Prediction accuracy and interpretability of a statistical model is enhanced by the LASSO regression technique by performing the variable selection and regularization. Robert Tibshirani (1996) first introduced it. Initially it was used for least squares estimates. One of the main problems with least squire method is that, we have non-zero coefficient all the time whatever the true beta coefficients are. Lasso keeps good features of both subset selection and ridge regression and it does this by shrinking some of the coefficients and setting the rest to zero.

In least squire method we minimize the residual sum of squires. Prediction accuracy and interpretation are the two main reasons for which the data analyst is usually not satisfied with the least squire estimates. They usually have low bias and large variance. Sometimes we can improve the prediction accuracy by shrinking/setting some coefficients to zero. A little bias is sacrificed when we decrease the variance of the predicted values. In this way overall prediction accuracy can be increased.

Subset selection and ridge regression are two standard techniques better than least squire method. But these methods have some drawbacks. Subset selection, however, can be extremely variable because of the discreteness of the process (either retain or drop out the regressor) but it provides interpretable models. Prediction accuracy is reduced since a small changes in the data can result in very different models being selected. On the other hand ridge regression is more stable which is a continuous process and shrinks the coefficients. Although it does not give an interpretable model since it does not set any coefficients to zero.

Lasso regularization was initially used for least squire estimation but it is applicable to a wide variety of statistical models, namely proportional hazards models, generalized linear models, generalized estimating equations, and M-estimators. At first we run the

Let us consider the following linear regression model:

y= βTx+ ϵ

Here ϵ~*Nn* (0, σ2I) and X~ *Np*(0, V). n is the number of observations and p is the number of predictors. In this example p=8, σ2=1, V= Ip and β = (3,1.5,0,0,2,0,0,0)T. We used 10 repetitions with n= 50 and 100 respectively.

In lasso we minimize the following objective function

arg min subject to

β

Where t 0 is the tuning parameter which controls the quantity of shrinkage which is applied to the estimates.

The above equation is equivalent to

arg min

β

Where, 0 is a tuning parameter. Cross validation (CV) is used to find the optimal and then we get the beta coefficient.

For adaptive lasso we consider the following objective function

arg min

β

Where is known weights vector.

We applied 5-fold cross validation (CV), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to decide the value of the tuning parameter. Then we compare the performance of Lasso and Adaptive Lasso based on these criterion.

**Cross Validation (CV):**

To find the optimal value of use CV. In this procedure we split the datasets into two groups, namely training set and test set. Usually only one part is used to model the data set. The other part of the data is predicted by using the model selected in first part.

Here is an example of the cross validation procedure with K (K=5 or 10) folds:

Split the data into 5 equal subsets. Consider the first subset as the test set and fit the models to the rest of the data. Evaluate the predictions of the training set by using the test set. Then repeat the procedure for second, third, fourth and fifth subsets respectively (means repeat the procedure until each subsets is considered as training set and test set). Finally, calculate the average the performance of all the test sets. The process is a cross-validated estimate of generalization error for each model. The procedure is finally end up by picking the model which has the smallest error.

We have used following calculation to estimate the coefficient using Adaptive Lasso:

, where

= , where and

Where optimal is calculated by cross validation method.

Akaike information criterion (AIC) is defined as follows

AIC = -2ln(L) + 2P

Where, L =(, , and P= number of non-zero coefficients

Bayesian information criterion (BIC) is defined as

BIC = -2ln(L) + P ln(n)

Where L and P are defined in AIC.

The results with n=50, repetition = 10

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | | Avg. No. of 0 Coefficients | | Median MSE | Mean MSE | Frequency of getting correct model |
| Correct | Incorrect |
| Lasso | CV | 1.4 | 0 | 0.01850394 | 0.02008248 | 0 |
| AIC | 2.9 | 0 | 0.01921771 | 0.02018252 | 3 |
| BIC | 3.9 | 0 | 0.01875845 | 0.02149817 | 4 |
| ALasso | CV | 3.4 | 0 | 0.01446094 | 0.03002126 | 2 |
| AIC | 4.3 | 0 | 0.009899418 | 0.01502099 | 6 |
| BIC | 4.6 | 0 | 0.009899418 | 0.01380999 | 7 |
| Enet | CV | 4.4 | 0 | 0.02475081 | 0.02399247 | 7 |
| AIC | 4.2 | 0 | 0.0148878 | 0.01690168 | 6 |
| BIC | 4.4 | 0 | 0.01599222 | 0.01712257 | 7 |

The results with n=100, repetition = 100

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Method | | Avg. No. of 0 Coefficients | | Median MSE | Mean MSE | Frequency of getting correct model |
| Correct | Incorrect |
| Lasso | CV | 2.64 | 0 | 0.02008572 | 0.02090119 | 13 |
| AIC | 2.82 | 0 | 0.01973033 | 0.01990746 | 12 |
| BIC | 3.77 | 0 | 0.01899754 | 0.02082585 | 30 |
| ALasso | CV | 3.79 | 0.04 | 0.01578133 | 0.0828695 | 39 |
| AIC | 4.08 | 0 | 0.01446375 | 0.01539818 | 46 |
| BIC | 4.77 | 0 | 0.00920333 | 0.01177625 | 81 |
| Enet | CV | 4.69 | 0 | 0.02812096 | 0.02962475 | 78 |
| AIC | 4 | 0 | 0.01361826 | 0.01527801 | 51 |
| BIC | 4.71 | 0 | 0.009905619 | 0.01299382 | 79 |

From the above table we see that Adaptive Lasso gives us the best performance in terms of getting/ identifying correct models compared to Lasso and Elastic Net.

**Reference**

1. Tibshirani, R. (1996) Regression Shrinkage and Selection via Lasso, *J. R. Statist. Soc* B*(Methodology),* 58, No. 1, 267-288.

2. Breiman, Leo. (1995) Better Subset Regression Using the Nonnegative Garrote. *Technometrics* 37 (4). 373–84. doi:10.2307/1269730.