A simulation study to compare variable selection methods

Group6: Kee-Young Shin, Weijia Xiong, Xinru Wang, Yimeng Shang

2/10/2020

Objective

When working with high-dimensional data, it is common practice to utilize variable selection methods to create an opitmal model that balances model fitness and model complexity. Many selection methods use p-values as the main criteria, which can pose many issues. Various factors such as sample size, correlation and variance can inflate or deflate p-values. Therefore, many traditional selection methods often struggle in the presence of weak predictors.

For our simulation study, we compared two variable selection methods: Stepwise Forward Selection and the LASSO regression. In the process, we wanted to (Task 1) study the proportion of signals identified by the two methods and (Task 2) determine the impact of removing weak predictors on parameter estimates.

Statistical Methods

Stepwise Forward Selection

The Stepwise Forward Selection is a type of selection method in which variables are iteratively added starting from an empty model if they improve the model under a certain criteria. The selection method in this project used the AIC as its criteria:

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

The AIC is a goodness of fit measure that favors smaller residual error, while also penalizing for increased predictors, and variables were added iteratively if it reduced the AIC. This method performs variable selection by comparing models.

LASSO Regression

The LASSO regression, on the other hand, is a regression analysis method that works to optimize the following loss function:

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

The λ represents the shrinkage parameter with a higher lambda value imposing a bigger constraint on the coefficients. A λ of 0 would lead to the cost function of a normal linear regression. This regularization technique can lead to zero valued coefficients, effectively neglecting the prediction in the evaluation of output. In other words, LASSO's variable selection is done within the model, unlike the Stepwise Forward Method, through its parameter estimation.

Data

Signals Definition

To effectively analyze the study objectives, data that contained a combination of strong, weak-and-independent (WAI), weak-and-correlated (WBC), and null predictors had to be generated. The signals were created using the following criterias:

Strong signals

$$S_{strong} = j : |\beta_j| > c\sqrt{\frac{log\,(p)}{n}}, \text{ for some c} > 0, 1 \leq j \leq p$$

Weak-but-correlated signals

$$S_{WBC} = j : |\beta_j| \leq c\sqrt{\frac{log\,(p)}{n}}, \text{ for some c} > 0, corr(X_j, X_j) \neq 0, \text{for some j' in } S_{strong}, 1 \leq j \leq p$$

Weak-and-independent signals

$$S_{WBC} = j : |\beta_j| \leq c\sqrt{\frac{log\,(p)}{n}}, \text{ for some c} > 0, corr(X_j, X_j) = 0, \text{for some j' in } S_{strong}, 1 \leq j \leq p$$

Null signals

$$S_{null} = j : \beta_j = 0, 1 \leq j \leq p$$

Methods for generating data

The general idea was to generate a dataset which contains a combination of "strong", "weak-but-correlated" and "weak-and-independent" predictors. We chose to build a dataset with 50 predictors and a sample size of 100. To do this we first created a 50x50 positive-definite variance-covariance matrix with WBC variables being correlated to the first strong predictor with $corr(X_j, X_j) = 0.3$. Then a random matrix X was created from a multivariate normal distribution with mean 0 and sigma equal to the variance-covariance matrix. Next, the matrix of true coefficient values was created with the strong signals set to 5 and the weak predictors (WAI and WBC) set to the threshold value defined by the changing c value. Finally, we generated the linear response Y values:

$$Y = 1 + X\beta + \epsilon$$

Scenarios to be investigated

To fully compare the variable selection methods we utilized various scenarios created by ranging the c values in the threshold and the ratio of strong and weak predictors. For Task 1, we varied both the c value and the ratio. We tested c values from 1 to 3, and three ratios: (1) 10 strong signals, 10 WAI, and 10 WBC, (2) 10 strong signals, 5 WAI, and 5 WBC, and (3) 5 strong signals, 10 WAI, and 10 WBC.

For Task 2- the influence of removing weak predictors on parameter estimates- we varied the ratio of strong and weak signals using the same three ratios, but kept the c value the same at 1.

For the LASSO Regression, the lambda tuning parameter was determined through cross-validation and was set to 1.

Performance Measures

Task 1: Comparison of identification ratio

In the comparison of variable selection methods, we used the identification rate as the performance measure. The identification rate was defined as the proportion of types of signals detected after the selection process.

Task 2: Parameter estimations

The study's Task 2 required us to determine the effect of removing weak predictors on parameter estimates. To do this we used the Mean Squared Error as the performance measure:

$$MSE = \frac{1}{n}(\bar{\beta}_{sim}\beta)^2$$

As weak parameters were removed one by one, we ran simulations repeatedly on the new number of predictors to determine the MSE of the estimates. Since we created the dataset, we know the true coefficient values and are able to calculate the MSE.

Simulation Results

Task 1: Comparison of identification rate

For Task 1, we ran 1000 simulations under the different scenarios. As the number of weak predictors increased, the identification rate for WBC predictors decreased, while the rates for strong predictors and WAI predictors remained around the same (Figure 1). Since the WBC variables are correlated to a strong predictor, it is not unexpected that we see these variables being less identified as the number of weak predictors increase.

The identification rate of WBC variables for LASSO seemed to decrease much faster than that of Forward Selection. This may be due to the fact that LASSO is regularization technique and therefore may be more sensitive to collinearity. In other words, LASSO regression when facing correlated variables will often keep one and throw out the other.

Strong predictors were much easier to identify than weaker signals (Figure 2) for both variable selection methods. Additionally, WAI signals were more easily identified than WBC signals. As can be seen by the bottom violin plots in Figure 2, the WAI identification rate was higher the WBC identification rate for both methods.

When increasing the threshold c values, the identification rates increased for both WAI and WBC signals, though the rate for WAI signals increased much faster (Figure 3 & Figure 4). And again, the Stepwise Forward Method seemed to perform slightly better in identifying WBC signals compared to LASSO. This increase in identification rate is most likely due to the fact that increasing the c in the threshold value is essentially increasing the signal strengths, making them easier to detect.

Task 2: Parameter estimations

To see how removing weak predictors impacted the parameter estimates, we ran 500 simulations under the different scenarios and focused on the impact on the strong predictors. The weak signals were removed sequentially starting from the weak-and-independent predictors. Figures 5, 6 and 7 depict the MSE for the first five strong predictors, since the MSE patterns were similar for the remaining strong predictors.

The MSE was higher for the first strong predictor than the other strong predictors (Figure 5 & Figure 6). This was due to the fact that only the first strong predictor was set to be correlated with the WBC variables, while the remaining strong predictors were all independent. Additionally, the MSE for the correlated strong predictor seemed to increase as more weak predictors were removed. LASSO had a lower MSE for this correlated strong predictor compared to that of the Stepwise Forward Selection. We believe this may be due to the fact that LASSO is a model estimator method while Stepwise Forward compares models for variable selection. In other words, LASSO addresses the collinearity by shrinking estimates.

When changing the ratio of weak and strong parameters, the patterns are mostly consistent. However, in the last scenario in which the ratio of strong:WAI:WBC signals was 10:5:5, the MSE was much lower compared to the other scenarios (Figure 7). This is most likely due to the fact that there was a greater number of strong predictors than WAI and WBC predictors, allowing for better estimations.

Conclusions

For our simulation study, we compared two selection methods in their ability to identify signals and make accurate parameter estimations in the presence of correlated variables. Overall, the Stepwise Forward Selection had higher identification rates compared to those of LASSO. LASSO, being a shrinkage method, estimates its values through a shrinkage parameter. Lambda was set at 1, putting a higher constraint on the coefficients, thereby leading to more zero coefficients. This may have played a factor in LASSO's relatively lower performance in identification rate. However, LASSO may be a better method to use if the goal is to estimate the true values. In terms of parameter estimation, LASSO showed lower MSE for correlated variables compared to those of the Stepwise Forward Selection.

There were some limitations of our study. Lacking computational power, we were only able to perform a relatively lower number of simulations. Additionally, we tested only on one value of lambda and correlation. For future studies, we intend to broaden our search parameters, comparing the two methods using a wider range of lambdas, correlations, and ratios of strong and weak predictors.

Figures

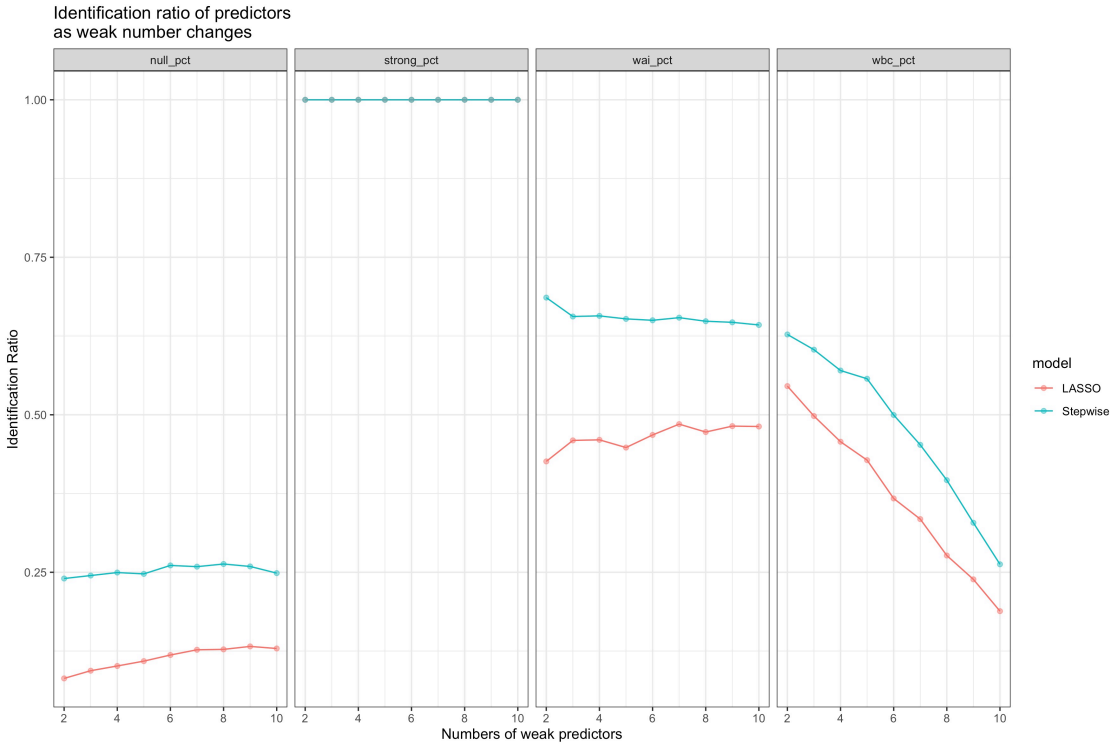

Figure 1: Identification ratio when number of weak predictors increase

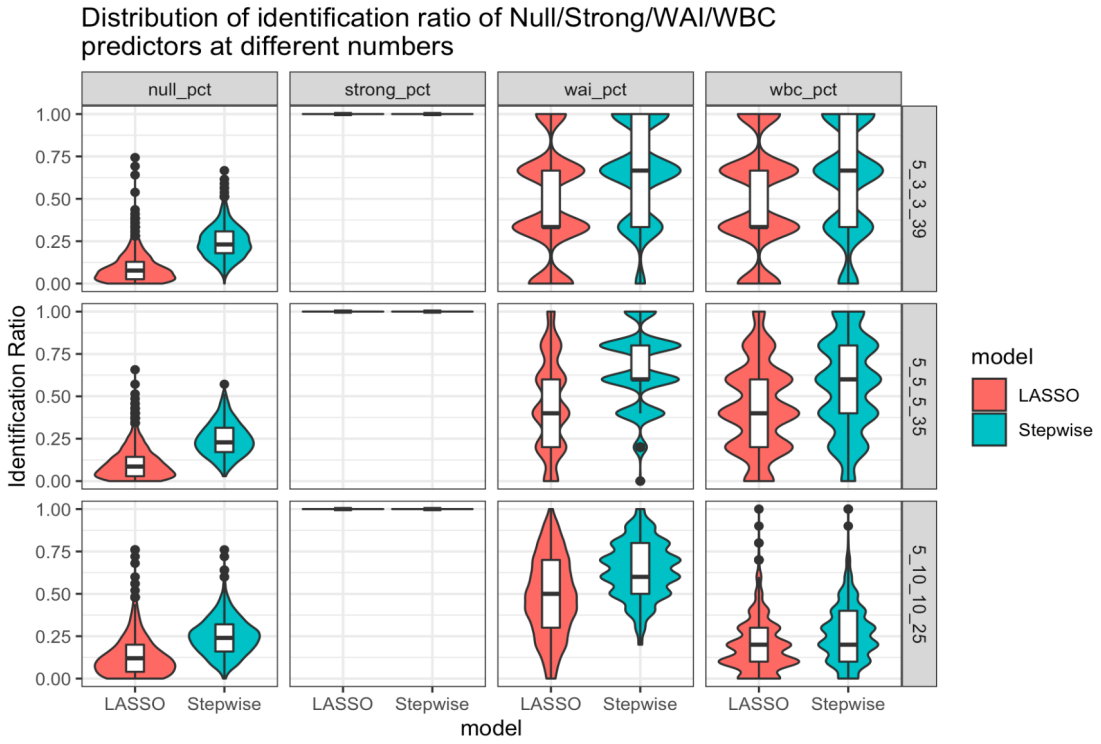

Figure 2: Distribution of identification ratio of Strong/WAI/WBC predictors at different numbers

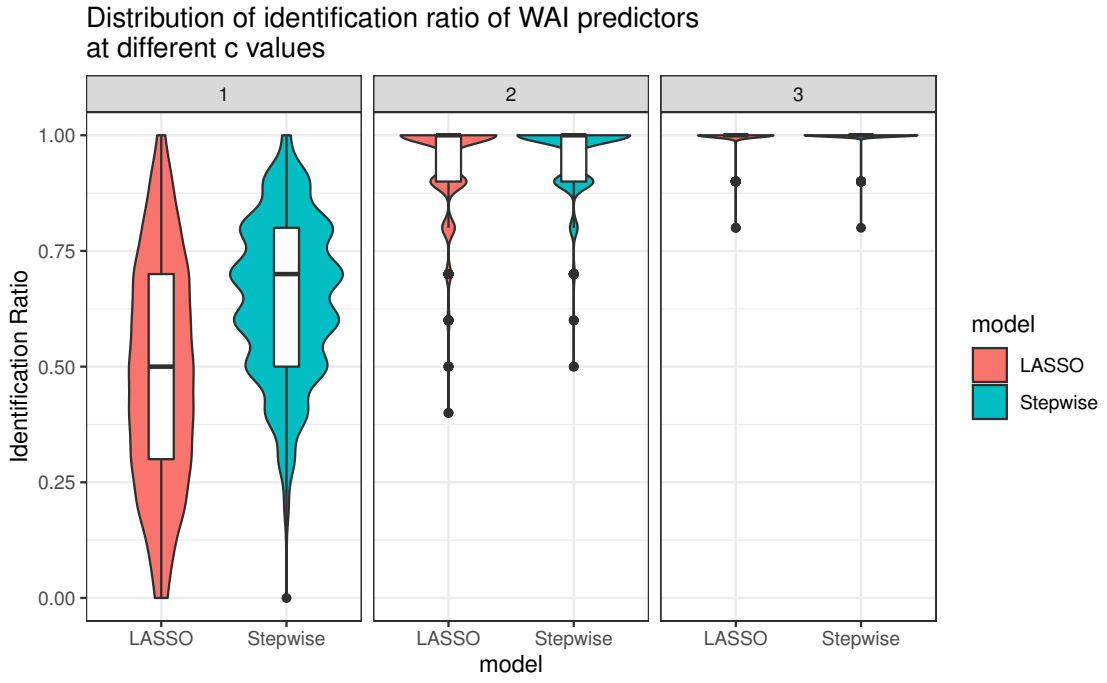

Figure 3: Distribution of identification ratio of WAI predictors at different c values

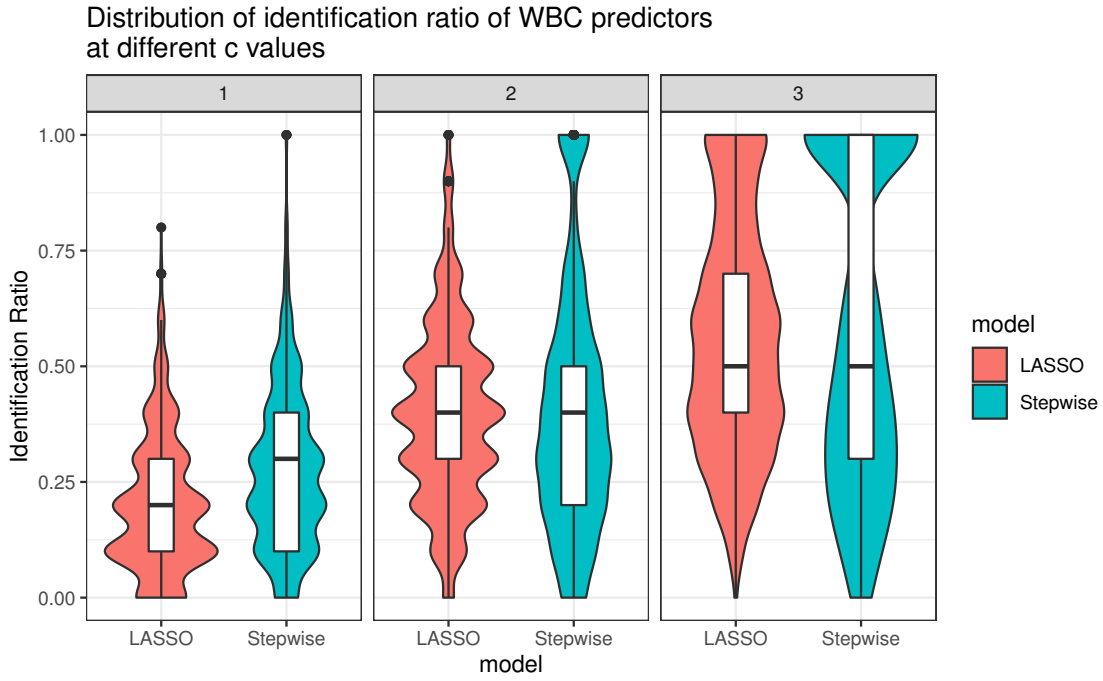

Figure 4: Distribution of identification ratio of WBC predictors at different c values

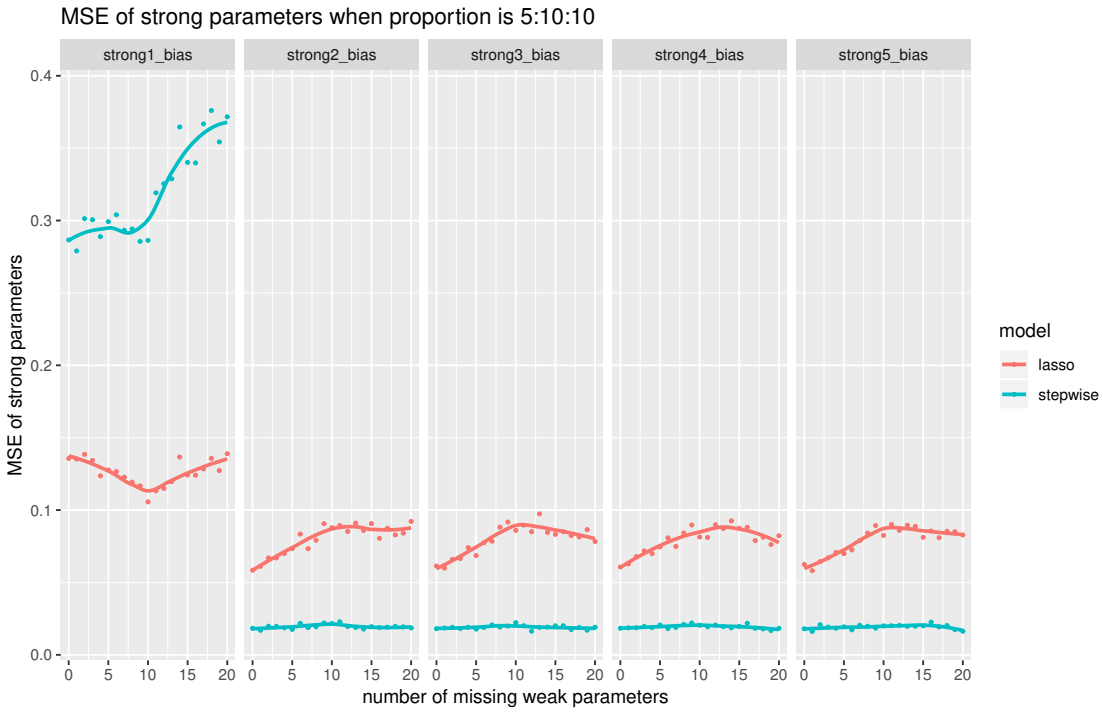

Figure 5: MSE of strong parameters when proportion is 5:10:10

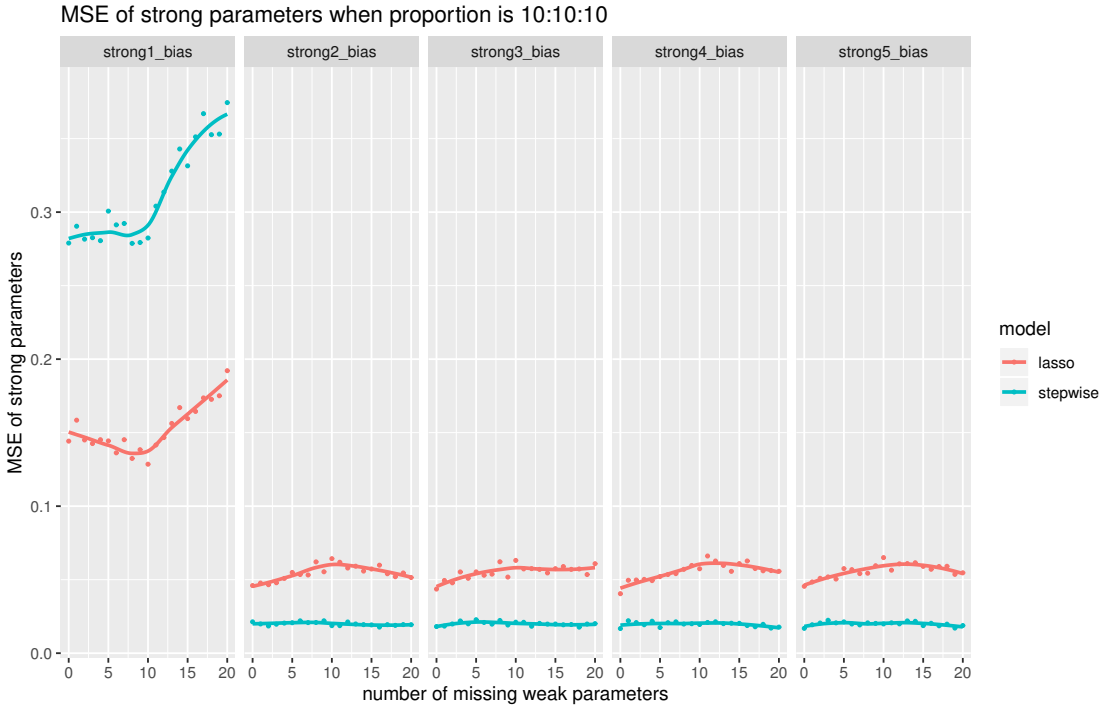

Figure 6: MSE of strong parameters when proportion is 10:10:10

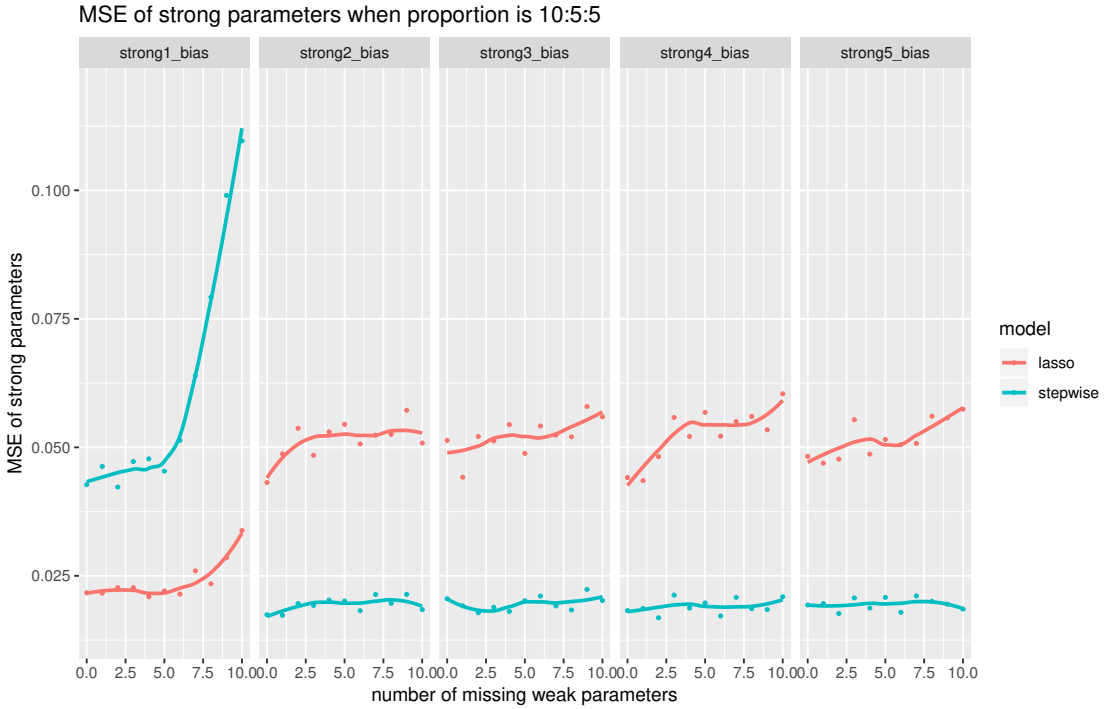

Figure 7: MSE of strong parameters when proportion is 10:5:5

Reference

1. Andrea Burton, Douglas G. Altman, Patrick Royston The design of simulation studies in medical statistics Statistics in Medicine 2006; 25:4279–4292

2. Li Y, Hong HG, Ahmed SE, Li Y. Weak signals in high-dimensional regression: Detection, estimation and prediction. Appl Stochastic Models Bus Ind. 2018;1–16. https://doi.org/10.1002/asmb.2340

3. Halabi S, Singh B. Sample size determination for comparing several survival curves with unequal allocations. Statistics in Medicine 2004; 23:1793–1815.