IEOR 265 – Lecture 6
Lasso Regression

1 Lasso Regression

The M-estimator which had the Bayesian interpretation of a linear model with Laplacian prior

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

has multiple names: Lasso regression and $L1$-penalized regression.

1.1 Soft Thresholding

The Lasso regression estimate has an important interpretation in the bias-variance context. For simplicity, consider the special case where $X'X = \mathbb{I}_p$. In this case, the objective of the Lasso regression decouples

$$\begin{aligned}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 &= Y'Y + \beta'X'X\beta - 2Y'X\beta + \lambda\|\beta\|_1 \\ &= Y'Y + \textstyle\sum_{j=1}^p \left(\beta_j^2 - 2Y'X_j\beta_j + \lambda|\beta_j|\right),\end{aligned}$$

where X_j is the j-th column of the matrix X. And because it decouples we can solve the optimization problem separately for each term in the summation.

Note that even though each term in the objective is not differentiable, we can break the problem into three cases. In the first case, $\beta_j > 0$ and so setting the derivative equal to zero gives

$$2\hat{\beta}_j - 2Y'X_j + \lambda = 0 \Rightarrow \hat{\beta}_j = Y'X_j - \lambda/2.$$

In the second case, $\beta_j < 0$ and so setting the derivative equal to zero gives

$$2\hat{\beta}_j - 2Y'X_j - \lambda = 0 \Rightarrow \hat{\beta}_j = Y'X_j + \lambda/2.$$

In the third case, $\beta_j = 0$.

For reference, we also compute the OLS solution in this special case. If we define $\hat{\beta}_j^0$ to be the OLS solution, then a similar calculation to the one shown above gives that $\hat{\beta}_j^0 = Y'X_j$. And so comparing to OLS solution to the Lasso regression solution, we have that

$$\hat{\beta}_j = \begin{cases} \hat{\beta}_j^0 + \lambda/2, & \text{if } \hat{\beta}_j = \hat{\beta}_j^0 + \lambda/2 < 0 \\ \hat{\beta}_j^0 - \lambda/2, & \text{if } \hat{\beta}_j = \hat{\beta}_j^0 - \lambda/2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

This can be interpreted as a soft thresholding phenomenon, and it is another approach to balancing the bias-variance tradeoff.

2 Dual of Penalized Regression

Consider the following M-estimator

$$\hat{\beta} = \arg\min_{\beta}\{\|Y - X\beta\|_2^2 : \phi(\beta) \leq t\},$$

where $\phi : \mathbb{R}^p \to \mathbb{R}$ is a penalty function with the properties that it is convex, continuous, $\phi(0) = 0$, and $\phi(u) > 0$ for $u \neq 0$. It turns out that there exists λ such that the minimizer to the above optimization is identical to the minimizer of the following optimization

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\phi(\beta).$$

To show this, consider the first optimization problem for $t > 0$. Slater's condition holds, and so the Langrange dual problem has zero optimality gap. This dual problem is given by

$$\begin{aligned} &\max_{\nu\geq 0}\min_{\beta} \|Y - X\beta\|_2^2 + \nu(\phi(\beta) - t) \\ \Rightarrow &\max_{\nu}\{\|Y - X\hat{\beta}^{\nu}\|_2^2 + \nu\phi(\hat{\beta}^{\nu}) - \nu t : \nu \geq 0\}. \end{aligned}$$

Let the optimizer be ν^* and define $\lambda = \nu^*$, then $\hat{\beta}^{\lambda}$ is identical to $\hat{\beta}$.

This result is useful because it has a graphical interpretation that provides additional insight. Visualizing the constrained form of the estimator provides intuition into why the $L2$-norm does not lead to sparsity, whereas the $L1$-norm does.

3 Variants of Lasso

There are numerous variants and extensions of Lasso regression. The key idea is that because Lasso is defined as an M-estimator, it can be combined with other ideas and variants of M-estimators. Some examples are given below:

3.1 Group Lasso

Recall the group sparsity model: Suppose we partition the coefficients into blocks $\beta' = \begin{bmatrix}\beta^{1'} & \dots & \beta^{m'}\end{bmatrix}'$, where the blocks are given by:

$$\begin{aligned} \beta^{1'} &= \begin{bmatrix}\beta_1 & \dots & \beta_k\end{bmatrix} \\ \beta^{2'} &= \begin{bmatrix}\beta_{k+1} & \dots & \beta_{2k}\end{bmatrix} \\ &\vdots \\ \beta^{m'} &= \begin{bmatrix}\beta_{(m-1)k+1} & \dots & \beta_{mk}\end{bmatrix}. \end{aligned}$$

Then the idea of group sparsity is that most blocks of coefficients are zero.

We can define the following M-estimator to achieve group sparsity in our resulting estimate:

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{m} \|\beta^j\|_2.$$

However, this estimator will not achieve sparsity within individual blocks β^j. As a result, we define the *sparse group lasso* as

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{m} \|\beta^j\|_2 + \mu\|\beta\|_1.$$

3.2 Collinearity and Sparsity

In some models, one might have both collinearity and sparsity. One approach to this situation is the *elastic net*, which is

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \mu\|\beta\|_1.$$

An alternative approach might be the Lasso Exterior Derivative Estimator (LEDE) estimator

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\Pi\beta\|_2^2 + \mu\|\beta\|_1,$$

where Π is a projection matrix that projects onto the $(p - d)$ smallest eigenvectors of the sample covariance matrix $\frac{1}{n}X'X$.

A further generalization of this idea is when there is manifold structure and sparsity: The Nonparametric Lasso Exterior Derivative Estimator (NLEDE) estimator is

$$\begin{bmatrix} \hat{\beta}_0[x_0] \\ \hat{\beta}[x_0] \end{bmatrix} = \arg\min_{\beta_0, \beta} \left\| W_h^{1/2} \left(Y - \begin{bmatrix} 1_n & X_0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \right) \right\|_2^2 + \lambda\|\Pi\beta\|_2^2 + \mu\|\beta\|_1,$$

where $X_0 = X - x_0' 1_n$, Π is a projection matrix that projects onto the $(p - d)$ smallest eigenvectors of the sample local covariance matrix $\frac{1}{nh^{d+2}} X_0' W_h X_0$, and

$$W_h = \operatorname{diag}\left(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h) \right).$$

4 High-Dimensional Convergence

One important feature of Lasso regression is consistency in the high-dimensional setting. Assume that X_j is column-normalized, meaning that

$$\frac{X_j}{\sqrt{n}} \leq 1, \forall j = 1, \ldots, p.$$

We have two results regarding sparse models.

1. If some technical conditions hold for the s-sparse model, then with probability at least $1 - c_1 \exp(-c_2 \log p)$ we have for the s-sparse model that

$$\|\hat{\beta} - \beta\|_2 \leq c_3 \sqrt{s} \sqrt{\frac{\log p}{n}},$$

   where c_1, c_2, c_3 are positive constants.

2. If some technical conditions hold for the approximately-s_q-sparse model (recall that $q \in [0, 1]$) and β belongs to a ball of radius s_q such that $\sqrt{s_q}(\frac{\log p}{n})^{1/2 - q/4} \leq 1$, then with probability at least $1 - c_1 \exp(-c_2 \log p)$ we have for the approximately-s_q-sparse model that

$$\|\hat{\beta} - \beta\|_2 \leq c_3 \sqrt{s_q} \left(\frac{\log p}{n}\right)^{1/2 - q/4},$$

   where c_1, c_2, c_3 are positive constants.

Compare this to the classical (fixed p) setting in which the convergence rate is $O_p(\sqrt{p/n})$.