Ch 6.2: Shrinkage - The Lasso

Lecture 12 - CMSE 381

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

March 4, 2024

Announcements

Last time:

- Ridge Regression

This time:

- The Lasso

Announcements:

- Midterm score release tomorrow
- Can ask questions about grade before the end of recitation this week
- Mean: 72.5 Median: 72.5, curve?

Section 1

Last time

Goal

Test MSE = Train MSE + over fitting

- 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
- Lasso

subset selection

Before:

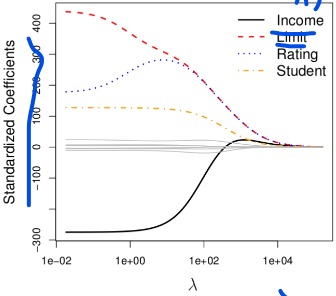
$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

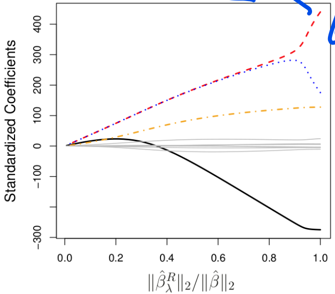
After:

hyper parameter

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

shrinking coeffs





Scale equivariance (or lack thereof)

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

$$\sum \left(y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2\right)^2 + \lambda\left(\beta_1^2 + \beta_2^2\right)$$

$1000\beta_1 \frac{X_1}{1000}$

- Least squares is scale equivariant
- Ridge regression is not

Solution: standardize predictors

centering

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

change unit $1 → $1000

- Ex: `income variable`. If you put in income by $1,000's of dollars, the answer you would get is quite different. Might emphasize a different coefficient.
- Least squares is scale equivariant
- Ridge regression very much is not
- $X_j\hat{\beta}^R_{j,\lambda}$ depends not only on λ but also on values of other predictors

Section 2

The Lasso

Same goal as before

- Fit model using all p predictors
- Aim to constrain (regularize) coefficient estimates

Ridge
- Shrink the coefficient estimates towards 0 *Ridge shrunk but DID NOT GET RID OF which is what we actually want*

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

- Ridge regression
- Lasso

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:

hyper parameter

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

$$\| \beta \|_{L_1} \text{ or } \|\beta\|_1$$

- ℓ_1 penalty instead of ℓ_2 penalty
- ℓ_1 norm: $\|\beta\|_1 = \sum|\beta_j|$
- ℓ_2 norm: $\|\beta\|_2^2 = \sum\beta_j^2$

$$\beta = \begin{bmatrix}\beta_1\\ \vdots \\ \beta_p\end{bmatrix}$$

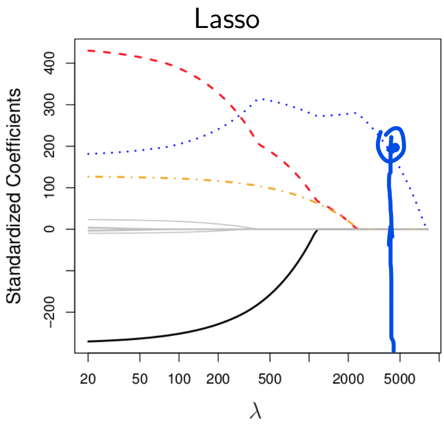
Subsets with lasso

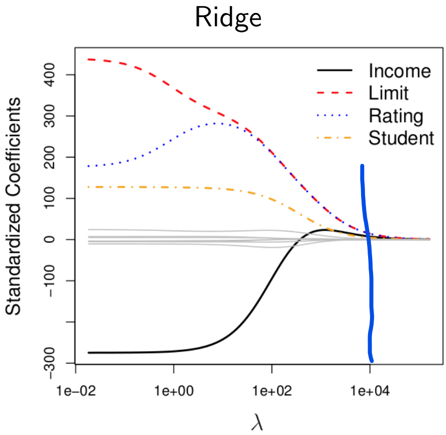
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

- Result of ℓ_1 penalty is that it forces some coefficients to be exactly 0
- Can be interpreted as something more like subset selection, but here called *variable selection*
- Yield are *sparse* models, that is those using only a subset of variables
- Selecing good λ is critical

- $\hat{\beta}$ from subset selec
- $\hat{\beta}^L$ $\leftarrow$ lots of components are 0
- $\hat{\beta}^R$ $\leftarrow$ none of component is strictly 0

An example on `Credit` data set





- $\lambda = 0$: Both are least squares
- $\lambda = \infty$: Everything is 0
-

Why Lasso shrinks coefficients to 0, - a mathematical verification on a simple example

$x_j := \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$

consider a simple special case with

- $n = p$,
- $x_j = e_j$ the j-th standard basis vector
- no intercept. $\beta_0 = 0$

$e_j := \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ ← jth

Then the usual least squares objective becomes

$$\min_{\beta_j} \sum_{i=1}^{p} (y_i - \beta_i)^2$$

The Lasso objective is

$$\sum_{i=1}^{p} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{p} |y_i|$$

the Ridge regression objective is

$$\sum_{i=1}^{p} (y_i - \beta_i)^2 + \lambda \sum_{i=1}^{p} y_i^2$$

- the least squares solution is $\hat{\beta}_j = y_j$
- the ridge solution is $\hat{\beta}_j^R = \frac{y_j}{1+\lambda}$
- the Lasso solution is

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2}, & y_j \geq \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2}, & y_j \leq -\frac{\lambda}{2} \\ 0, & |y_j| < \frac{\lambda}{2} \end{cases}$$

Scale equivavariance (or lack thereof)

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

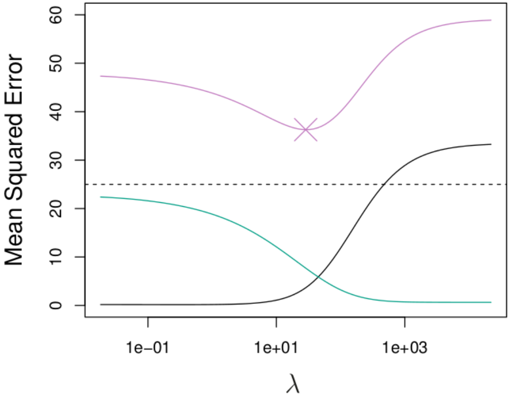
Least squares **is** scale equivariant.
Ridge/Lasso **are very much not**.

Solution: standardize predictors

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

- Ex: `income variable`.
- Least squares is scale equivariant
- Ridge regression very much is not
- $X_j \hat{\beta}^R_{j,\lambda}$ depends not only on λ but also on values of other predictors

Bias-Variance tradeoff



$\lambda \uparrow \iff$ complexity $\downarrow$

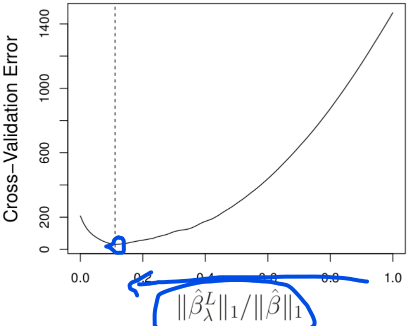
Squared bias (black), variance (green), and test mean squared error (purple) for simulated data.

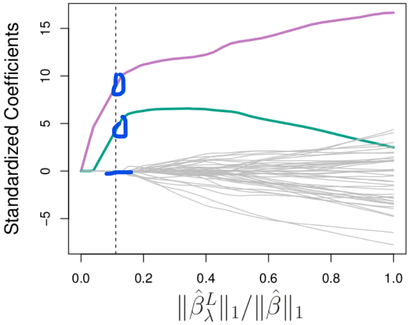
Horizontal dashed line is minimum possible test MSE

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





- Data generated with $p = 45$ variables, $n = 50$ data points, only 2 variables actually involved
- The vertical dashed line: choice of λ
- Right: Two colored lines are predictors related to the response (signal), grey are unrelated (noise)
- Correctly gives much larger coefficient estimates to the two signal predictors, but also the min CV error corresponds to a set of coefficient estimates for which only the signal variables are non-zero.
- Note that least squares solution doesn't even emphasize one of the variables (green one)

Section 3

Optimization Formulation

Another formulation for Ridge Regression

unconstrained

Find β to minimize:

penalty

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

$\Longleftrightarrow$

For any λ, there exists a s such that they have the same sol.

constrained

Find β to minimize

RSS

subject to

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

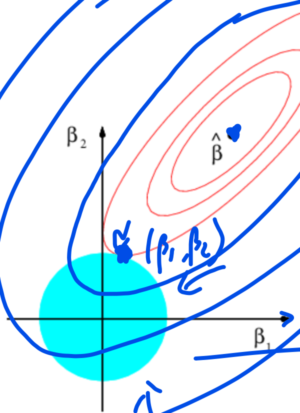
- For every λ, there is an s to make the right side solution the same as the one on the left.
- Think of s as a budget for how large $\|\beta_j\|_2$ can be.
- Large enough s and least squares answer is available, so would just return least squares.
- Smaller s means we might not get quite the optimal RSS, but do well enough. See next slide.

Visualization using disks

$\min_x f(x)$

f : RSS

x : β_0, β_j



(β_1, β_2)

$\beta_1 \neq 0$

$\beta_2 \neq 0$

$p = 2$

Find β to minimize

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

$j = 1, p$

subject to

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

budget

$\beta_1^2 + \beta_2^2 \leq s$

- Red lines are the RSS levelsets
- Blue ball has radius s^2.
- If s is big enough, encloses the minimum of RSS function
- Smaller s means have to be in the blue ball

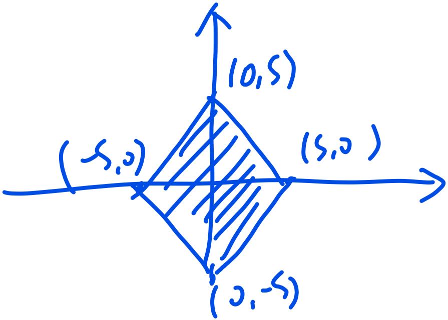
What about ℓ_1?

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

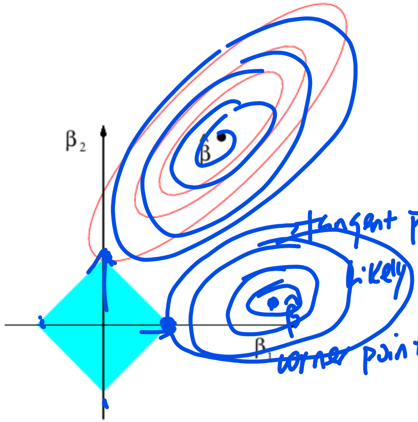
$\beta_1^2 + \beta_2^2 \le S$

$|\beta_1| + |\beta_2| \le S$

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



Same game for Lasso



tangent points → likely to be on the corner

corner point → lies along axis

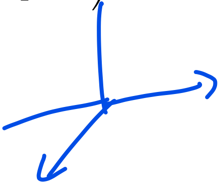
- Diamond sides at $(0, \pm s)$, $(\pm s, 0)$

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

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



- If s is big enough, encloses the minimum of RSS function
- Smaller s means have to be in the blue ball
- Min likely to happen at corner, on axis, where something is 0

Same game for subset selection

Find β to minimize

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

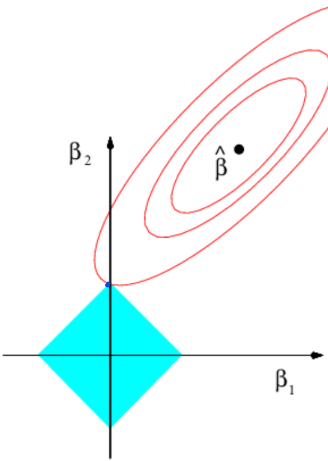
subject to

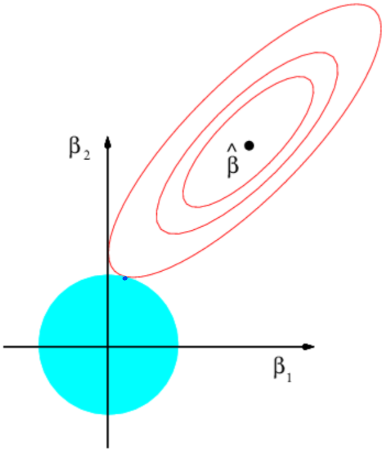
$$\sum_{j=1}^{p} \mathrm{I}(\beta_j \neq 0) \leq s$$

NP-hard.

- Similar format but computationally infeasible since still requires checking all subsets containing s predictors
- Interpretation: ridge regression and lasso are computationally feasible versions of subset selection

Using this visual to understand why lasso gets us zero values





Because ridge regression has circular constraint with no sharp points, intersection doesn't happen on axis line. But because of the diamond shape of ℓ_1, much better chance of that.

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:

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

Summary

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:

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

Also, find best choice of λ or s using CV