

# The Lasso Method of Parameter Selection

C. Kinstley, A. Nabar, T. Williams, C. Vollmer

April 26, 2017



# Improving OLS

- Recall the ordinary least squares procedure (OLS) estimates unknown coefficients in a linear regression model by minimizing the **squared difference** between the **predicted** and **actual** responses.
- We would like to fit  $n$  data points to a model of the form

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

- Let  $X$  denote the  $n \times p$  design matrix where the  $x_{ij}$ th entry is the  $i$ th point of sample data corresponding to the  $j$ th dependent variable and  $Y$  the response vector
- So long as  $n \geq p$  OLS gives us a best-fit hyper-plane as follows:

- The **difference** between the actual and predicted response for each data point  $i \in 1 \dots n$  is

$$\epsilon_i = Y_i - X\beta$$

- We minimize the sum of the squared differences, i.e minimize the function

$$E(\hat{\beta}) = \sum_{i=1}^n \epsilon_i^2 = \|Y - X\beta\|^2$$

- It can be shown that the minimizers are  $\hat{\beta} = (X^T X)^{-1} X^T Y$  but these solutions are usually approximated

# Problems with OLS

Two common problems with OLS estimates are as follows:

- Imprecision
- The expected value of the error term is 0 but the variance may be large.
- How does this happen? Recall that the least squares estimation  $\hat{\beta} = (X^T X)^{-1} X^T Y$ .
- If  $(X^T X)$  is near-singular, then small changes in the  $X$  might lead to large changes in  $\hat{\beta}$ .
- So, even if our  $\hat{\beta}$  fits one sample well, there is no guarantee it will fit other samples well, let alone the population!

- Interpretation
- A large number of independent variables can make the model difficult to interpret, especially when we want to isolate the "most important" variables.
- Do we care about variables with very small coefficients?

# Improving OLS: Subset Selection

## Subset Selection

- Simply ignore one or more of the independent variables! That is, set the coefficient(s) to 0.
- This helps with interpretability, if only because there is less to interpret.
- Drawback: Subset Selection is a discrete process. Regressors are either kept or dropped; there is no in-between. Small changes in the sampling data can thus result in very different models.
- Not computationally practical for high dimensional data

# Improving OLS: Ridge Regression

## Ridge Regression

- In Ridge regression we again minimize  $E(\beta) = \|Y - X\beta\|^2$
- But subject to the constraint that

$$\sum_{i=1}^{p-1} \beta_i^2 = \|\beta\|_2^2 \leq t \text{ for } t \geq 0$$



# Improving OLS: Ridge Regression

## Ridge Regression

- This is equivalent to the unconstrained minimization of

$$E^R(\beta) = \|Y - X\beta\|^2 + \lambda \|\beta\|^2$$

where  $\lambda$  is a function of  $t$

# Improving OLS: Ridge Regression

## Ridge Regression

- In terms of the coefficients,  $\hat{\beta}^R$ , it is equivalent to adding a small constant value  $\lambda$  to diagonal entries of  $(X^T X)$  in the OLS solution
- This prevents the matrix from being singular or near-singular.
- Drawback: Reduces the variance, but does not set any coefficients to 0, so it doesn't help with interpretability. It also adds bias.

# Enter the Lasso

- L.A.S.S.O: Least Absolute Shrinkage and Selection Operator.
- Like Ridge we minimize  $E(\beta)$  under a constraint, the lasso coefficients are found by minimizing

$$E^L(\beta) = \|Y - X\beta\|^2 \text{ subject to } \|\hat{\beta}\|_1 \leq t$$

- We are using the 1 norm, instead of the 2 norm used in ridge
- As in Ridge this can be written as an unconstrained optimization problem with  $\lambda(t)$

# Variable Selection

- Lasso has the desirable property that it will usually set some of the coefficients equal to zero
- Like in subset selection this leads to a more interpretable model overall
- In Ridge and OLS the minimizing coefficients are almost surely non-zero

# Why is this?

Consider the constrained optimization problems for lasso and ridge in the case of two variables

*pictureofconstrainedproblem*

- As in Ridge and subset selection the Lasso enjoys lower variance in the magnitude of coefficients than OLS
- Why? I'm not sure...

# Tuning parameter

- $t$  or  $\lambda(t)$  determines the amount of shrinkage we apply to the OLS estimates

- Let  $\hat{\beta}^o = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$  be the OLS coefficients

- Define  $t_o = \sum_{j=1}^p |B_j|$ , at  $t \geq t_o$  we recover the OLS estimates

# something

- Setting  $t \leq t_0$  will **shrink** the solutions toward 0.
- For example, setting  $t = \frac{t_0}{2}$  is similar to selecting the best subset containing half of the regressors.



# Why is this?

In the 2-D case this looks like

*pictureofconstrainedproblem*

# Lasso vs. Subset Selection vs. Ridge Regression

- Small number of large effects.
  - Subset selection is best, Lasso is second best, Ridge Regression is worst.
- Small to moderate number of moderate-sized effects.
  - Lasso is best, Ridge regression is second best, Subset selection is worst.
- Large of small effects.
  - Ridge regression is best, Lasso is second best, Subset selection is worst.

# Lasso vs. Subset Selection vs. Ridge Regression

- Do these results make sense?
  - Recall that the Lasso was designed to work like Ridge regression but with some of the benefits of Subset selection.
  - It thus makes sense for the Lasso to fall between Ridge regression and Subset selection on extreme cases and to beat both of them on cases not well-suited to either.
- **CAUTION!**
  - These results refer to **prediction accuracy**
  - As for **interpretability**:
    - Subset selection  $>$  Lasso  $>$  Ridge regression, always
    - Why? More nonzero coefficients = more interpretable, always