STA315 Final Project

Bhatt, Devansh; Student ID - 1001442209
Tan, Charles; Student ID - 1003518142

20 April 2020

TOPIC -
Sure Independence Screening for Ultrahigh Dimensional Feature Space

We declare that this assignment is solely our own work, and is in accordance with the University of Toronto Code of Behaviour on Academic Matters.

This submission has been prepared using LaTeX.

Abstract

Ultra-high dimensional data has become increasingly more common in the face of many areas of scientific discoveries, most notably in genomics and proteomics. When the dimension of the feature space is significantly larger than the sample size, the technique of sure screening, that in which fundamentally encapsulates the very concept of variable selection, successfully reduces the dimension to a smaller scale. In this paper, we carry out Tibshirani's LASSO as well as execute two variations of Fan and Lv's Sure Independence Screening (SIS) method which include Smoothly Clipped Absolute Deviation (SIS-SCAD) and Dantzig Selection (SIS-DS) to assess their accuracy and speed in accomplishing high dimensional reduction. Simulations show that SIS-DS is most effective in sample size reduction whilst SIS-SCAD is ultimately the optimal choice with respect to error rate and runtime analysis.

Introduction

The main concerns in "Sure Independence Screening for Ultrahigh Dimensional Feature Space" written by Jianqing Fan and Jinchi Lv revolve around high dimensional data and the unforeseeable problems that it brings. High dimensionality occurs when the number of predictors is conspicuously larger than the sample size. This is a notable issue because it is more than likely that not all of these predictors have an influence on the response which is why variable selection is so important. Variable selection is responsible for retaining a subset of the predictors that actually matter to the response whilst discarding the remaining ineffectual excess variables. This leads to many benefits such as enhanced model interpretability and significantly improved computation time, both products of a simplified model representation.

The rapidly increasing prevalence of high-dimensional data in the medical field especially in genomics and proteomics is what intrinsically motivated Fan and Lv to tackle this issue of high dimensionality. As a result, they proposed their very own variation of variable selection known as Sure Independence Screening (SIS). SIS is performed in order to preserve all the important predictors during the variable selection process with a probability tending to 1 (Fan and Lv, 2008). As we will go into more depth in the following section, SIS is founded on the idea of correlation learning as it "ranks the importance of features according to their marginal correlation with the response variable and filters out those that have weak marginal correlations with the response variable" (Fan and Lv, 2008, p. 853). After carrying out the SIS procedure, it is extremely beneficial to pair SIS in conjunction with a lower-dimensional variable selection technique such as LASSO, SCAD or the Dantzig Selector (DS) which can effectively reduce the parameter space even further.

Despite the soundness of the proposed SIS method, it does unfortunately come with three flaws:

Problem #1: Some irrelevant predictors that are strongly correlated with important predictors may assume a higher priority in being selected by SIS over the important predictors that have a weaker relationship to the response.

Problem #2: An important predictor that is marginally uncorrelated with other variables but jointly correlated with the response cannot be chosen by SIS and hence will not be included in the final model.

Problem #3: The issue of collinearity increases the difficulty of the variable selection problem.

Fan and Lv propose an iterative variant of their SIS method known as Iterative Sure Independence Screening (ISIS). ISIS is their solution to rectify the three issues listed above as it refines the methodological power of SIS.

This paper is organized as follows. In the next section we discuss the mathematical foundation of SIS as well as their newly proposed ISIS method and how it quantitatively improves the former method. Subsequently, we will delve into our replicated simulations that originate from the given literature which include SIS-SCAD and SIS-DS along with their corresponding results as shown in R. Finally, some concluding remarks will be given in the closing section.

Fan and Lv's SIS and ISIS

The general SIS model as implemented by Fan and Lv is carried out as follows:

1. We start off with an n-vector for our response Y and a n × p matrix for our variable X.

2. We generate models by regressing Y over each X_p individually.

3. By utilizing the Pearson Correlation Coefficient r, we can find the correlation coefficient for each of the p models.

4. We get the absolute value of each of the p correlation coefficients then rank them from largest to smallest.

5. Given that d is strictly less than n, we get the top d correlation coefficients to serve as our "surviving" chosen variables. This reduces the dimensions of our x variable from n × p to n × d.

6. After this process is completed, we can continue to reduce the feature space further by using one of the potential low-dimensional methods such as the Smoothly Clipped Absolute Deviation (SCAD), the Dantzig Selector or LASSO.

However, as we expressed earlier, the SIS is not a perfect variable selection technique due to three notable drawbacks explored earlier. Hence, the authors introduced an enhanced variation known as Iterative Sure Independence Screening (ISIS), which serves as a plausible improvement to their original Sure Independence Screening technique.

ISIS is carried out as follows:

1. Using one of the SIS based model selection methods such as SIS-SCAD or SIS-LASSO, we can select a subset of k_1 variables $\{X_{i_1}, \ldots, X_{i_{k1}}\}$ which we can denote as A_1.

2. After regressing the response Y over X_{i_1} to $X_{i_{k_1}}$, this results in an n-vector of residuals.

3. Let the n-vector of residuals serve as the new response then we repeat the previous steps on the remaining p - k_1 variables. This would result in a new subset $\{x_{j_1}, \ldots\ k_{j_{k_2}}\}$ which we can denote as A_2.

   Note #1: By fitting the residuals on X_1 to X_p, this can significantly lower the selection priority of the unimportant variables that are highly correlated with the response as the residuals are uncorrelated with the selected variables in the subset A_1. This ultimately solves problem #1.

   Note #2: The predictors which were initially missed due to having a low correlation with the response are now able to survive the screening process. Hence, this also solves problem #2.

   Note #3: After the variables in A_1 enter the model, the marginally weakly correlated variables to the response should now be correlated with the residuals. Which finally solves problem #3.

4. We continue to repeat this process until we get ℓ disjoint subsets A_1, ...,A_ℓ where the union $\bigcup_{i=1}^{\ell} A_i$ has a size of d. Similar to the process of SIS, the size of d should be strictly less than the sample size n.

5. We can slowly approach the true sparse model by extending the variable selection process above by once again using a lower dimensional method such as the SCAD, Dantzig selector or LASSO.

Numerical Studies

For the simulation we decided to replicate the settings from which the authors proposed in the paper. Fan and Lv (2008) devised to select a linear model ($Y_i = \beta_0 + \beta_1 X_i + ... + \beta_p X_i + \epsilon$) with Normally distributed predictors to create a $n \times p$ matrix X. We created two models with $(n, p) = (200, 1000)$ and $(n, p) = (800, 20000)$. We followed the authors in selecting the true size of the model, s, to be 8 and 18 (Fan and Lv, 2008). The non-zero components of the p-vectors β were chosen by using $(-1)^u(a+|z|)$, where $a = 4log(n)/n^{1/2}$ and $5log(n)/n^{1/2}$, u is drawn from a Bernoulli distribution with parameter 0.4 and z was drawn from the standard Normal distribution (Fan and Lv, 2008). Finally, we used the p-vector β to create the n-vector of response variables Y.

In order to verify the results the authors attained, we applied 4 of the same methods they employed: Dantzig Selector, LASSO using the LARS algorithm, SIS-SCAD and SIS-DS. For the SIS-SCAD and SIS-DS methods we allowed SIS to first reduce the parameter space to $n/log(n)$, and then let SCAD and DS reduce the parameter space further (Fan and Lv, 2008).

To test the performance of each method, we collected the median model sizes, including the intercept, from 200 runs and median estimation errors. We decided to use the square root of the square-loss (RMSE) function to calculate the estimation errors; contrary to what the authors used in the article. We define RMSE as $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2}$. Additionally, we decided to also collect data on the computation time for 200 runs of each test. We opted to not partition the dataset into a training and test set because there was no evidence of the authors' doing so.

p	Dantzig Selector	Lasso	SIS-SCAD	SIS-DS
1000	6	225	9	6
20000	–	–	19	12

Table 1: Results of simulation. Medians of selected model sizes.

p	Dantzig Selector	Lasso	SIS-SCAD	SIS-DS
1000	5.87	6.96×10^{-15}	2.31×10^{-5}	5.87
20000	–	–	3.55×10^{-5}	8.99

Table 2: Results of simulation. Medians of estimation errors.

p	Dantzig Selector	Lasso	SIS-SCAD	SIS-DS
1000	2 mins	2 min	18 secs	25 secs
20000	–	–	7.8 mins	2.9 mins

Table 3: Results of simulation. Total run-time.

As shown in all three tables above, we could not include the results corresponding to the columns of the Dantzig Selector and LASSO at p=20000 due to the extraneous computational costs at predicting the p-vector β using those methods. From Table 1 we can see that the only method capable of accurately reducing the parameter space to s is SIS-SCAD at both levels of sparsity. This method is also exceptional at reducing the estimation errors to a very small scale. The methods which rely on the Dantzig Selector are reducing the parameter space more than s which leads them to getting very large estimation errors as observed in Table 2. As we can observe from the computation times in Table 3, SIS-SCAD does very well for $p = 1000$ but it takes the most amount of computation time for $p = 20000$. On the contrary, for SIS-DS at the $p = 20000$ level, it does a good job at reducing the computation time but the large estimation errors and inaccuracy at reducing the parameter space fail to make a strong case for the method. Though the computational time is very high for SIS-SCAD at estimating an extremely large parameter space, it has proven to be the optimal solution for such a job. We would like to add that computation times will vary for each simulation run and different computing setups. It is interesting to see the difference in times for each simulation run on one comuter. Our results are closely in line with what the authors had gathered in their study. They had also concluded that SIS-SCAD is the overall best method. However, they were dealing with different selected model sizes and estimation errors for each method. This leads us to believe that we did not exactly recreate the testing environment for which the authors had operated in throughout their investigation.

Conclusion

After exploring the solutions presented in “Sure Independence Screening for Ultrahigh Dimensional Feature Space”, there is a general consensus regarding the numerous problems brought forth by ultra high-dimensional datasets. High dimensionality is becoming the norm of datasets in the fields of gene study, physics, finance and social media. As companies start using vast nets to capture data, they are inevitably going to face the problem of large feature spaces. In these circumstances, they can rely upon the suggestions as proposed by Fan and Lv along with the statistical software packages that had been developed to solve their problems. Fan and Lv’s solution of Sure Independence Screening can reduce the feature space to a size smaller than the sample size which can enable analysts to use some existing, well-defined low-dimensional techniques to reduce the feature space to a more manageable, interpretable, and computationally efficient level. Finally, the method of Sure Independence Screening ultimately provides us with another option to prediction and classification in a world that is rapidly being consumed by increasingly complex data.

Reference

- Fan, J., Lv, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society. Series B (Statistical Methodology), 70(5)*, 849-911. Retrieved February 12, 2020, from `www.jstor.org/stable/20203862`