Project 1-Topic 3 Report

Hongjie Liu, Jiajun Tao, Shaohan Chen

02/27/2023

Objectives

When dealing with high-dimensional data

Statistical Methods Studied

Stepwise Forward Selection

Step-wise forward selection

LASSO

LASSO regression

Scenarios Investigated

Several parameters

Methods for Generating Data

Data is generated using a 4-step method:

Performance Measures

Ability of Model to Identify Weak and Strong Predictors

One method of

Effect of Missing Weak Predictors

On the other hand ## Results

According to Table 2, forward stepwise always includes more predictors than the number of true predictors in original data while LASSO is more selective.

Scenarios 1-4: Varying total number of weak predictors

Type I error and power

While both methods do are decent at selecting true predictors for the model, LASSO is much better at excluding null predictors than forward selection (Table 3).

Coefficient estimation

Although the median SE of the coefficient estimates for forward stepwise is lower than LASSO, its MSE is higher when there are few true weak predictors in the data. ### Scenarios 4-8: Varying degree of correlation

Type I error and power

As the correlation of the WBC predictors increases, the type I error for forward stepwise and LASSO remain fairly constant around 0.35 and 0.05, respectively. Meanwhile, the power of forward selection increases, while the power of LASSO decreases slightly. (Figure 6)

Coefficient estimation

Although forward stepwise selection includes more non-null predictors on average (Table 3), it has increasing and higher MSE compared with LASSO, showing a growing trend when correlation becomes larger. The MSE of LASSO is controlled under 0.1, whereas forward selection has an inflated MSE (over 0.2) when the correlation gets very high.

Conclusion

General

Forward stepwise selection always includes more predictors than LASSO, and the model size given by each method increases as the number of total true predictors in the original data increases.

Varying total number of weak predictors

The SE variance for forward stepwise is much larger than LASSO when the number of weak predictors is small

Varying degree of correlation

With increased correlation

Discussion

As the degree of correlation increases

Figures and Tables