Lasso-SIM: An empirical study of the advantages of combining Least Absolute Shrinkage and Subset Operator with Single Index Model as compared to traditional logistics regression using AIC.

Mr. Murphy Choy
School of Information System
Singapore Management University

Summary

LASSO has been the preferred model selection procedure for a long time in regression models. Single Index Models (SIM) or Projection Pursuit Regression with a single term is a useful model where they are powerful classifiers in situation where classical logistics regression cannot handle. In this particular article, we will try to compare the use of LASSO in conjunction with SIM and other possible combinations of model such as AIC-SIM or AIC logistics regression in terms of predictive power.

Introduction

Logistic regression is the most common model used in analytics to classify object or response based on some factors that are assumed to be related to the object or response. However, logistics regression is relatively inflexible being constrained by specify link function that is defined by the user. This inflexibility results in the inability of logistics regression to predict object or responses where the positive and negative results are not so clearly defined. To overcome this lack of flexibility, single index model is proposed in this case to replace the logistic regression. However, SIM, traditionally uses almost all covariates to achieve a high predictive power. To reduce the reliance on too many covariates, LASSO method will be used prior the construction of SIM to identify the important covariates and subsequently building a SIM using these covariates.

LASSO

The Lease Absolute Shrinkage and Selection Operator (LASSO) is a model selection and variables estimation technique proposed by Tibshirani (1996). The LASSO method applies an $L_1$-type of penalty on the regression coefficients which results in the shrinkage of many of the coefficients near to zero to be set to zero. This results in many of coefficients of the less important regressors to be set to zero which leads to the important regressors to be retained in the model. The LASSO method is a more stable and accurate method of model selection as compared to traditional variable selection methods (Tibshirani, 1996).

Let us assume a Gaussian linear regression model,

$$y = X\beta + \varepsilon$$  

where $y = (y_1, ..., y_n)^T$ are the responses, $\beta = (\beta_1, ..., \beta_d)$ are the regression coefficients, $X = (x_1, ..., x_d)$ are the covariate matrix, and errors in the form $\varepsilon = (\varepsilon_1, ..., \varepsilon_n) \sim N(0, \sigma^2 I_n)$. 


Without loss of generality, the covariates are standardized to a mean 0 and unit length, and the response has mean 0.

The LASSO estimate is the solution to

\[
\min_{\beta} (y - X \beta)^T (y - X \beta), \quad \text{s.t. } \sum_{j=1}^{d} |B_j| \leq t.
\]

In this case, \( t \geq 0 \) is the tuning parameter. Assume that \( \hat{\beta}^0 \) be the ordinary least square (OLS) estimate and 

\[ t_0 = \sum |\hat{\beta}^0_j| \]

Thus values of \( t < t_0 \) will shrink those solutions toward 0. Tibshirani (1996) demonstrated that the solution to the LASSO estimate for the orthonormal matrix design above has the form of

\[
\hat{\beta}_j = \text{sign}(\beta^0_j) \left( |\beta^0_j| - \gamma \right)^+, \quad j = 1, \ldots, d, \text{ when } X^T X = I
\]

However, to select the appropriate model, we still need to select the appropriate selection criteria in order to select the correct model.

**Single Index Model**

Assume that \( Y \) is the response variable and \( X = (x_1, x_2, x_3, \ldots, x_p)^T \) are covariates. The single index model is written as

\[
Y = g(x^T \theta^0) + \varepsilon
\]

Where \( E(\varepsilon | X) = 0 \). \( g \) is an unknown link function and \( \theta^0 \) is an unknown unit vector with nonzero components positive for identification purposes. Most of the research into SIM have been mainly concerned with the estimation of the index parameters and the non parametric link function with particular focus on the root-n consistency of the index parameters (Powell et al., 1989; Härdle & Stoker, 1989; Härdle et al., 1993; Xia et al., 2002). The most popular methods of estimation are the average derivative estimation method proposed by Härdle & Stoker (1989) and Härdle et al. (1993).

In most cases, the research assumes that the covariates are important and contain information needed to predict the response. However, in situations where the number of covariates are extremely large or in cases of high dimensional environments (Naik & Tsai, 2000), the accuracy of the parameters estimation and prediction are compromised (Altham, 1984). To ensure good predictive power of the model, it is necessary to reduce the number of unnecessary covariates. Splice inverse regression was used with success in Naik & Tsai (2001), however, the predictors are constrained by those that continuous and elliptically symmetric. These strong constrains results in the use of cross validation methods in selecting variables in SIM. However, in Kong & Xia (2007), it is shown that cross validation method display a different behavior in SIM and thus they proposed the use of a new method called separated cross validation method. However, the cross validation methods are computer intensive. To save up on computation time, LASSO is proposed to be used as the main
variable selection method in conjunction with the use of information criterion to enhance its effectiveness.

**Model selection criteria for LASSO**

To select the appropriate model, various selection criterions were adopted to select the correct model. Akaike’s information criterion and Bayesian information criterion are among the commonly used criteria to select models. In our situation, we will try Akaike’s information criterion, Bayesian information criterion and their variants in selecting the most appropriate model. Mallow's Cp is computed by the LARS method in R programming. Thus for the definition of Mallow's Cp, it is recommended that the reader be referred to the writer of the method in R.

The Akaike’s information criterion is proposed by Akaike(1969) as a mean of selecting the appropriate model size. It is defined as,

\[
AIC = n \log\left(\frac{RSS}{n}\right) + 2k
\]  

(4)

A variant of AIC, known as AICc (Hurvich & Tsai, 1989 ;Burnham and Anderson, 2004) is also used as a way to select the appropriate sample size. It is defined as,

\[
AICc = n \log\left(\frac{RSS}{n}\right) + 2k + \frac{2k(k+1)}{n-k-1}
\]  

(5)

Schwarz (1978) proposed the Bayesian information criterion as an alternative to AIC. It is defined as,

\[
BIC = n \log\left(\frac{RSS}{n}\right) + k \log(n)
\]  

(6)

A variant of BIC by Hannan and Quinn (1979) which is called BICc in this article is used as an alternative to BIC. It is defined as,

\[
BICc = n \log\left(\frac{RSS}{n}\right) + k \log(\log(n))
\]  

(7)

The ideal model is the one selected with the lowest value for each of the criterion. However, in the examples that follows, we will use the various criterion and compared the terms selected and the how each criterion performs. It is noted that in literature (Hurvich & Tsai, 1989) that AIC and its variants tend to perform poorly in small datasets.
Empirical Examples

Example 1: Caravan Insurance Policy Holder Identification

Origin: This dataset is owned and supplied by the Dutch data mining company Sentient Machine Research, and is based on real world business data. This dataset has been used in the CoIL Challenge 2000 datamining competition. The objective is to identify the caravan insurance policy holders.

In the original competition, the main priority is given to identification of policy holders and the best model will be the one which indicates the highest number of positive result from the policy holders. In the competition, the winner is Charles Elkan from University of California using naïve Bayes classifier which identify 121 policy holders.

Using all the variables, both the logistics regression and single index model achieves a 50% identification rate which yield 119 positive results. Using AIC, the single index model yield a 50% positive rate using 33 variables while the logistics regression yield a 40% positive rate using the same variables. When the various information criterion are used in conjunction with LASSO, they yielded different variables being selected. Both AIC and AICc selected 30 variables and produced a positive rate of 60% in logistics regression and 40% in SIM. BIC, on the other hand, uses 15 variables and has a positive rate of 0.4 using logistics regression and 0.6 using SIM. BICc uses 29 variables with a positive rate of 0.6 for logistics regression and 0.33 for SIM. It can be seen that LASSO using BIC in conjunction with SIM produces very good result which is comparable to the top winner with minimum resources and easily outmatch the runner ups who got 115 positive (using a combination of neural network like classifier) and 112 positive (using a combination of 20 naïve bayes classifiers).

Therefore, the LASSO-SIM model performs better than logistics regression and at the same time reduces the amount of resources needed to perform such a prediction.

<table>
<thead>
<tr>
<th>Model</th>
<th>LASSO Used</th>
<th>AIC/AICc/ BIC/BICc/ None</th>
<th>Number of Terms</th>
<th>% Positive Result (Positive)</th>
<th>% Negative Result (Positive)</th>
<th>%Positive Result (Negative)</th>
<th>%Negative Result (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>N</td>
<td>None</td>
<td>86</td>
<td>0.500</td>
<td>0.500</td>
<td>94.094%</td>
<td>5.906%</td>
</tr>
<tr>
<td>SIM</td>
<td>N</td>
<td>None</td>
<td>86</td>
<td>0.500</td>
<td>0.500</td>
<td>94.094%</td>
<td>5.906%</td>
</tr>
<tr>
<td>GLM</td>
<td>N</td>
<td>AIC</td>
<td>33</td>
<td>0.400</td>
<td>0.600</td>
<td>94.093%</td>
<td>5.907%</td>
</tr>
<tr>
<td>SIM</td>
<td>N</td>
<td>AIC</td>
<td>33</td>
<td>0.500</td>
<td>0.500</td>
<td>94.116%</td>
<td>5.884%</td>
</tr>
<tr>
<td>GLM</td>
<td>Y</td>
<td>AIC/AICc</td>
<td>30</td>
<td>0.600</td>
<td>0.400</td>
<td>94.118%</td>
<td>5.882%</td>
</tr>
<tr>
<td>SIM</td>
<td>Y</td>
<td>AIC/AICc</td>
<td>30</td>
<td>0.400</td>
<td>0.600</td>
<td>94.093%</td>
<td>5.907%</td>
</tr>
<tr>
<td>GLM</td>
<td>Y</td>
<td>BIC</td>
<td>15</td>
<td>0.400</td>
<td>0.600</td>
<td>94.093%</td>
<td>5.907%</td>
</tr>
<tr>
<td>SIM</td>
<td>Y</td>
<td>BIC</td>
<td>15</td>
<td>0.500</td>
<td>0.500</td>
<td>94.094%</td>
<td>5.906%</td>
</tr>
<tr>
<td>GLM</td>
<td>Y</td>
<td>BICc</td>
<td>29</td>
<td>0.600</td>
<td>0.400</td>
<td>94.117%</td>
<td>5.882%</td>
</tr>
<tr>
<td>SIM</td>
<td>Y</td>
<td>BICc</td>
<td>29</td>
<td>0.333</td>
<td>0.667</td>
<td>94.070%</td>
<td>5.929%</td>
</tr>
</tbody>
</table>
Example 2: Rear Wheel Drive Vehicle Identification

Origin: Specifications are given for 428 new vehicles for the 2004 year. The variables recorded include price, measurements relating to the size of the vehicle, and fuel efficiency. Kiplinger's Personal Finance, December 2003, vol. 57, no. 12, pp. 104-123.

In this dataset, there are a good mixed of variables both of categorical and continuous nature. The original purpose of this dataset is to estimate car prices. However, in this example, we will attempt to identify whether a vehicle is rear wheel drive or not based on some variables.

There are 428 rows of data and 19 columns of variables. After cleaning the dataset, we are left with 387 rows of data and 11 columns of variables. We than split the dataset into the training set and validation set. Below is the table showing the results of the models. In this case, SIM produces the best positive result and negative result using only 5 variables.

Table 2: Model Results

<table>
<thead>
<tr>
<th>Model</th>
<th>LASSO Used</th>
<th>AIC/AICc/BIC/BICc</th>
<th>Number of Terms</th>
<th>% Positive Result (Positive)</th>
<th>% Negative Result (Positive)</th>
<th>% Positive Result (Negative)</th>
<th>% Negative Result (Negative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>N</td>
<td>None</td>
<td>11</td>
<td>0.2567</td>
<td>0.7432</td>
<td>0.6017</td>
<td>0.3982</td>
</tr>
<tr>
<td>SIM</td>
<td>N</td>
<td>None</td>
<td>11</td>
<td>0.3454</td>
<td>0.6545</td>
<td>0.6623</td>
<td>0.3376</td>
</tr>
<tr>
<td>GLM</td>
<td>N</td>
<td>AIC</td>
<td>4</td>
<td>0.4333</td>
<td>0.5667</td>
<td>0.7007</td>
<td>0.2992</td>
</tr>
<tr>
<td>SIM</td>
<td>N</td>
<td>AIC</td>
<td>4</td>
<td>0.3827</td>
<td>0.6173</td>
<td>0.6886</td>
<td>0.3113</td>
</tr>
<tr>
<td>GLM</td>
<td>Y</td>
<td>AIC/AICc/BIC/BICc</td>
<td>5</td>
<td>0.4736</td>
<td>0.5263</td>
<td>0.6912</td>
<td>0.3087</td>
</tr>
<tr>
<td>SIM</td>
<td>Y</td>
<td>AIC/AICc/BIC/BICc</td>
<td>5</td>
<td>0.4756</td>
<td>0.5244</td>
<td>0.7619</td>
<td>0.2381</td>
</tr>
</tbody>
</table>

Example 3: Counterfeit Banknote Identification

Origin: Extracted from Kong and Xia (2007). The dataset uses several measurements to identify the genuine from the counterfeit notes.

In this small dataset, there are 200 rows of data and 6 columns of variables. It is smaller than the previous 2 datasets and keeping in mind that AIC and its variants performs poorly in small sample. We will be using Cp as the selection criterion. Using the same condition as specified in Kong and Xia (2007), the average number of misspecification is 0.5039 which is slightly superior to that using the separated cross validation shown below:

<table>
<thead>
<tr>
<th>Method</th>
<th>Ave. No of Misspecification</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Variables</td>
<td>0.5787</td>
</tr>
<tr>
<td>Cross Validation</td>
<td>0.6223</td>
</tr>
<tr>
<td>Separated Cross Validation</td>
<td>0.5100</td>
</tr>
<tr>
<td>Principal Component Analysis</td>
<td>0.5411</td>
</tr>
<tr>
<td>LASSIM</td>
<td>0.5039</td>
</tr>
</tbody>
</table>
Conclusion

LASSO-SIM provides a viable alternative to classical logistics regression if the correct information criterion is used. LASSO in conjunction with information criterion also provides a competitive alternative to separated cross validation method in variable selection for SIM.

References: