Lecture on Penalized Regression

Li Xing

June 22, 2023

body{ font-family: Helvetica; font-size: 16pt; }

Issues Due to High Dimensionality

- The least squares estimates have relatively low bias and low variability especially when the relationship between Y and X is linear and the number of observations n is way bigger than the number of predictors p.
- But, when $n \approx p$, the least squares fit can have high variance and may result in over fitting and poor estimates on unseen observations.
- And, when $n < p$, the variability of the least squares fit increases dramatically, and the variance of these estimates in infinite

Example in Bioinformatics

- **Genome-Wide Association Studies**
    - There are millions of SNPs predictors $n << p$.
- **RNA Sequencing Data**
    - $\approx 20,000$ protein coding genes $n \approx p$.

Information Sparsity

- When we have a large number of variables X in the model there will generally be many that have little or no effect on Y.
- Leaving these variables in the model makes it harder to see the “big picture”, i.e., the effect of the “important variables”.
- The model would be easier to interpret by removing (i.e. setting the coefficients to zero) the unimportant variables.

Solutions

- Subset Selection
  - Identifying a subset of all p predictors X that we believe to be related to the response Y, and then fitting the model using this subset
  - E.g. best subset selection and stepwise selection

Solutions

- Dimension Reduction
    - Involves projecting all p predictors into an M-dimensional space where $M < p$, and then fitting linear regression model
    - E.g. Principle Components Regression

Solutions

- Shrinkage
  - Involves shrinking the estimates coefficients towards zero
  - This shrinkage reduces the variance
  - Some of the coefficients may shrink to exactly zero, and hence shrinkage methods can also perform variable selection
  - E.g. Ridge regression, Lasso, and Elastic Net Model

Penalized Regression

- Ordinary Least Squares (OLS) estimates by minimizing

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

- Ridge Regression

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

- Lasso Regression

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

Another Formulation

Lasso (6.8) and Ridge (6.9) can be formulated as below with a constraint associated with parameter s.

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le s \tag{6.8}$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \le s, \tag{6.9}$$

Geometry Illustration

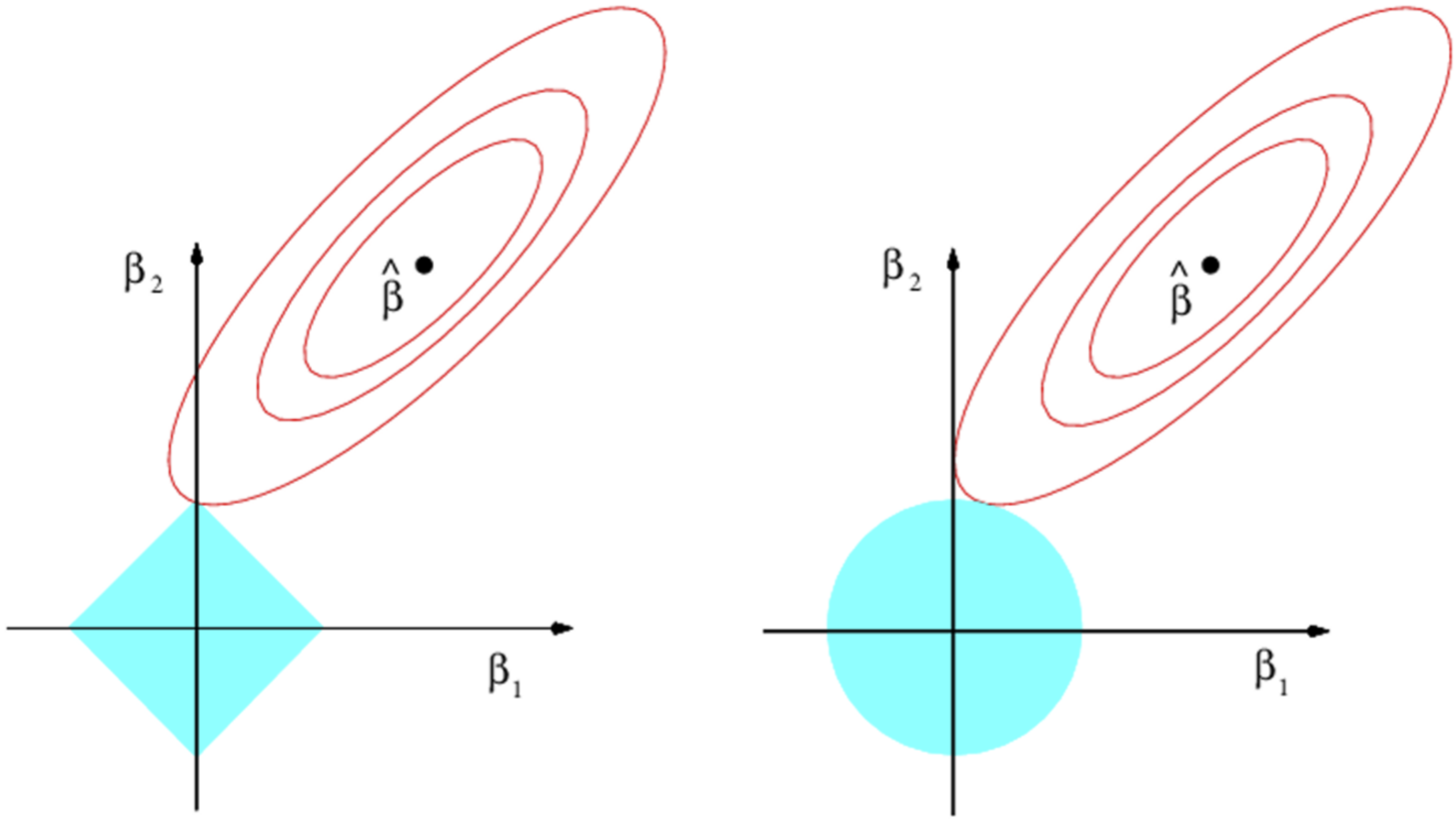

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \le s$ *and* $\beta_1^2 + \beta_2^2 \le s$*, while the red ellipses are the contours of the RSS.*

R Package

glmnet 4.1-7 | Get started | Reference | Articles | Changelog

An Introduction to `glmnet`

Trevor Hastie

Junyang Qian

Kenneth Tay

March 27, 2023

R Package

This vignette describes basic usage of glmnet in R. There are additional vignettes that should be useful:

- "Regularized Cox Regression" describes how to fit regularized Cox models for survival data with `glmnet`.
- "GLM `family` functions in `glmnet`" describes how to fit custom generalized linear models (GLMs) with the elastic net penalty via the `family` argument.
- "The Relaxed Lasso" describes how to fit relaxed lasso regression models using the `relax` argument.

`glmnet` solves the problem

$$\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)\|\beta\|_2^2/2 + \alpha\|\beta\|_1\right],$$