

Reducing Model Complexity

Hossein Falaki
Computer Science Department,
University of California, Los Angeles
falaki@cs.ucla.edu

1 Introduction

Too complex models are prone to *over-fitting* risk, i.e., they might interpret the noise or coincidence. Therefore methods and principles to reduce model complexity are needed.

If the regressors are orthogonal, finding the best subset of estimators is fairly easy using soft or hard thresholding. The problem of finding the optimal subset of estimators for arbitrary regressors is NP-hard, because the number of possible subsets increases exponentially with the model complexity, p .

In this note we introduce two techniques for finding the approximate optimal subset of estimators, and the underlying principles. To bound the model complexity we will constrain $\|\beta\|$. We will study L_2 norm for constraining $\|\beta\|$ and the resulting method which is called *ridge regression*. We will also study two methods with L_1 norm: *small-step stagewise regression* and *least angle regression* and the underlying principle that is called *LASSO*.

2 Ridge Regression

The easiest way to constrain $\|\beta\|$ is using the L_2 norm, thus the regression would be:

$$\text{minimize } \|Y - X\beta\|^2 \text{ subject to: } \|\beta\|^2 < t \quad (1)$$

Solving this constraint optimization is equivalent to solving the following Lagrangian equation:

$$\text{minimize } \|Y - X\beta\|^2 + \lambda\|\beta\|^2 \quad (2)$$

Which is straight forward:

$$\begin{aligned} \frac{\partial}{\partial \beta_j} : 2X(Y - X\beta) + 2\beta_j\lambda &= 0 \\ \Rightarrow \tilde{\beta} &= (X^T X + \lambda I)X^T Y \end{aligned}$$

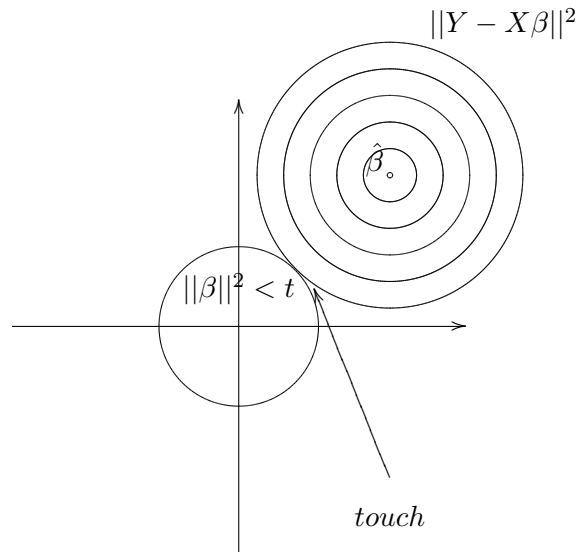


Figure 1: Visualization of an L_2 regularization

This equation is illustrated in Figure 1. The solution to the Lagrangian equation is where the contour of $\hat{\beta}$ touches the $\|\beta\|^2 < t$ boundary.

Ridge regression is a general form of *shrinkage*. In other words it guarantees that the model will not arbitrarily follow noise, but it does not limit the number of estimators. Therefore if the number of regressors is too high, we will end up with a large number of estimators each one of which is relatively small.

As mentioned before, finding the optimal subset of these β_j s to keep and set the rest to zero is an NP-hard problem. Using the L_1 norm will give an approximation of the optimal subset of estimators to keep.

3 LASSO

The *Least Absolute Shrinkage & Selection Operator* uses a square shape regularize $\|\beta\|_1$. This can be achieved by using the L_1 norm for constraining $\|\beta\|_1$:

$$\text{minimize } \|Y - X\beta\|^2 \text{ subject to: } \|\beta\|_1 < t \quad (3)$$

Where:

$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

It can be easily seen in Figure 2 that the contour will touch the limit mostly in the corner points. At these points, some of β_j s are zero, therefore

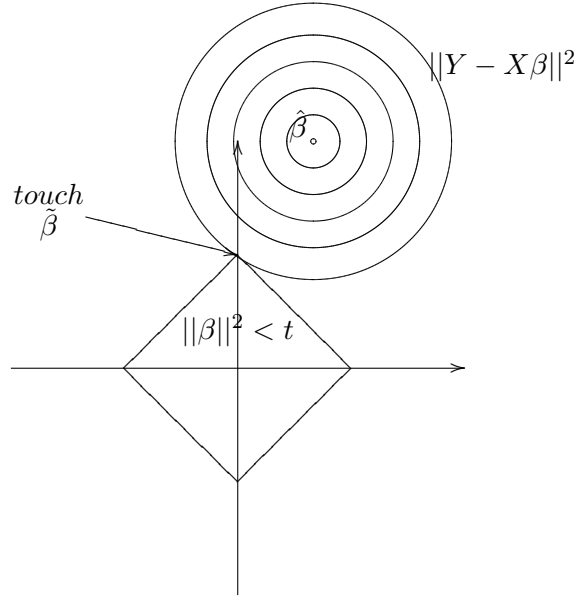


Figure 2: Visualization of an L_1 regularization

LASSO is performing shrinkage and subset selection at the same time. In what follows we will demonstrate and argue that LASSO results a close to optimal solution.

3.1 Small-step Stage-wise Regression

Stage-wise regression minimizes the residual of regression in small steps in an effort to find the most effective estimators. Assuming that $\|X_j\|^2 = 1$ details of the algorithm follows:

1. $\beta = 0, R = Y$
2. Chose X_j that maximizes $|\langle X_j, R \rangle|$:
 $\beta_j = \beta_j \pm \delta$ (+ or - depending on the sign of $\langle X_j, R \rangle$)
 $R = R \pm \delta X_j$
3. Go to 1

By choosing a small enough δ this algorithm guarantees that at each stage:

$$\forall i, j \in \text{active set} : |\langle X_i, R \rangle| \doteq |\langle X_j, R \rangle| \quad (4)$$

As the algorithm runs, more X_i s will be added to the active set and the complexity of the model increases. We can find the optimal active set through

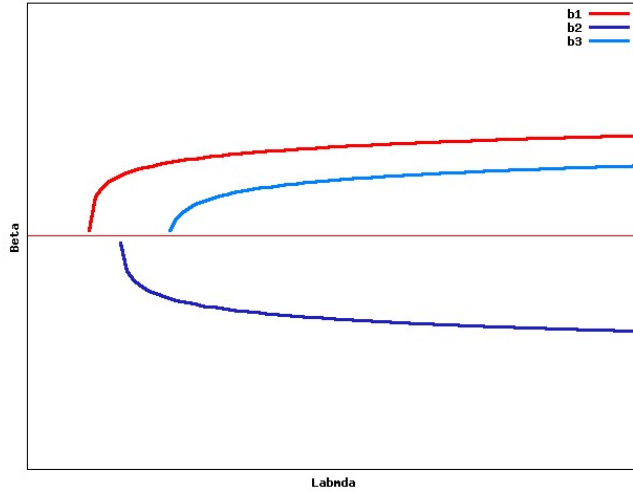


Figure 3: β_j values as λ increases

cross-validation ¹.

The observation in Equation 4 can be used to increase the speed of the regression process. This new method is called *least angle regression*.

3.2 Least Angle Regression

Instead of increasing β_j s in the active set one step at a time we can use the following equation:

$$\forall i \in \text{active set} : | \langle X_i, R \rangle | = \lambda \quad (5)$$

We can increase λ until a new β_j enters the active set as seen in Figure 3. Similar to the stage-wise regression, we can find the optimal active set of β_j s as the final solution through cross-validation.

3.3 Relation with LASSO

We can show why stage-wise regression with small steps and least angle regression are in effect implementing LASSO. For both these methods:

$$\begin{aligned} | \langle X_j, (Y - X\beta) \rangle | &= \lambda && \text{for active } \beta_j \\ \beta_j &= 0 && \text{for inactive } \beta_j \end{aligned}$$

This setting is equivalent to the solution of the LASSO equation 3. Suppose β_j is picked by LASSO and without loss of generality assume $\beta_j > 0$.

¹Dividing the training set into “training” and “testing” subsets, and testing the result of the learning algorithm on the training subset using the testing subset.

Then:

$$\frac{\partial}{\partial \beta_j} : 2X_j(Y - X\beta) + \lambda' = 0$$

In other words, stage-wise regression and least angle squares are in practice solving Equation 3 and thus performing LASSO.

4 Acknowledgment

This note is based on Prof. Ying Nian Wu's lectures on Theoretical Statistics at UCLA.