# Stat 151a Lecture Notes 22

Chae Yeon Lee

April 18, 2023

## 1 Overview

Today:

- Shrinkage (Finish Ridge and Start on Lasso)
- Generalized Linear Models

Relevant Readings:

- James et al. 6.2
- Fox 15.1

# 2 Shrinkage Methods, Continued

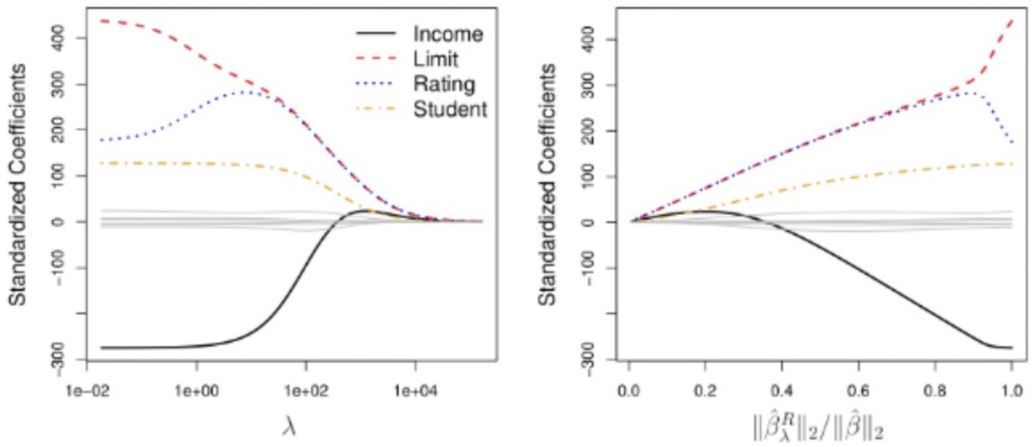

**FIGURE 6.4.** *The standardized ridge regression coefficients are displayed for the* Credit *data set, as a function of $\lambda$ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.*

As $\lambda$ increases, all coefficients shrink to zero. Also, there is no monotonicity. For blue (Rating), the coefficient increase before decreasing. Therefore, there is no guaranteed trajectory of t coefficients.

## 2.1 Ridge Regression

**Ridge Pros:**

- Analytical solution for each $\lambda$. (i.e. there is a linear form of matrix multiplication to get optimal solution). However, there isn't one for Lasso.
- Theoretically, it can always outperform OLS.
- No matter how many coefficients we have, we can run for Ridge.

**Ridge Cons:**

- There is no explicit variable selection. (no 0's in the $\vec{\beta}$)
- There is no clear inference approach. (i.e. adding regularization term to the model, we don't have understanding of how to do likelihood model around Ridge / Lasso.) Implicitly, it is a Bayesian framework with a "shrinkage" prior.

## 2.2 Lasso Regression

Lasso does not have closed form solution

$$\min_{\vec{\beta}} \|\vec{y}* - Z\vec{\beta}\|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

We cannot compute optimal $\hat{\beta_\lambda}$ on paper. However, it is still convex, and so there is a single local optimum, and there are fast numerical solvers. If numerically or iteratively, I converge on a local optimum, then that is the global optimum.
These are some numerical approaches:

- Coordinate descent: optimize the problem one parameter at a time. Fix all other $\beta$'s and then optimize each $\beta$ one at a time.
- Iterative WLS
- Proximal Gradient

*glmnet()* is the go-to package for solving LASSO problems.
Lasso coefficients will actually shrink to zero. Therefore, we will have explicit variable selections.

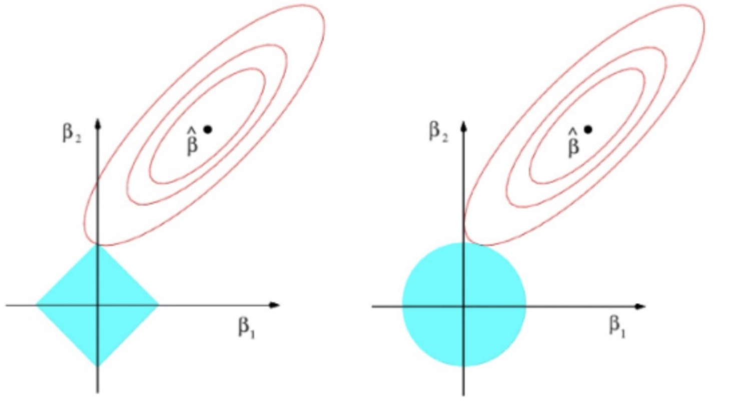

**FIGURE 6.7.** *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \le s$ and $\beta_1^2 + \beta_2^2 \le s$, while the red ellipses are the contours of the RSS.*

$\beta$ is the OLS solution. Level sets, $\|\vec{y} - X\vec{\beta}\|^2$, are ellipses. Often times, Lasso will touch the tip of diamond, which is why Lasso solution tend to zero out. On the other hand, Ridge solutions will not zero out the coefficients.
**Lasso Pros:**

- Performs Variable Selection
- Is convex / computationally feasible

**Lasso Cons:**

- There is no "linear / analytical" solution to analyze.
- Still no clear (non-Bayesian) inference direction. (Though you can refit OLS on the selected variable from LASSO. However, this refitting is an issue for inference because we have already used data to select these features. [post-selection inference] In practice, if you have a lot of data, then you do variable selection on a portion of the data and run the selected model on the remaining data.)

In practice, we tune the optimal $\lambda$ value for Lasso or Ridge is using cross-validation.

### 2.3 Introduction to GLM

To this point, we've assumed that $\vec{y}$ is continous. What if outcome is binary?
For example, I could be trying to:

- predict if a tumor is benign or not
- predict if a team will win or lose a game

The model $\vec{y} \sim N(x\vec{\beta}, \sigma_\epsilon^2 \mathbb{1}_n)$ doesn't really make sense.
Issues: Let's say that any model we create predict the mean of $y_i$ conditional on $\vec{X}_\lambda^T$.
Then, $E[y_i|\vec{X_i}^T] = P(y_i = 1|\vec{X_i}^T) = \pi_i$
The problems are:

- The variance of $\vec{y}$ is no longer constant. The variance of the Bernoulli random variable $var(y_i|\vec{x_i}^T) = \pi_i(1-\pi_i)$.
- The errors are now dichotomous. If $y_i = 1$, then $\epsilon_i = 1 - E[y_i|\vec{x_i}^T] = 1 - \pi_i$. If $y_i = 0$, then $\epsilon_i = 1 - E[y_i|\vec{x_i}^T] = -\pi_i$
- If we model $E[y_i|\vec{x_i}^T] = \vec{x_i}^T\vec{\beta}$ then we may generate predictions outside of [0,1].

**Generalized Linear Model (GLM)** will alleviate these concerns. GLM is a framework to allow $\vec{y}$ to follow various parametric distributions.

- **Bernoulli:** $\vec{y}$ is binary
- **Poisson, Negative Binomial:** $\vec{y} \in \mathbb{Z}^+$ (count data)
- **Exponential, Gamma:** $\vec{y} \in \mathbb{R}^+$ (continuous data)
- **Multinomial:** $\vec{y}$ is a count of occurences of K different outcomes.

#### 2.3.1 GLM Procedure

1. Specify the conditional distribution of $y_i|\vec{x_i}^T$. For example $y_i \sim$ Bernoulli($\pi_i$).
2. Consider the parameter/s of the distribution in 1, there relations to $E[y_i|\vec{x_i}^T]$ and their domain.
3. Find a link function, g() that transforms $E[y_i|\vec{x_i}^T]$ into the range $(-\infty,\infty)$