BSTA 670 Statistical Computing: Project 2

John Pluta

Fall 2014

1 Introduction

An important step in producing stable and interpretable statistical models is variable selection. First introduced in [1], LASSO is a penalized regression method that is especially useful in cases with many variables, as it leads to sparse solutions where only a few variables are given non-zero values, and can accommodate correlated data by shrinking coefficient estimates. Furthermore, the LARS algorithm [2] provides an extremely computationally efficient method for solving the LASSO problem and defines the entire coefficient path. For these reasons, LASSO has become a popular tool for variable selection and model building. While it is undoubtedly a useful tool in many applications, it is important to understand its limitations as well. Work by [3] begins this discussion by investigating how the performance of LASSO relates to the underlying structure of the data it is measuring. In this work, we seek to verify the results of one of the experiments from this paper, and then provide an extension.

2 Irrepresentable Condition

First, we define some of the important terminology of the experiment. [3] present an 'Irrepresentable Condition' that quantifies certain properties of the data in question. This is defined as follows:

The design matrix, X, is arranged so that the first q columns are of the data associated with $\beta(1)$, and the remaining $p-q$ columns follow. Then X can be partitioned into $X(1)$ and $X(2)$, where $X(1)$ includes only columns 1 through q, and $X(2)$ contains columns $q+1$ through p.

Since X is normally distributed with 0 mean, we can state that:

$$Cov(X) = E[X^T X] \rightarrow \tfrac{1}{n} X^T X = C$$

C can be partitioned into the sub matrices:

$$C_{11} = \tfrac{1}{n} X(1)^T X(1)$$
$$C_{21} = \tfrac{1}{n} X(2)^T X(1)$$

where C_{11} is the covariance of the relevant variables, and C_{21} is the covariance between the relevant and irrelevant variables. Then, the Irrepresentable Condition can be stated mathematically as:

$$||(C_{11}^{-1} C_{12} sign(\beta(1)||_\infty < 1 - \eta,\ \eta \in [0, 1)$$

Thus if the relevant variables are more closely related to each other than they are the irrelevant variables, we will get a value of less than 1, and the Irrepresentable Condition will be met. We can also express this condition in terms of η:

$$\eta = 1 - ||(C_{11}^{-1}C_{12}sign(\beta(1)||_\infty$$

In this form, the value of η represents the degree to which the Irrepresentable Condition is met or violated, with larger positive values implying it is more strongly met, and smaller negative values implying it is more strongly violated.

3 Simulation

In this work, we aim to replicate the results of section 3.2 of [3], where the authors illustrate quantitatively the relationship between the magnitude of the Irrepresentable Condition and the effect on model selection.

For the simulation, we have $n = 100$, $q = 5$, $p = 32$, and $\sigma^2 = 0.1$. The true β parameters are: $\beta(1) = [7, 4, 2, 1, 1]^T$ and $\beta(2) = 0$, where $\beta(2)$ contains all of the remaining coefficients. This is the model that we seek to recover, and is used in generating the dependent variable Y.

Next, a matrix $S \sim Wishart(p, p)$ is defined. From this, we generate 100 design matrices (denoted X), where $X \sim N(0, S)$. This creates random datasets, with random correlation induced between the columns of X. This step was run several times until a set of X matrices was found with a suitable distribution of η- we want the range of η to cover at least -.2 to .2, to mimic the results in [3].

For each X, we sample 100 values of $\epsilon \sim N(0, \sigma^2)$, and compute the outcome vector in the standard manner, as $Y = XB + \epsilon$. With, Y, X, and β defined, the LARS algorithm is run and the entire solution path is computed. Then, the solution path is examined for a solution that is sign consistent with the true beta values. The sign function is defined as:

$$sign(\beta) = \begin{cases} 1 & \beta > 0 \\ 0 & \beta = 0 \\ -1 & \beta < 0 \end{cases}$$

Thus we are looking for a LASSO solution that chooses the correct variables, and weights these variables in the correct direction. The result, a binary success or failure, is recorded, and this process is repeated 1000 times for each design matrix. The final outcome is the total proportion of successfully matched models compared to the total number of models run; in essence, this is the probability that LASSO can recover the correct model, in relation to the value of η. Results from [3] are presented in Figure 1.

From Figure 1, we observe a strong relationship between η and the probability of LASSO recovering the true model. The threshold for meeting the Irrepresentable Condition is at $\eta = 0$, and there is a sharp increase in recovery rate at this point. Below the threshold (when the condition is violated), the probability of recovery quickly drops off to 0, and conversely, quickly rises to one when the condition is strongly met. These results imply a strong relationship between the structure of the design matrix and the utility of LASSO as a variable selection tool.

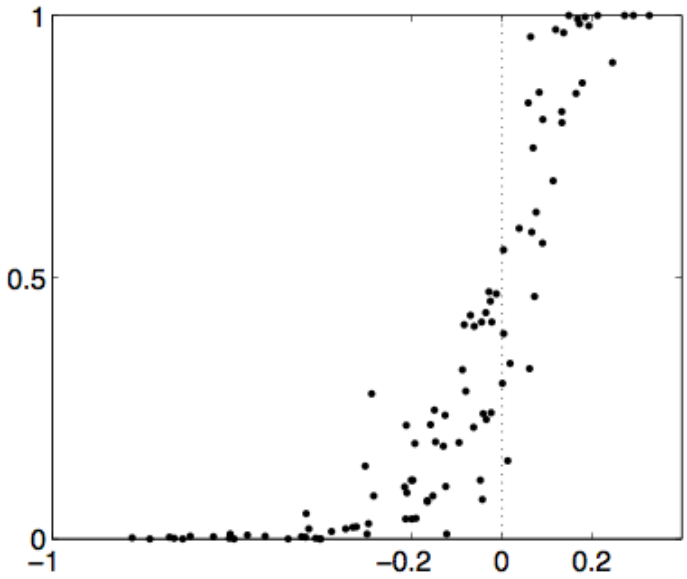

Figure 1: Probability of correctly recovering model as a function of η. Figure taken from (Zhao and Yu, 2006).

4 Reproducibility

In this section, we attempt to replicate the results of the presented experiment. Figure 2 shows the results of the reproduced simulation. Each iteration (e.g. creation of a design matrix and 1000 simulations) took on average 0.01 seconds. The resulting plot is quite similar to that presented in the original paper, confirming that the results of [3] are indeed reproducible.

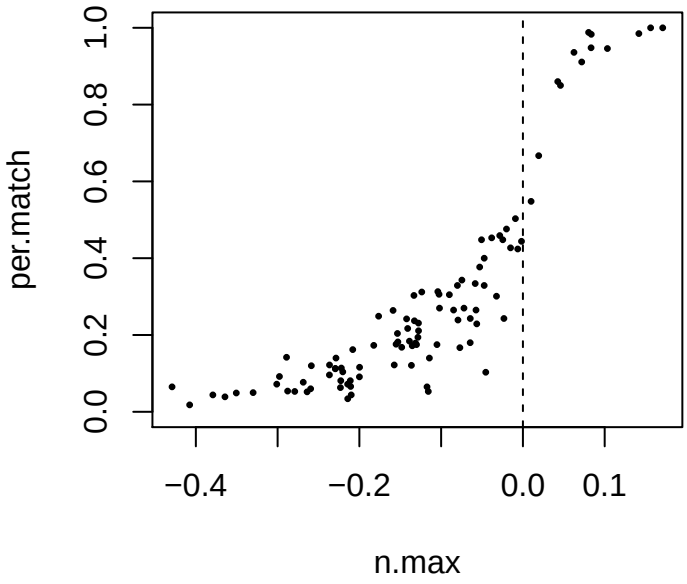

Figure 2: Probably of correctly recovering model as a function of η: replicated results.

5 Extension

The results of the simulation show that the probability of LASSO recovering the true model is closely related to the value of η, with the probability quickly approaching 0 if the Irrepresentable Condition is violated. In practice, this is difficult to test a-priori, since we have no way of partitioning X into $X(1)$ and $X(2)$. While [3] do provide some guidance for how to test for the Irrepresentable Condition without being able to distinguish the relevant and irrelevant variables, the results seem to suggest that LASSO is not suited for many kinds of problems.

However, the presented simulation provides no information on the magnitude of the error. If LASSO completely fails at picking the correct variables for negative values of η, then indeed it must be used with caution. We extend the results of the previous experiment by quantifying the degree of error as a function of η. The simulation is unchanged, but we now search the entire coefficient solution path and find the best fitting model, even if it is not exact. That is, for each iteration, we find the minimum error and store this value. For each design matrix, we report the average minimum error, and plot this over η to examine the relationship between the two. Results are summarized in Figure 3.

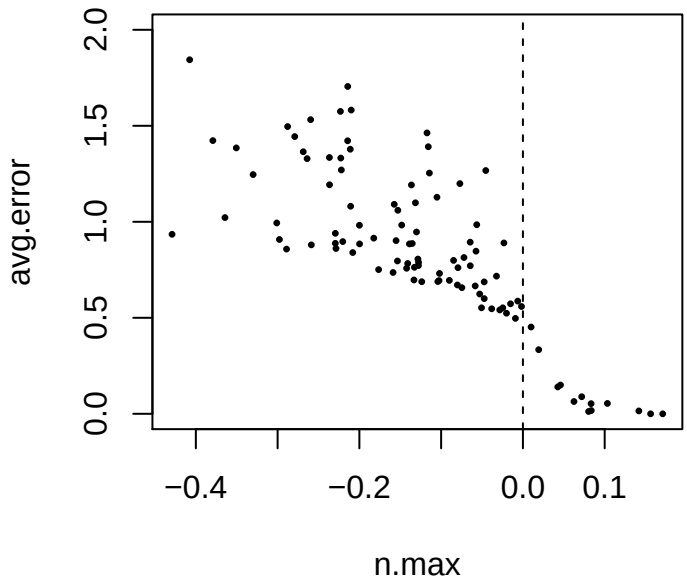

Figure 3: Average error as a function of η

6 Discussion

The relationship between η and average error is unsurprising; it is largest for negative values of η, when the Irrepresentable Condition is violated the most, and drops sharply at 0. It is practically zero for positive values of η, when the condition is satisfied. The important finding from this experiment is that the magnitude of the error never exceeds 2, and is typically around 1. This implies that in all but the worst cases, LASSO is picking at most one variable incorrectly; this corresponds to choosing the first 5 variables correctly and adding one extraneous variable incorrectly.

In the listed example, with $p = 32$ variables, selecting only one incorrectly is still a promising result and would likely lead to reasonable coefficient estimates. Furthermore, LASSO is often used as a preliminary step for variable selection, and further regression techniques are then used on the selected variables. It is quite possible that post-hoc testing, such as the significance test for LASSO results proposed in [4], would eliminate the erroneously chosen variable. Regardless, the results of the extension make the results of the initial simulation more promising; even when the underlying data is not well-conditioned, LASSO is still an appreciable tool for variable selection.

References

[1] Tibshirani, R. 1996. *Regression shrinkage and selection via the lasso.* J. R. Statist Soc. 58(1):267-288.

[2] Efron, B., Hastie, T., Johnstone, Iain, Tibshirani, R. 2004. *Least angel regression.* The Annals of Statistics. 32(2):407-499.

[3] Zhao, P., Bin, Y. 2006. *On model selection consistency of Lasso.* Journal of Machine Learning Research 7:2541-2563.

[4] Lockhart, R., Taylor, J., Tibshirani, R., Tibshirani, R. 2014. *A significance test for the LASSO.* The Annals of Statistics. 42(2):413-468.