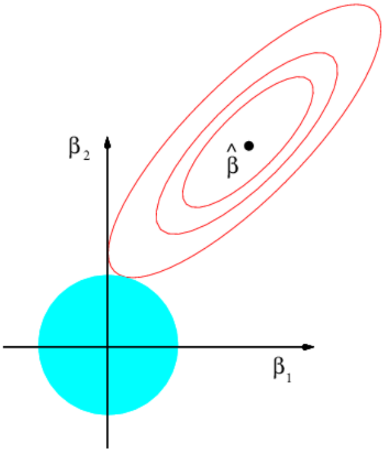
$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

subject to

$$\sum_{j=1}^{p} \mathrm{I}(\beta_j \neq 0) \leq s$$

Using this visual to understand why lasso gets us zero values





TL;DR - Original forumlation

Least Squares:

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

Ridge:

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

The Lasso:

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

TL;DR

Find β to minimize

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

subject to:

Least Squares:
No constraints

Ridge:

$$\sum_{j=1}^{p} \beta_j^2 \leq s$$

The Lasso:

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

Next time

10	M	Oct 3	Leave one out CV	5.1.1, 5.1.2	
11	W	Oct 5	k-fold CV	5.1.3	
12	F	Oct 7	More k-fold CV,	5.1.4-5	
13	M	Oct 10	k-fold CV for classification	5.1.5	HW #4 Due
14	W	Oct 12	Resampling methods: Bootstrap	5.2	
15	F	Oct 14	Subset selection	6.1	
16	M	Oct 17	Shrinkage: Ridge	6.2.1	HW #5 Due
17	W	Oct 19	Shrinkage: Lasso	6.2.2	
18	F	Oct 21	[No class, Dr Munch out of town]		
	M	Oct 24	No class - Fall break		
19	W	Oct 26	Dimension Reduction	6.3	
20	F	Oct 28	More dimension reduction; High dimensions	6.4	HW #6 Due
	M	Oct 31	Review		
	W	Nov 2	***Midterm #2***		