LECTURE 10
Sparse regression

We have seen previously that for the case that $p \gg n$ the following ridge regression model allows us stable inference

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|^2, \quad \lambda > 0.$$

Often we don't just want a good predictive model but we also want to know which variables are relevant to the prediction. The problem of simultaneously inferring a good regression model as well as selecting variable is called simultaneous regression and variable selection. In this lecture we will state some standard methods for simultaneous regression and variable selection.

We fist state the standard model

$$Y_i = (\beta^*)^T x_i + \varepsilon_i,$$

however we now assume that the regression coefficients are zero for the majority coordinates ($i = 1, .., p$). The subset of non-zero coordinates for the true model $\mathcal{A}_* = \{j : |\beta_*^{(j)}| \neq 0\}$ and the number of non-zero coefficients is denoted as $|\mathcal{A}_*|$. Our objective is given data $D = \{(x_i, y_i)_{i=1}^n\}$ to infer $\widehat{\beta}$ such that

(1) Selection consistency: The non-zero subset of $\widehat{\beta}$ is denoted as $\widehat{\mathcal{A}} = \{j : |\widehat{\beta}^{(j)}| \neq 0\}$. We would like the two subsets $\mathcal{A}_*$ and $\widehat{\mathcal{A}}$ to be close for any finite n and identical as $n \to \infty$.
(2) Estimation consistency: How well do the coefficients in the selected set converge:

$$\lim_{n \to \infty} \widehat{\beta}_{\mathcal{A}_*} = \beta^*_{\mathcal{A}_*}$$

The approach we will use for simultaneous regression and variable selection is the following minimization problem

$$\arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_q^q, \quad \lambda > 0,$$

where $\|\beta\|_q^q$ is a penalization by the q-norm. We've already seen the result of minimizing the 2-norm leads to ridge regression. We will now explore two other norms: the 1-norm and the 0-norm.

We start with the zero norm

$$\widehat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta)^2 + \lambda\|\beta\|_0^0, \quad \lambda > 0,$$

This is equivalent to the following minimization problem

$$\widehat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^p I(\beta_j \neq 0), \quad \lambda > 0,$$

which is suggesting minimizing the square error using the fewest variables possible with λ acting as the tradeoff between the number of variables and the error. The above minimization problem is NP-hard as it reduces to to exact cover by three sets. This means we can't practically implement the above optimization problem with any efficiency, even $p = 10$ requires a search over a massive space.

10.1. LASSO: Least Absolute Selection and Shrinkage Operator

The idea behind the lasso procedure is to minimize

$$\widehat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1, \quad \lambda > 0,$$

where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. For reasons we will discuss minimizing the above penalized loss function results in variable selection and regression, results in regression coefficients that are exactly zero. An argument has been made that minimizing the 1-norm regularized problem is a good approximation of the 0-norm minimization problem. We will explore both why this minimization problem approximates the 0-norm as well procedures to minimize the 1-norm.

10.1.1. The geometry of polytopes

Recall that there is an equivalence between

$$\arg\min_{\beta\in\mathbb{R}^p} \tfrac{1}{n}\textstyle\sum_{i=1}^n (y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1,$$
$$\arg\min_{\beta\in\mathbb{R}^p} \tfrac{1}{n}\textstyle\sum_{i=1}^n (y_i - x_i^T\beta)^2 \text{ subject to } \|\beta\|_1 \leq \tau.$$

We will contrast the following two minimization problems

$$\arg\min_{\beta\in\mathbb{R}^p} \tfrac{1}{n}\textstyle\sum_{i=1}^n (y_i - x_i^T\beta)^2 \text{ subject to } \|\beta\|_1 \leq \tau.$$
$$\arg\min_{\beta\in\mathbb{R}^p} \tfrac{1}{n}\textstyle\sum_{i=1}^n (y_i - x_i^T\beta)^2 \text{ subject to } \|\beta\|_2^2 \leq \tau.$$

Solutions to the upper problem is constrained to a 1-norm ball around the origin and the solutions to the lower problem is constrained to a 2-norm around the origin. Consider the true β vector to be β_*, the geometry of the square loss has ellipses as contours of equal loss. The minimizer is the smallest loss value that intersects the boundary of the p-norm ball.In the figure below we show this for two variables.

A minimizer with a sparse solution will touch/intersect the contours of the error ellipses on the axes that is sparse faces of the p-dimensional polytope. For example, when the constraint is the 2-norm ball around the origin it is very unlikely that the intersecting point will be concentrated on the axes. The geometry of the 1-norm ball especially in high dimensions intersects the ellipse at a few points. For example, the 0-norm is a star or spike that is on the axes so it will always be sparse.

Although we have considered the constrained optimization 1-norm problem the same results hold for lasso.

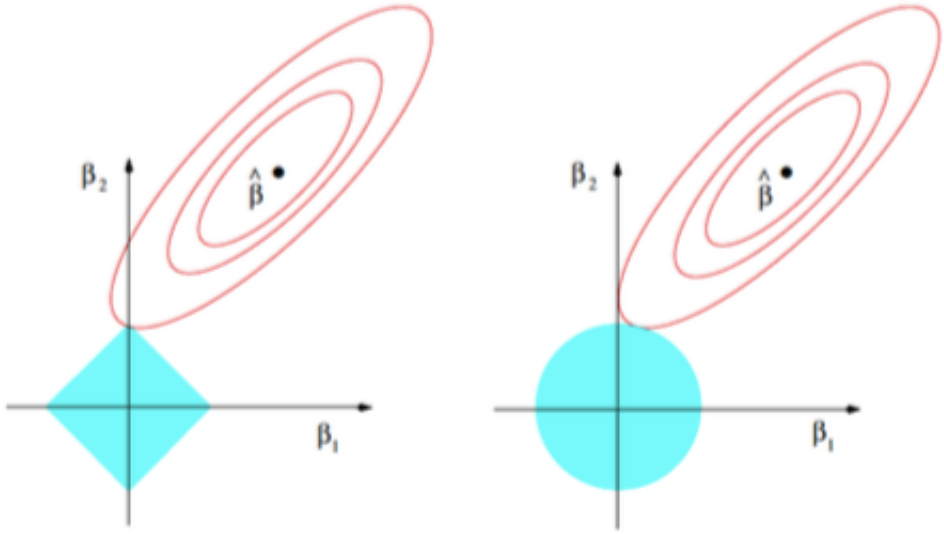

Figure 1. The 1-norm minimization for two variables is on the left and the 2-norm minimization is on the right.

10.1.2. The regularization path

Recall the optimization problem

$$\widehat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1, \quad \lambda > 0,$$

It is known that for $\lambda = 0$ the solution $\widehat{\beta}_{\text{LASSO}} = \widehat{\beta}_{\text{OLS}}$ and for $\lambda = \infty$ the solution is $\widehat{\beta}_{\text{LASSO}} = 0$. The regression coefficients β_λ for the lasso with regularization parameter λ is a p-dimensional vector with many of its values set to zero for larger values of λ. The idea of the regularization path is to examine how the β's change with λ the picture one should consider is λ as the x-axis and the β's on the y-axis. It is a mathematical fact that the graph of the β's will be piecewise continuous and approach zero at some point.

The idea behind the regularization path is to help select how many variables to keep in the model. In the ridge model it is hard to interpret a regularization parameter as coefficients are not sent to zero and the changes are slow. This is somewhat mitigated in the lasso model.

In the figure below we consider two regression analyses, one using ridge and the other with lasso, the dataset is a prostate cancer related problem. The response variable is PSA, a biomarker marker for prostate cancer, the covariates are several other biomarkers.

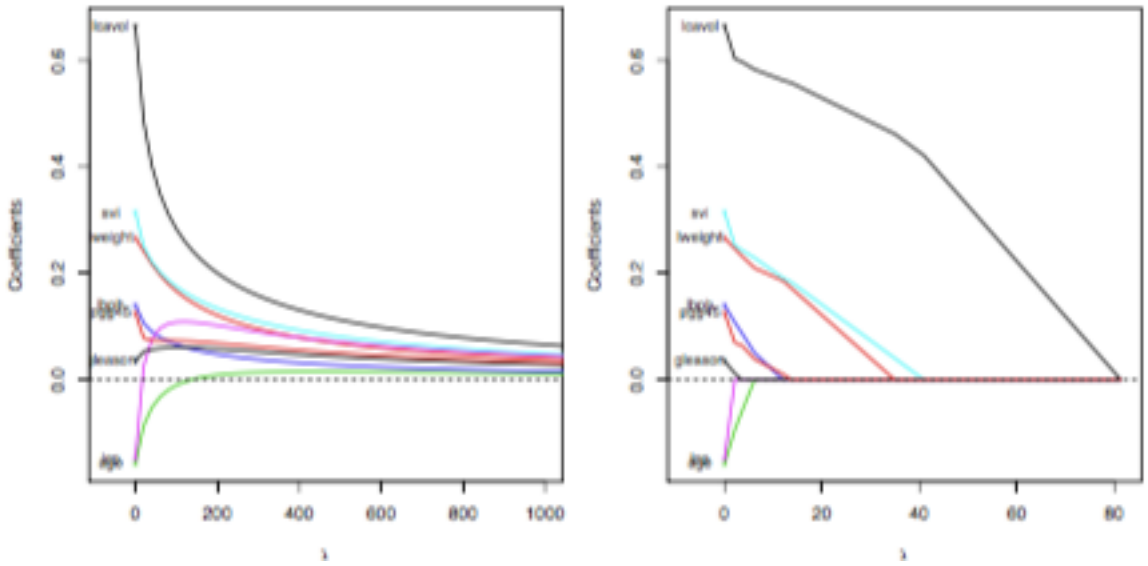

Figure 2. The left figure is the regularization path for ridge regression, the x-axis is the λ parameter and the y-axis is plotting the coefficients. The right figure is the same plot but for lasso. The response variable is PSA, a biomarker marker for prostate cancer, the covariates are a several biomarkers.