High-Dimensional Sparse Models and Introduction to Lasso Regression:

We have a regression model,

$$Y = \beta' X + \epsilon = \sum_{j=1}^{P} \beta_j X_j + \epsilon$$

Here X are independent variables or regressors and β's are coefficients associated with them. So we have found $\beta' X$ is the best linear predictor of Y using X.

We have p regressors over here, generally, we call it p-dimensional. Mathematically we can write as,

$X = (X)_{J=1}^{P}$

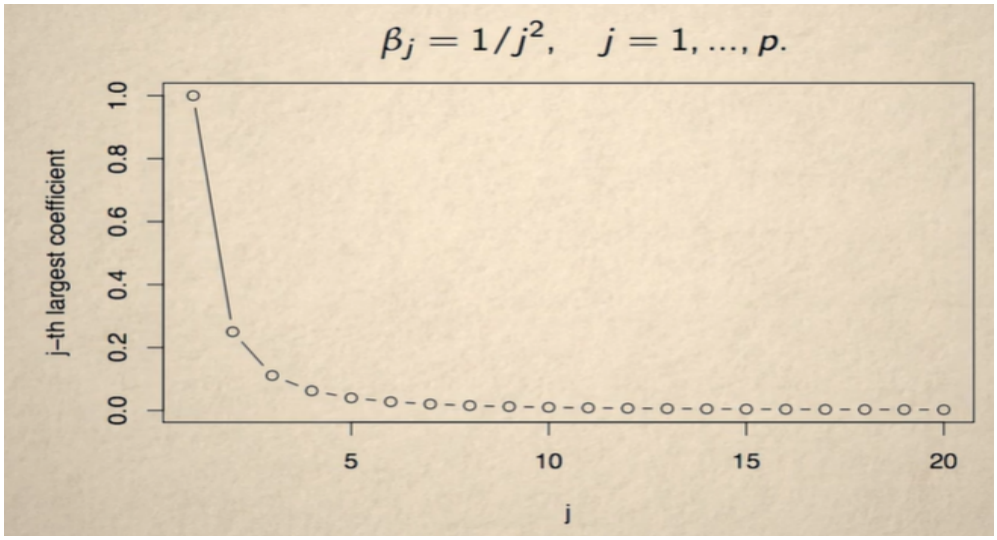
Now we have p which is much larger than n .i.e number of regressor are larger than the number of observations. For this kind of situation classical linear regression or ordinary least squares fails in these settings, because it overfits the data. Overfitting is when a statistical model fits exactly against its training data. When this happens, the algorithm, unfortunately, cannot perform accurately against unseen data, defeating its purpose.

We need to make some assumptions and modify the classical regression method to deal with this problem, Once such assumptions are Approximate Sparsity.

Approximate Sparsity :

This informally means that there is a small group of regressors that have relatively large coefficients, that can be used to approximate the best linear predictor $\beta' X$ quite well. The rest of the regressors have relatively small coefficients.

For example, An approximate sparsity is a linear model with the coefficients β_j's are given by 1 over j^2.



In this graph x-axis(j) represents the number of coefficients and the y-axis represents the value of the coefficients. we can see, the coefficients decrease quite fast, with only 3 or 4 regression coefficients that appear to be large other than these all are very small and closer to zero.

The key takeaway from here is while finding the best linear predictor when we have a large number of regressors(p) all the coefficients are not required as some of the regressors have large coefficients and the rest of the coefficients are very small. So if we include all coefficients while finding the BLP then the model loses its generalization power.

Now let's define the approximate sparsity formally,

Approximate sparsity: The sorted absolute values of the coefficients decay fast enough, namely the j-th largest coefficient (in absolute value) denoted by $|\beta|_{(j)}$ obeys:

$|\beta|_{(j)} \leq Aa^{-j}, a > 1/2$

Aa^{-j} can be also written as $\frac{A}{a^j}$.

Formally, the approximate sparsity means that the sorted absolute values of the coefficients decrease to zero fast enough, namely the j-th largest, in absolute value, the coefficient is at most of size j into the power of -" a" times a constant, where "a" is greater than 1/2. Here the constant measures the speed of decay.

Now let's build a best linear predictor when p is greater than n or p/n is not small,
For estimation purposes, we have a random sample of Yi and Xi, where i ranges from 1 to n. We have discussed that $\beta' X$ is the best linear predictor of Y using X when p/n is small.

Here we are finding the good linear predictor $\hat{\beta}\ X$which works well when p/n is not small. We can construct $\hat{\beta}\ X$ as the solution of the following penalized regression problem, one of such penalized regression algorithms are Lasso.

Lasso :
The word "LASSO" stands for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularisation of data models and feature selection. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

In lasso,

$$\min_{b \in \mathbb{R}^p} \quad \sum_i (Y_i - b' X_i)^2 + \lambda \cdot \sum_{j=1}^{P} |b_j|,$$

In lasso, we are minimizing the sample mean squared error that results from predicting Y_i with $\beta' X$ plus a penalty term, which penalizes the size of the coefficients $|\beta|_{(j)}$ by their absolute values. We control the degree of penalization by the penalty level λ.

The amount of the penalty can be fine-tuned using a constant called λ . Selecting a good value for λ is critical. When λ=0, the penalty term has no effect, and lasso regression will produce the classical least square coefficients. However, as λ increases to infinity, the impact of the shrinkage penalty grows, and the lasso regression coefficients will get close to zero.

The value of λcan be selected in two ways,

1. Theoretically justified method

$$\lambda = \sqrt{\mathrm{E}\epsilon^2}2\sqrt{2nlog(pn)}.$$

Where $E\epsilon^2 = \sum Y - \widehat{Y}$, n is the number of observations and p is the number of regressors.

2. Cross-validation - Another good way to pick penalty level is by cross-validation (which uses repeated splitting of data into training and testing samples to measure predictive performance).

This penalty level ensures that the Lasso predictor $\widehat{\beta}\ X$ does not overfit the data and delivers good predictive performance under approximate sparsity.

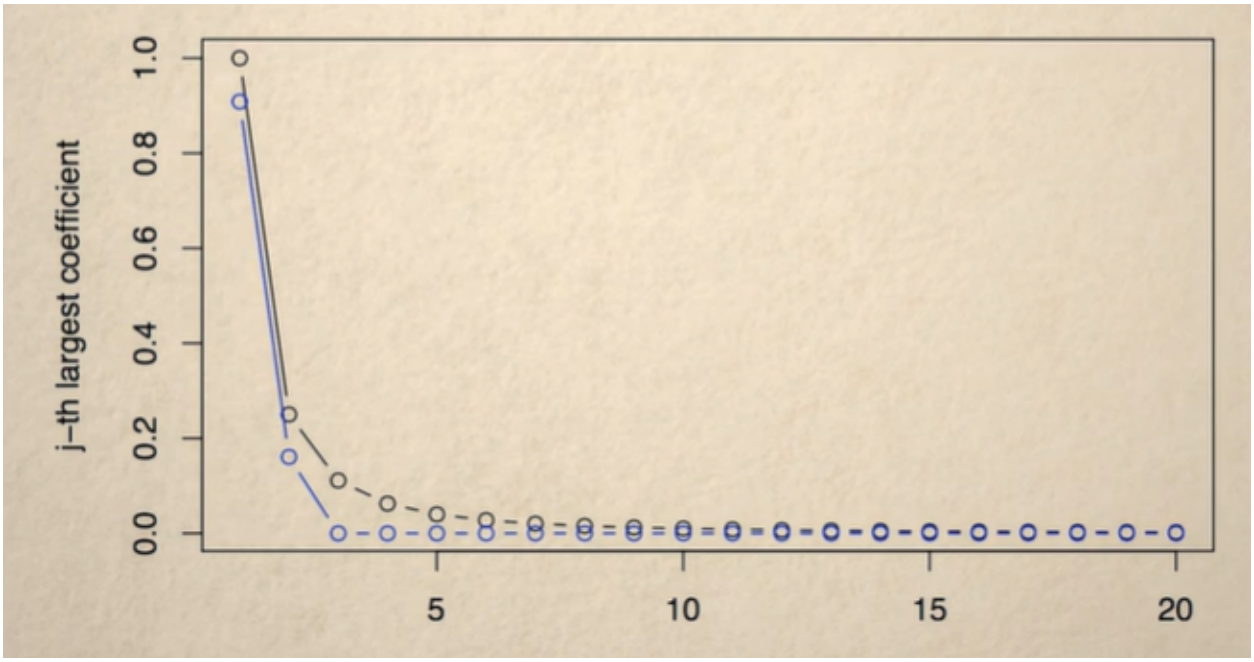
Intuitively, Lasso imposes the approximate sparsity on the coefficients $\widehat{\beta}$, just like in the assumption. It presses down all of the coefficients to zero as much as possible, without sacrificing too much fit, and it ends setting many of these coefficients to zero.

Let's understand through the example,

We have a simulation example of n=300 and p=1000,

$$Y = \beta' X + \epsilon, \quad X \sim N(0, I_p), \quad \epsilon \sim N(0, 1),$$

where X_j's and ϵ are standard normal variables and the true coefficients βj are equal to 1 over $1/j^2$.



The following figure shows that $\hat{\beta}$ in blue is indeed sparse and is close to the true coefficient vector β in black. Indeed, most of $\hat{\beta}_j$'s are set to zero, except for several coefficients that align quite well with the largest coefficients β_j of the true coefficient vector. This shows that Lasso can leverage approximate sparsity to deliver a good approximation to the true coefficients.

From the figure, we see that Lasso sets most of the regression coefficients to zero. It figures out approximately, though not perfectly, the right set of regressors.

The regressors selected by the lasso, using these regressors if we are going to fit the model by least squares, this method is called "least squares post-Lasso" or simply **post-Lasso**. Post-Lasso does not shrink large coefficients to zero as much as Lasso does and often improves over Lasso in terms of prediction.

Let's discuss the quality of prediction that Lasso and Post-Lasso methods provide. By definition, the best linear prediction rule (out-of-sample) is $\beta' X$, so the question is

Does $\hat{\beta}\ X$ provide a good approximation to $\beta' X$?

Here are trying to estimate j parameters β_1 through β_j, imposing the approximate sparsity via penalization. Under sparsity, only a few, say s, parameters will be "important", and we can interpret them as the effective dimension. Lasso

approximately figures out which parameters are important and estimates them. Intuitively, to estimate each of the "important" parameters well we need many observations per such parameter. This means that n/s must be large, or, equivalently, s/n must be small where n is the number of observations and s are effective dimensions.

Theorem

Under regularity conditions and under the approximate sparsity assumption, with probability approaching 1 as $n \to \infty$,

$$\sqrt{\mathrm{E}_X(\beta' X - \widehat{\beta}' X)^2} \leqslant const \cdot \sqrt{\mathrm{E}\epsilon^2} \sqrt{\frac{s \log(pn)}{n}},$$

where E_X *denotes expectation with respect to* X*, and the effective dimension is*

$$s = const \cdot n^{\frac{1}{2a}}.$$

Here the effective dimension s is equal to a constant times n into the power of 1 over 2a, where a is the rate of decrease of coefficients in the approximate sparsity assumption.

The key takeaway from this theorem is that, if n is large and the effective dimension s is much smaller than n/log(pn), for nearly all observations of the sample, the Lasso predictor gets really close to the best linear predictor.

Therefore, under approximate sparsity, Lasso/Post-Lasso will approximate the best linear predictor well. This means that Lasso and Post-Lasso won't overfit the data, and we can use the sample and adjusted R2 and MSE to assess out-of-sample or unseen predictive performance. We can verify the out-of-sample predictive performance by using test/validation samples.