P8160 Project 1: Simulation Study for Variable Selection Methods

Ngoc Duong, Crystal Li, Yuchen Qi

Introduction

Variables selection methods are always trying to balance between model fitness and model complexity in the high-dimensional setting. And in application of traditional variables selection methods, they often struggle with identifying weak signals. Some signals are weak but they are still of importance to the true model.

Objectives

In this project, we use simulations to compare two automated methods of variable selection, namely stepwise forward method and LASSO regression, in their ability to correctly identify relevant signals and estimating how missing weak signals impact coefficients of strong signals. Specifically, we aim to assess:
(1) How well each method performs in identifying weak and strong predictors (by calculating the percentages of strong and weak predictors being captured by each model), and
(2) How missing "weak" predictors impact the estimations of strong predictors (by calculating the bias and MSE between "true" strong coefficients and their estimates).

Statistical methods to be studied

Methods of interest in this report are the step-wise forward method and automated LASSO regression which are two popular methods for the variable selection.

Step-wise forward method: Starting with the empty model, and iteratively adds the variables that best improves the model fit. In this report, it is done by sequentially adding predictors with the largest reduction in AIC, where

$$AIC = n\ln(\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/n) + 2p,$$

where $\widehat{y}_i$ is the fitted values from a model, and p is the dimension of the model (i.e.,number of predictors plus 1).

Automated LASSO regression It estimates the model parameters by optimizing a penalized loss function:

$$\min_{\beta} \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\|\sum_{k=1}^{p}|\beta_k|$$

where λ is a tunning parameter. Here cross-validation (CV) is the chosen selection criteria for LASSO.

Scenarios to be investigated

First we give the definitions of "strong", "weak-but-correlated" and "weak-andindependent" signals.

Definition of strong signals —

$$S_1 = \{j : |\beta_j| > c\sqrt{log(p)/n},\ \text{some } c > 0, 1 \le j \le p\}$$

Definition of weak-but-correlated signals —

$$S_2 = \{j : 0 < |\beta_j| \le c\sqrt{log(p)/n},\ \text{some } c > 0, \text{corr}(X_j, X_j') \ne 0, \text{for some } j' \in S_1, 1 \le j \le p\}$$

Definition of weak-and-independent signals —

$$S_3 = \{j : 0 < |\beta_j| \leq c\sqrt{log(p)/n},\ \text{some } c > 0, \text{corr}(X_j, X_j') = 0, \text{for all } j' \in S_1, 1 \leq j \leq p\}$$

To narrow the scope of our simulations, some variables are fixed.
(1) We set the proportions of strong signals, weak and independent signals, and weak but correlated signals to be 10%, 20%, 20% respectively, then we have 50% null predictors.
(2) The coefficients of strong signals follow Uniform(5, 10) which is sufficiently larger than the bound, and the coefficients of strong signals follow Uniform(1/2bound, bound), where the bound is threshold by definition.
(3) The threshold multiplier c is set to be 1.

Then, we vary the amount of total predictors from 10 to 100, with step to be 10. We also choose the correlation value to be 0.3, 0.5, 0.7. For each scenario, we generate 100 datasets. And in each dataset, the sample size is 200.

Methods for generating data

Generating the predictor data matrix X

From the proportions of each type of signals and the number of total predictors, we get how many signals for each type. Then we generate a covariance matrix with the correlations set in this scenario following the definitions of each signal type. Whether the matrix is positive definite is also checked before passing it to the R function `mvrnorm`, which produces random numbers from a multivariate normal distribution.

Generating the response Y

We generate the response Y as a linear combination of four types of signals and an error term. The distribution of Y is

$$Y \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2)$$

where the variance is 1.

Performance Measures

Task 1: identify strong and weak predictors

We wanted to investigate both variable selection methods' ability to correctly identify strong and weak (both WAI and WBC) predictors and whether they do so consistently. Therefore, we measure their performances by calculating the percentages of captured strong, WBC and WAI predictors using these two methods as the number of parameters and correlation value changes.

Task 2; how missing "weak" predictors impacts the estimations of strong predictor

In order to see the effect of missing weak predictors on the coefficient estimates of strong predictors, before fitting the models, we deleted a certain number of weak signals (from 1 to 20) from the original data. We then calculated the MSE and bias between "true" strong coefficients and their estimates, where

- bias

$$\frac{1}{p_{strong}} \sum_{j=1}^{p_{strong}} (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$$

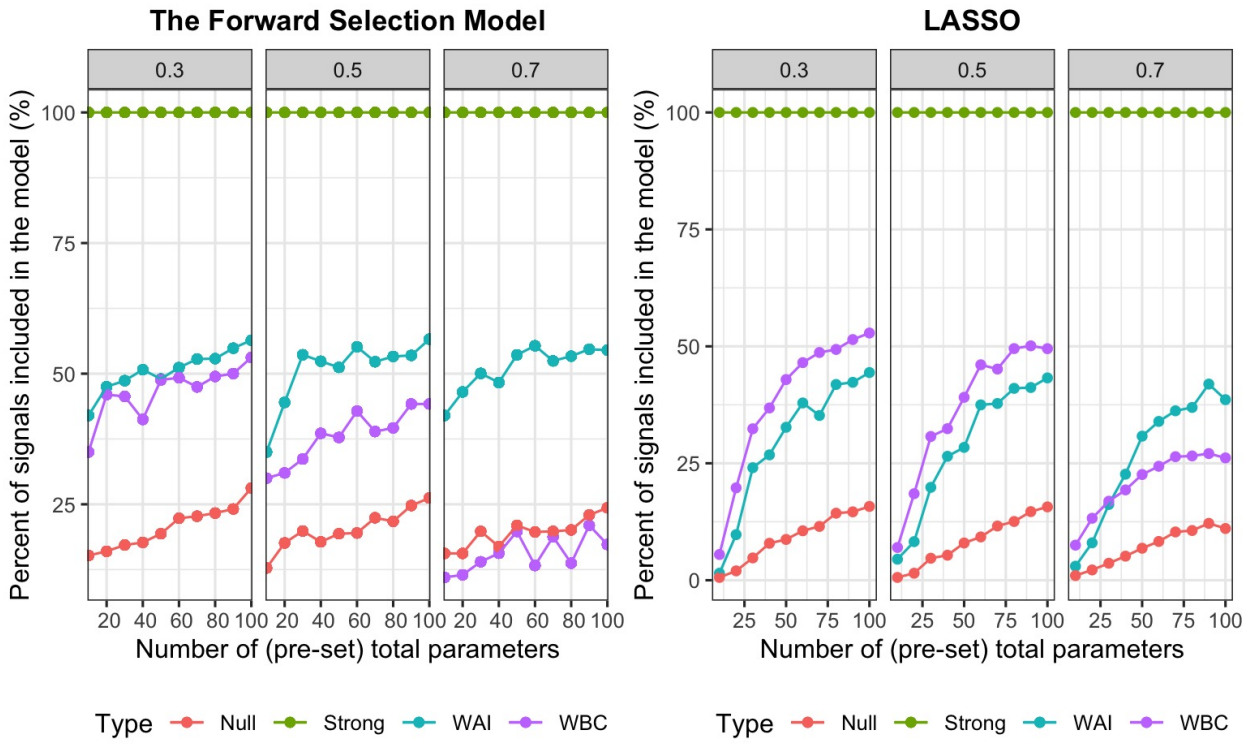

Figure 1: Percent of captured Signals

- MSE

$$\frac{1}{p_{strong}} \sum_{j=1}^{p_{strong}} (\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^2$$

Simulation results

Figures

Code

The gitbub link is https://github.com/qi-yuchen/Advancd_Computing_Project_1.