P8160 Project 1: Variable Selection Simulation Study

Group 3: Yujia Li, Ziqing Wang, Xicheng Xie

Mar 3. 2023

1. Objectives

When working with high-dimensional data, it is common to utilize variable selection methods to find out a model that optimally balances model complexity and fitness. Two widely-used classes of variable selection methods are subset selection and regularization methods. However, in the presence of both strong and weak predictors, both can suffer from failing to correctly select both while filtering out insignificant or null predictors. This behavior can lead to biased prediction, especially when there are more weak predictors in the data. The bias could be further complicated when correlation exists between strong and weak predictors. As such, the objective of this simulation study is two-fold. First, it aims to compare the performance in variable selection of two popular methods, stepwise forward selection and Automated LASSO regression, in the presence of weak predictors. Second, it aims to examine the impact of missing weak predictors on parameter estimations.

2. Statistical Methods To Be Studied

Stepwise regression is a variable selection procedure for independent variables (X_i) which consists of a series of steps designed to find the most useful X-variables to include in a regression model. The basis for selection could either be a) choosing a variable that satisfied the stipulated criterion, or b) removing a variable that least satisfies the criterion. Hereby, forward selection is applied, which iteratively examines the statistical significance of each independent variable in a linear regression model. The criterion of interest for each X-variable is chosen to be Akaike Information Criterion (AIC) defined as follows:

$$AIC = nln(\sum_{i=1}^{n}(y_i - \overline{y_i})^2/n) + 2p$$

where $\overline{y_i}$ is the predicted outcome, thus $\sum_{i=1}^{n}(y_i - \overline{y_i})^2$ is the sum squared error. $nln(\frac{SSE}{N})$ represents the uncertainty in the model and p is a penalty term for the number of parameters. Lower AIC values indicate a better-fit model, and a model with a ΔAIC (the difference between the two AIC values being compared) of more than -2 is considered significantly better than the model it is being compared to.

The adaptive LASSO is an alternative approach to improve variable selection. It seeks to minimize model coefficient magnitude and remove non-significant predictors through the following 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 the regularization parameter and the penalty is pen(λ) = λ|β|. By penalizing, some coefficients are shrunk to exactly zero, so the covariates associated with these coefficients are not retained in the model (Tibshirani, 1996). By controlling the amount of penalization, the λ parameter in the LASSO regression is closely related to the number of non-zero estimated coefficients. Although both Stepwise forward and LASSO selection methods penalize for the number of predictors involved, the constrained nature of LASSO could miss important variables and decrease sensitivity, especially if the underlying data structure is complex or nonlinear while Stepwise forward could overfitting by including irrelevant variables and thus decreasing specificity.

3. Methods

3.1 Defining strong and weak predictors

The strong, weak-but-correlated, and weak-and-independent predictors are defined as follows:

Strong predictors:

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

Weak but correlated (WBC) predictors:

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

Weak and independent (WAI) predictors:

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

Null predictors:

$S_4 = \{j: |\beta_j| = 0, 1 \le j \le p\}$

where p is the total number of predictors, n is the number of observations in the data, and c is a user-specified parameter that determines the threshold between strong and weak predictors.

3.2 Data Generation

To generate the dataset for simulation, firstly, the number of strong, weak, and null predictors are fixed. To fulfill the relationships defined between the predictors, a covariance matrix is initially set up as an identity matrix with the dimension equal to the total number of predictors. Then, the indices of the strong, weak, and null predictors are randomly chosen, and the correlation between the strong and WBC predictors is set to a chosen value. The design matrix *X* is then generated by sampling from a multivariate normal distribution with a zero mean and the aforementioned covariance matrix using the *mvrnorm()* function in R. The true regression coefficients for each predictor variable are set based on the definitions of strong, WBC, WAI, and null predictors, where the strong predictors are given larger coefficients than the

weak predictors, and the null predictors are given coefficients of 0. Finally, the outcome variable *Y* is generated by multiplying the design matrix by the true regression coefficient vector and adding a random error term that follows a standard normal distribution.

3.3 Scenarios to be investigated

To continue investigating the 2 problems interested, 4 scenarios with different predictors' structure and property are also adopted. The threshold multiplier c that defines strong and weak predictors as well as the correlation pattern between strong and WBC predictors are changeable in order to note whether their variation affects the model performance in detecting different predictors. The number of each signal type separately is also altered to see if it influences the model performance. Detailed numbers set up shown in the table below.

Scenario 1: Varying correlation between strong and WBC predictors (p)	Total number of WBC predictors: 10 Total number of WAI predictors: 10 Total number of null predictors: 40 - 5 - 10 - 10 = 15 c = 3 p = 0.1, 0.3, 0.5, 0.7, 0.9
Scenario 2: Varying the number of WBC predictors	Total number of WAI predictors: 10 c = 3 p = 0.3 Total number of WBC predictors: 2, 6, 10, 14
Scenario 3: Varying the number of WAI predictors	Total number of WBC predictors: 10 c = 3 p = 0.3 Total number of WAI predictors: 2, 6, 10, 14
Scenario 4: Varying the threshold parameter c	Total number of WAI predictors: 10 Total number of WBC predictors: 10 p = 0.3 c = 1, 3, 5

NOTE: For each setup within each simulation scenario, 100 simulation trials were repeated for both methods. The number of observations in each generated dataset is 10000. The number of predictors for all these trials is 40. The number of strong predictors for all these trials is 5. The number of strong, WBA, WAI, and null predictors adds up to the total number of predictors (40).

3.4 Performance measures

We use various measures to evaluate the performance of stepwise forward and LASSO in identifying and selecting weak and strong predictors. Two main measures are used to test the ability to select true predictors: Power and Type-1 error, as shown below.

- Power $= \frac{number\ of\ true\ predictors\ selected\ in\ the\ model}{total\ number\ of\ true\ predictors\ in\ the\ data} \times\ 100\%$

- Type-1 error= $\frac{number\ of\ null\ predictors\ selected\ in\ the\ model}{total\ number\ of\ null\ predictors\ in\ the\ data} \times\ 100\%$

Based on the definitions, A high power indicates that the method is effective in identifying strong and weak predictors. A low type 1 error indicates that the method is selecting few null predictors. We also measured the percentage of strong, WBC, and WAI predictors selected to evaluate the methods' ability to identify different types of predictors. Their definitions are shown below.

- % of strong predictors selected = $\frac{number\ of\ strong\ predictors\ selected\ in\ the\ model}{total\ number\ of\ strong\ \ predictors\ in\ the\ data} \times\ 100\%$
- % of WBC predictors selected = $\frac{number\ of\ WBC\ predictors\ selected\ in\ the\ model}{total\ number\ of\ WBC\ predictors\ in\ the\ data} \times\ 100\%$
- % WAI predictors selected = $\frac{number\ of\ WAI\ predictors\ selected\ in\ the\ model}{total\ number\ of\ WAI\ predictors\ in\ the\ data} \times\ 100\%$

Regarding the parameter estimations, we are interested in evaluating how missing weak predictors impact the parameter estimations for stepwise forward and LASSO. To access this, we measure the sum of squared error (SSE) and the mean squared error (MSE) of coefficient estimates in each simulation among all the scenarios according to the following equations:

- $SSE = \sum_{i=1}^{p} (\hat{\beta}_i - \beta_{i0})^2$
- $MSE = \frac{1}{p} \sum_{i=1}^{p} (\hat{\beta}_i - \beta_{i0})^2$

where p equals the total number of predictors, $\hat{\beta}_i$ is the estimated regression coefficient for the ith predictor, and β_{i0} is the true value of the regression coefficient for the ith predictor.

4. Simulation Results

Task 1: Evaluate the two method's performance in selecting different types of predictors across various scenarios

Scenario 1: We hypothesized that, as the correlation between strong and WBC predictors increases, the power of both methods will decrease. The intuition was that high correlation between strong and WBC might obscure the magnitude of the weak effect of WBC predictors on the response variable. As a result, both methods are more likely to select strong predictors only. The simulation results indeed showed a decrease in power between $p = 0.1$ and $p \geq 0.3$ for both methods, as shown in the median powers in Figure 1. Moreover, the forward selection method consistently has higher power on average than LASSO. However, forward selection generally has higher type-1 error rates than LASSO for all values of p, as suggested by Figure 2.

Both methods did well in selecting strong predictors (100% strong predicted selected) regardless of the value of p. However, p affects both method's performance in selecting WBC predictors and WAI predictors differently, according to Figure 3 and Figure 4. Specifically, for both methods, % WBC predictors selected decreased as the correlation increases, but % WAI predictors selected remained stable.

Scenario 2: Figure 5 suggests that, as the number of WBC predictors increases, the power of both methods decreases. Forward selection on average has a higher power and type-1 error rate than LASSO across different numbers of WBC predictors, as suggested in Figure 5 and Figure 6, respectively.

Both methods were able to select all strong predictors, However, as the number of WBC predictors increases, the % of WBC predictors selected decreases for both methods, while the % of WAI predictors selected by both methods were unaffected, according to Figure 7 and Figure 8. Forward selection selects more WBC and WAI predictors than LASSO, which is expected because forward selection selects more predictors.

Scenario 3: As the number of WAI predictors increases, the power for both methods slightly increases. Forward selection consistently has higher power and type-1 error rate than LASSO across all different numbers of WAI predictors examined, as indicated by Figure 9 and Figure 10.

Both methods selected all strong predictors. As for the weak predictors, increasing the number of WAI predictors slightly increased LASSO's, but not forward selection's performance in selecting WBC and WAI predictors, according to Figure 11 and Figure 12.

Scenario 4: Recall the multiplier c as a component of defining strong and weak predictors, we found that increasing c improves both stepwise and LASSO's ability on detecting weak predictors. It stands to reason that increasing c allows weak predictors to take on higher values, leveling the ability of the model to detect them. Since statistical power is affected by the size of the effect, where larger effects are more easily detected, we expect that increasing the threshold multiplier c increases the power. Indeed, both methods' power increases as the value of c increases according to Figure 13. However, changing the value of c did not alter either method's type-1 error rate, as the median type-1 error rate remained similar in Figure 14. Again, forward selection on average has higher power and type-1 error rate than LASSO across all different values of c considered.

Both methods selected all strong predictors in all simulation trials. It's worth noting that along with the increment in c , both methods detect more WAI and WBC predictors, as shown in Figure 15 and Figure 16. However, the % of WBC predicted selected on average is lower than the % of WAI predictors selected for both methods. It might be caused by WBC's correlation to strong predictors which hampered the model performance.

Task 2: Examine how missing weak predictors affects the two method's parameter estimations

We investigate the relationship between the percentage of missing weak predictors of every scenario and the corresponding SSE and MSE of the parameter estimations. From Figure

17 and 18, as the proportion of missing weak predictors increases in the model, the SSE and MSE of parameter estimations decreases in both forward selection and LASSO. For stepwise forward selection, when the proportion of missing weak predictors reaches to 0.1, it seem to be a cut-off point for the SSE and MSE of parameter estimations: when the proportion is larger than 0.1, the SSE and MSE decrease sharply and the variance shrink at the same time, as shown in Figure 17 and Figure 18. This seems counterintuitive since one would expect that missing more weak predictors in a model would increase the SSE and MSE. However, one possible explanation for this phenomenon is that, in order to miss fewer weak predictors, more variables need to be included in the model, increasing the probability of selecting null predictors. This increases the risk of making Type 1 errors, which can inflate the errors in the estimation of parameter coefficients. This can ultimately lead to an increase in SSE and MSE despite missing fewer weak predictors in the model.

Moreover, our simulation results for the LASSO method did not include any instances where the proportion of missing weak predictors fell within the range of 0 to 0.1, while the simulation results for the stepwise forward method did not encompass any scenarios where the proportion of missing weak predictors ranged from 0.8 to 1. This observation is reflected in scenarios 1-4, where forward selection consistently selects more weak predictors and more total predictors, compared to LASSO.

It is also worth noting that, for most proportions of missing weak predictors, LASSO generally has smaller MSE and SSE on average than forward selection. It suggests that LASSO may be more robust to missing weak predictors when compared to forward selection. One possible explanation is that the LASSO method is designed to perform regularization, which means that it can more effectively handle situations where there are many irrelevant or weak predictors in the data. In contrast, the stepwise forward method relies on a sequential process of adding predictors to the model, which can be more sensitive to missing predictors. However, it's important to note that these results are based on a specific set of scenarios used in this simulation study. A wider range of scenarios need to be investigated before making more conclusive comments.

5. Conclusions

In terms of model performance, stepwise forward selection consistently has higher power and higher Type-I error than LASSO. In scenarios where the predictors are strong, both methods perform well with a 100% success rate on selecting them. In scenarios examining how weak predictors act in the models, increasing the number of WAI predictors does not have much effect on either method's performance in selecting weak predictors. Yet, increasing the number of WBC predictors decreases the performance of both methods in selecting WBC predictors but not WAI predictors. In addition, the higher the correlation coefficient between strong and WBC predictors, the less the power of forward selection and LASSO models performs. To conclude, both methods are effective in selecting strong predictors. If one's main goal is to identify weak predictors or to

achieve a high statistical power, forward selection is the better option. However, if one's goal is to keep the type-1 error low, LASSO is the preferred method.

In the presence of missing weak predictors, the forward selection method may be more sensitive to model misspecification compared to LASSO. This is because forward selection relies on a stepwise approach of adding predictors based on individual significance, and if a predictor is missing, it may select an alternative predictor that is not the best predictor for the model. As a result, the predictive accuracy of the model may decrease, which can be reflected in higher MSE (Mean Squared Error) and SSE (Sum of Squared Error) values.

On the other hand, LASSO is more robust to missing predictors because it performs regularization, which helps to handle high-dimensional data with many weak or irrelevant predictors. By shrinking the coefficients of weak predictors towards zero, LASSO can effectively handle missing predictors without adversely affecting the performance of the model. Therefore, on average, LASSO may produce models with smaller MSE and SSE compared to forward selection when missing weak predictors.

6. Contributions

Yujia Li: Preliminary simulation setup, simulation scenario 4, slides & report writing
Ziqing Wang: Data generation, simulation scenarios 1, 2, 3, slides, report writing & editing
Xicheng Xie: Data generation, parameter estimation, report writing & editing

Appendix

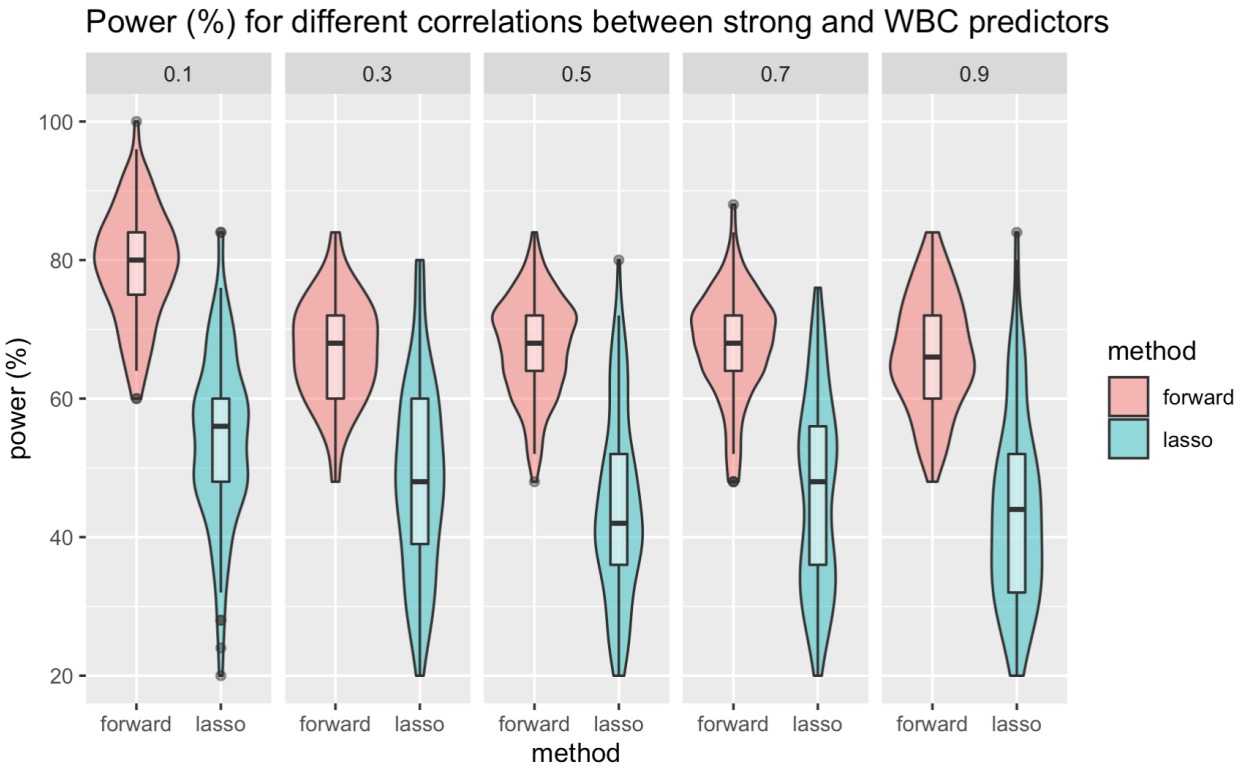

Figure 1: Power (%) for different correlations between strong and WBC predictors

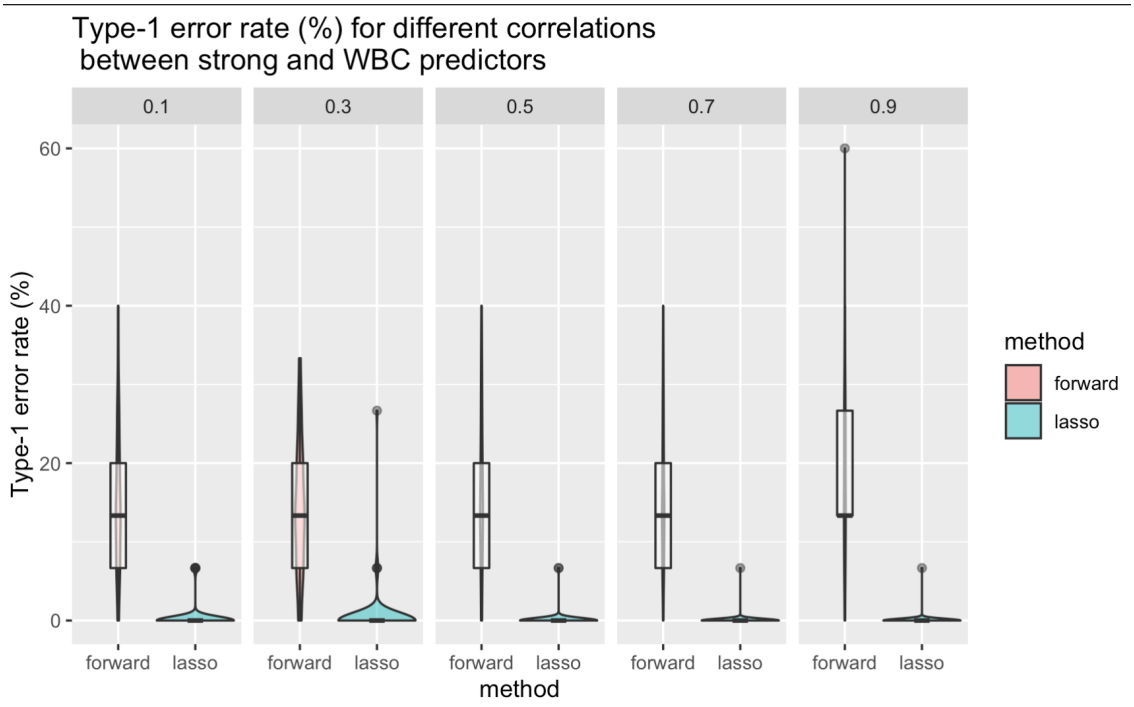

Figure 2: Type-1 error rate (%) for different correlations between strong and WBC predictors

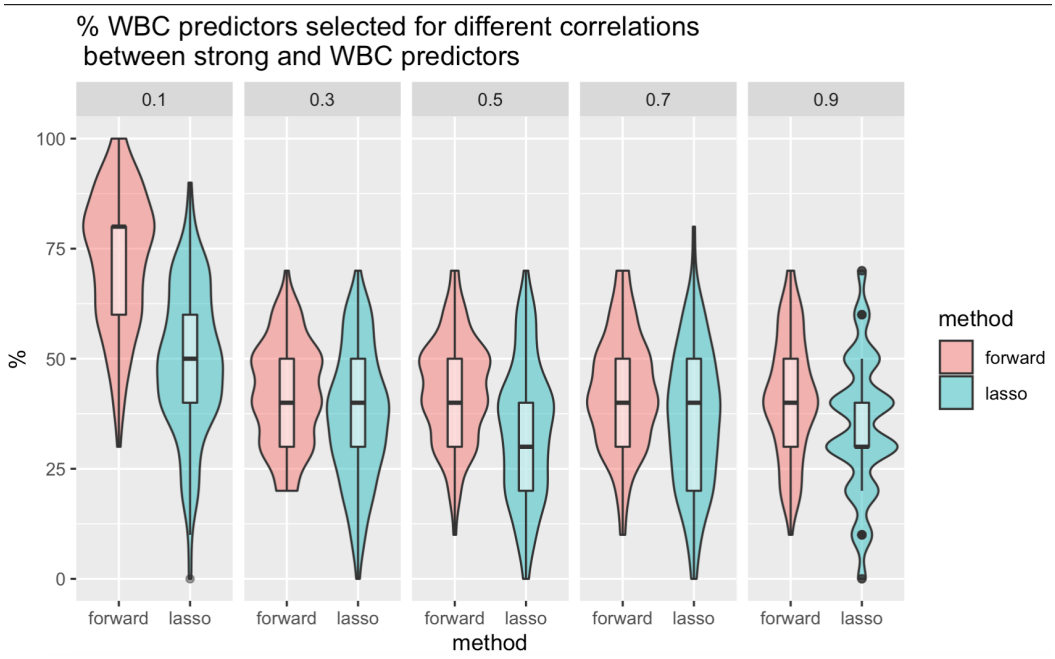

Figure 3: % WBC predictors selected for different correlations between strong and WBC predictors

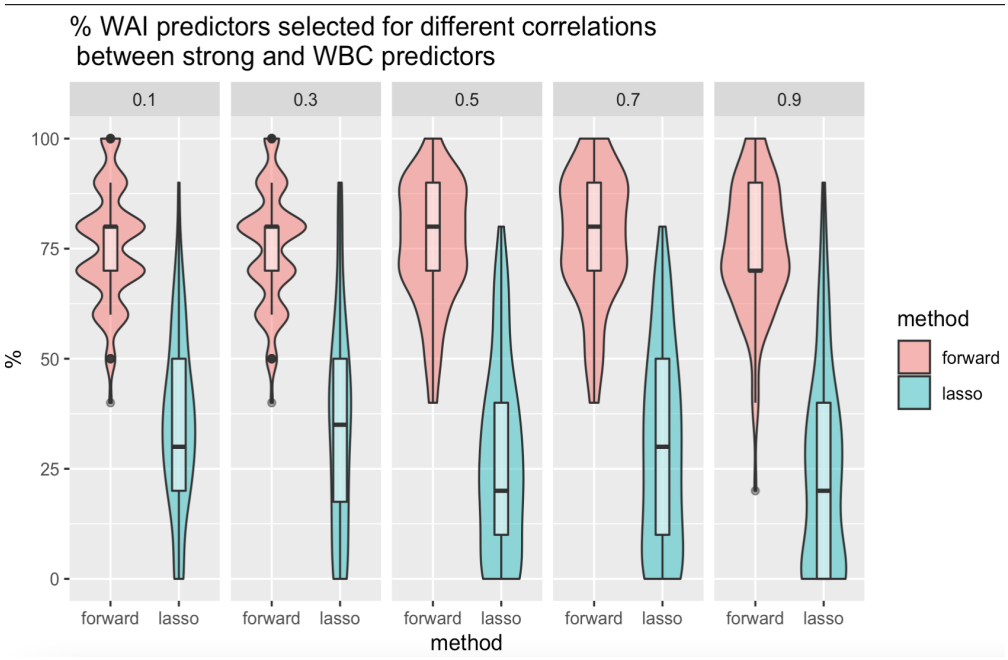

Figure 4: % WAI predictors selected for different correlations between strong and WBC predictors

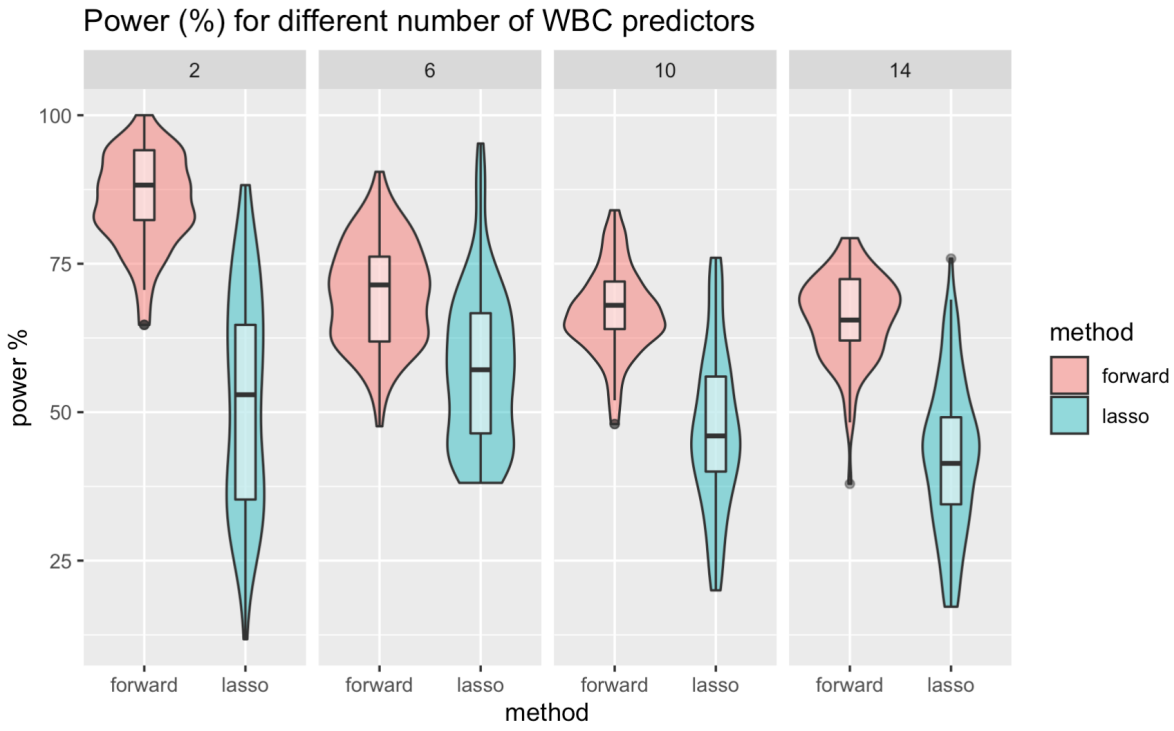

Figure 5: Power for different number of WBC predictors in the data

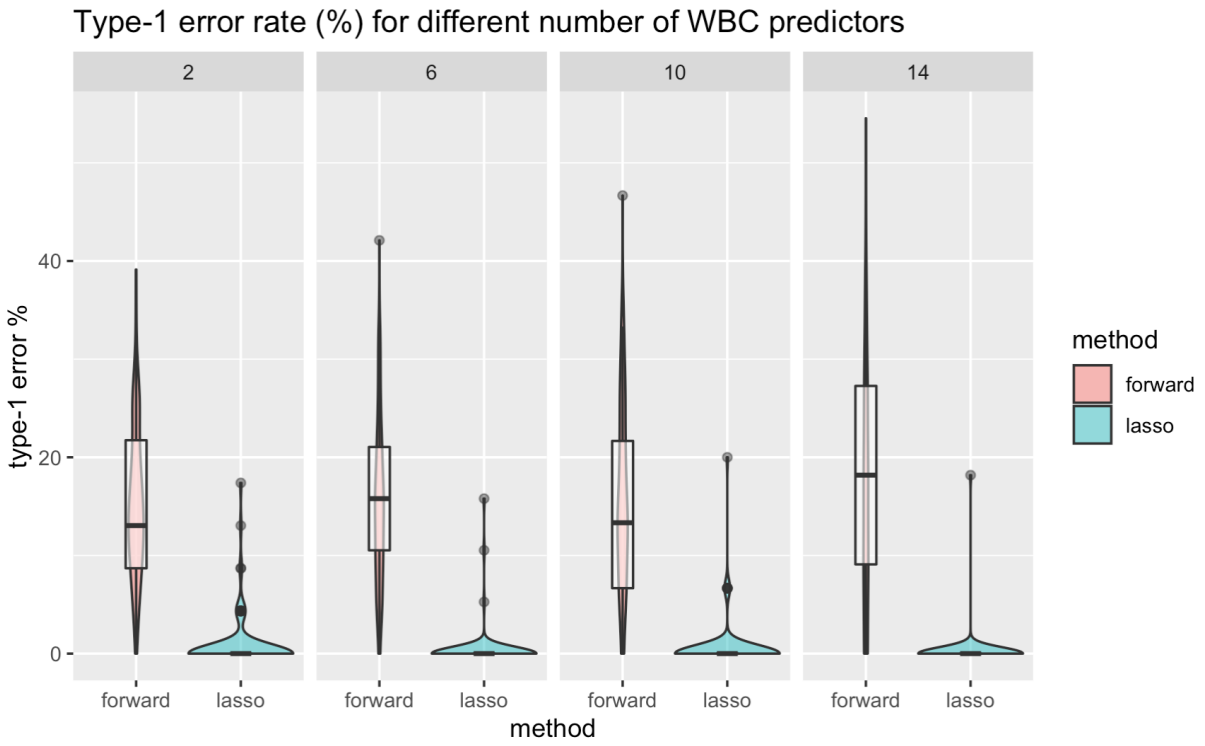

Figure 6: Type-1 error rate for different number of WBC predictors in the data

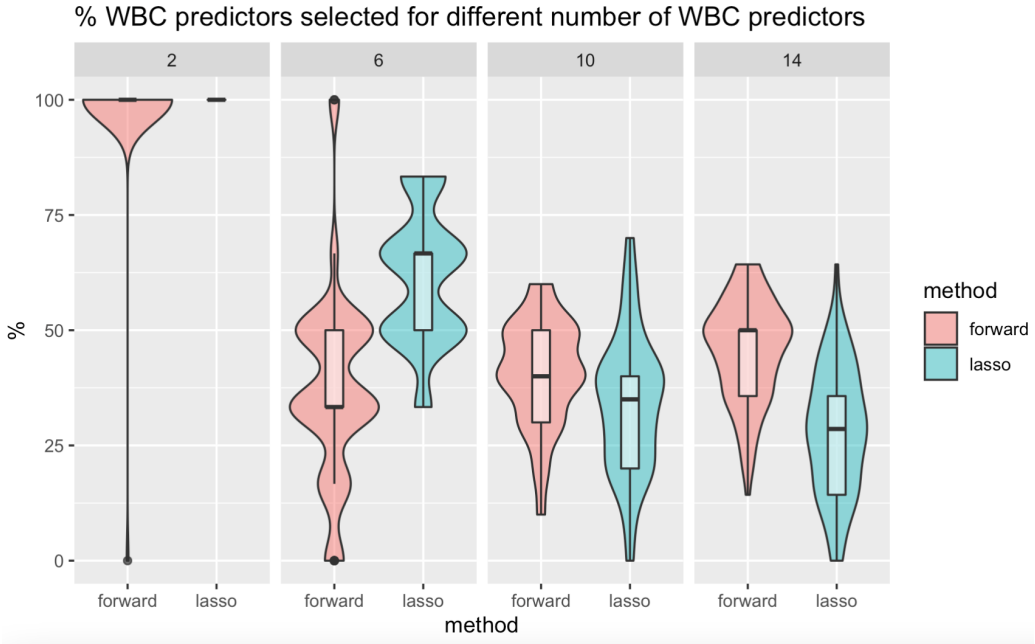

Figure 7: % WBC predictors selected for different number of WBC predictors in the data

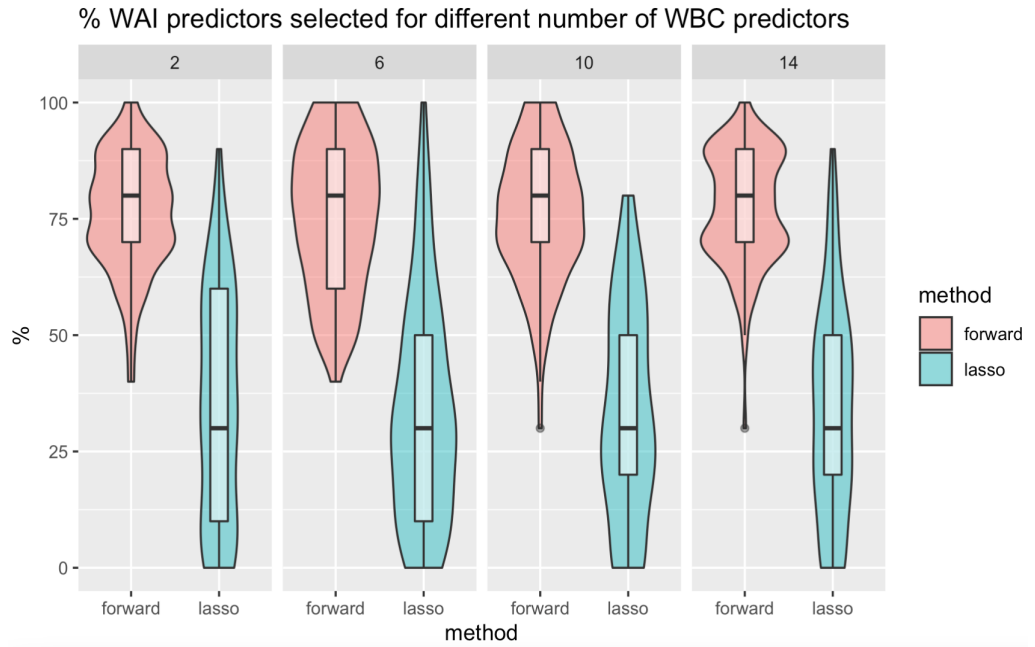

Figure 8: % WAI predictors selected for different number of WBC predictors in the data

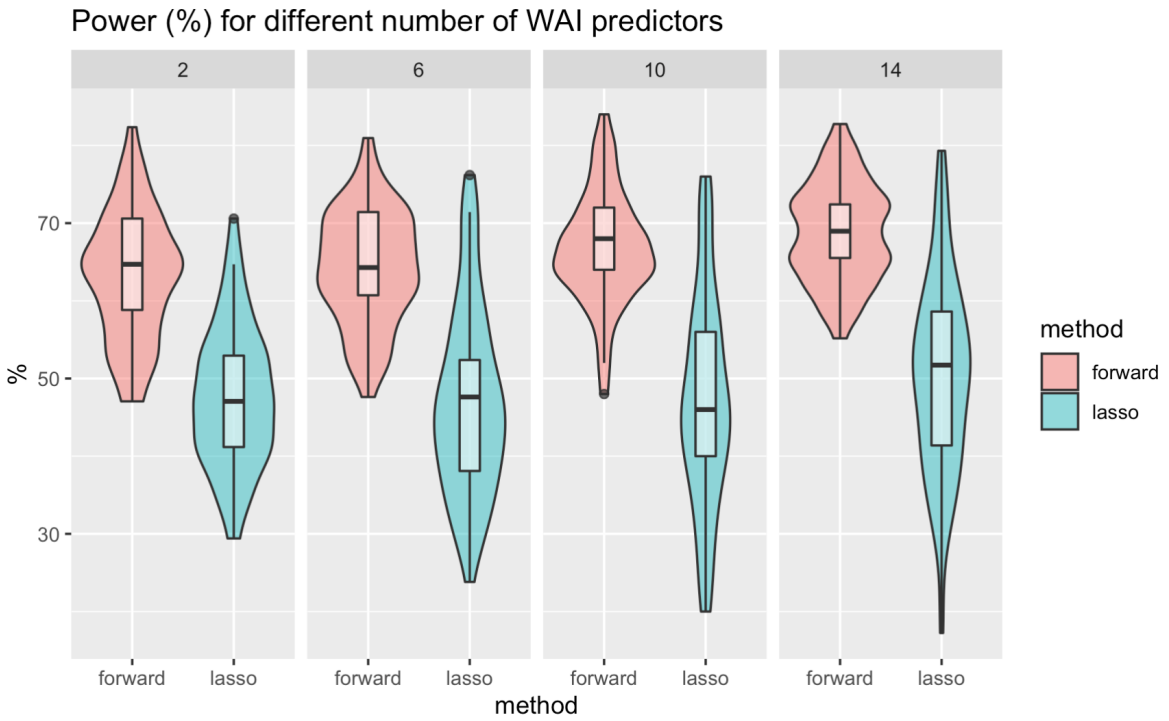

Figure 9: Power for different number of WAI predictors in the data

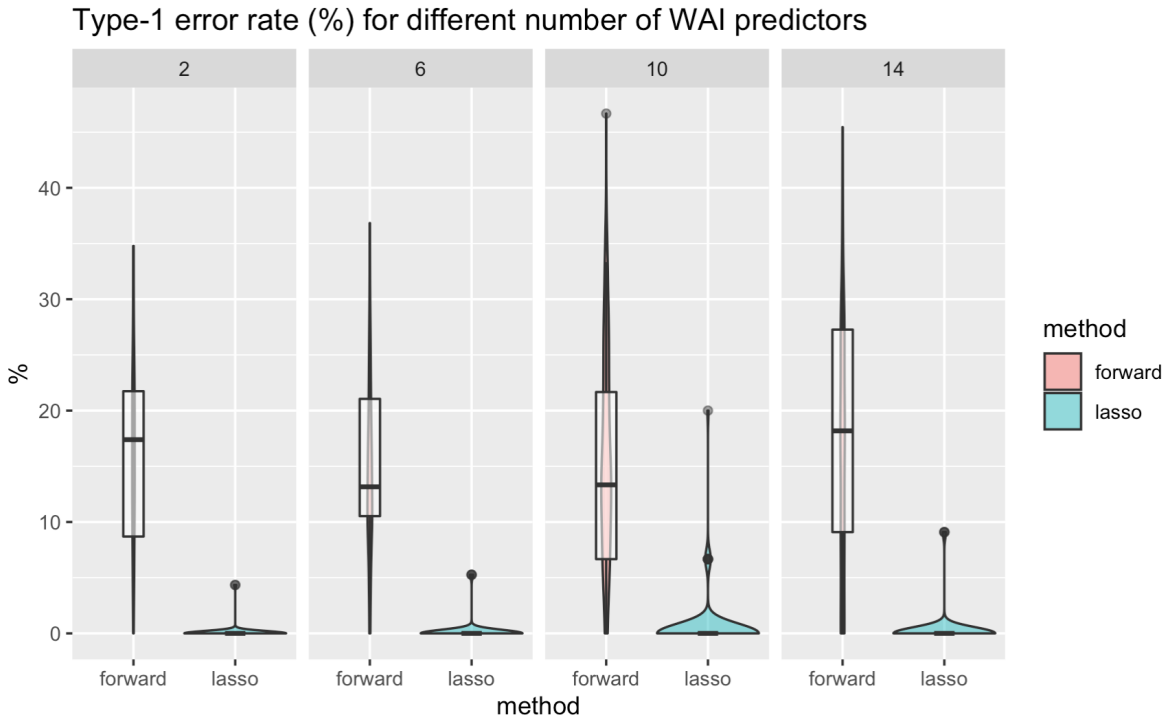

Figure 10: Type-1 error rate for different number of WAI predictors in the data

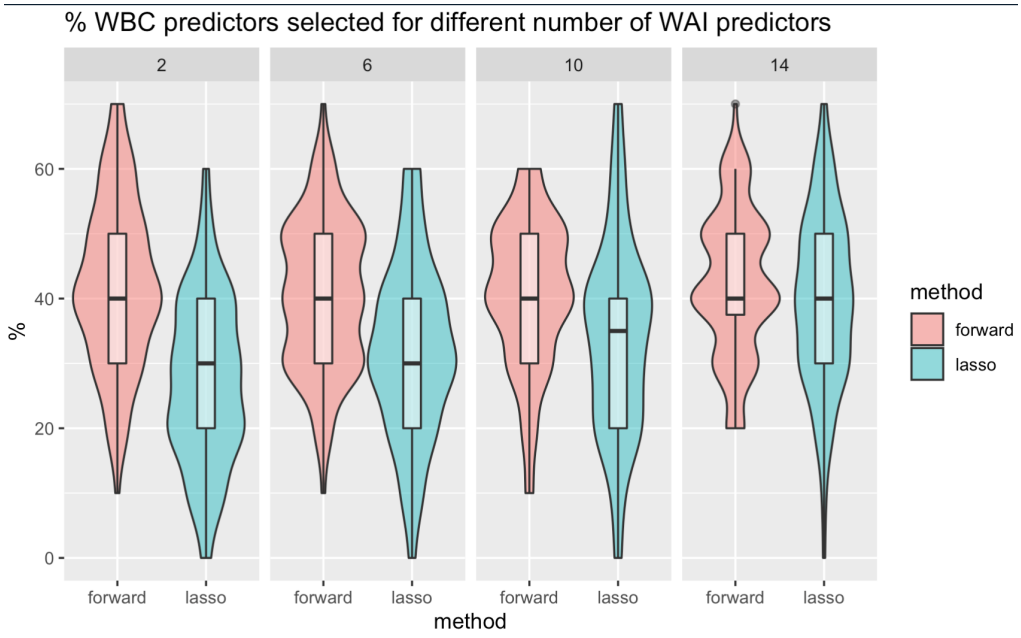

Figure 11: % WBC predictors selected for different number of WAI predictors in the data

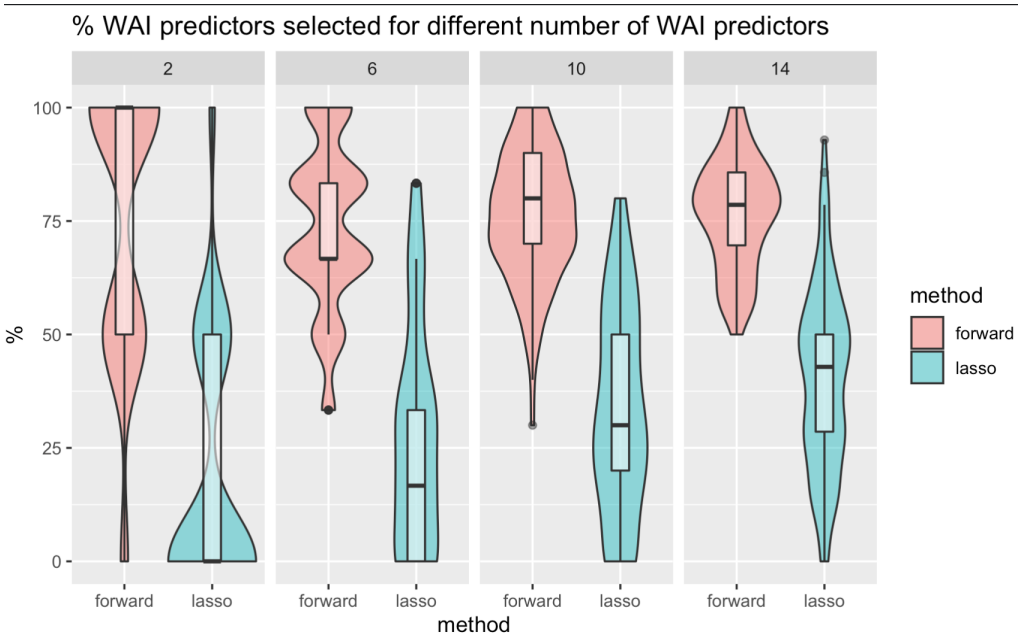

Figure 12: % WAI predictors selected for different number of WAI predictors in the data

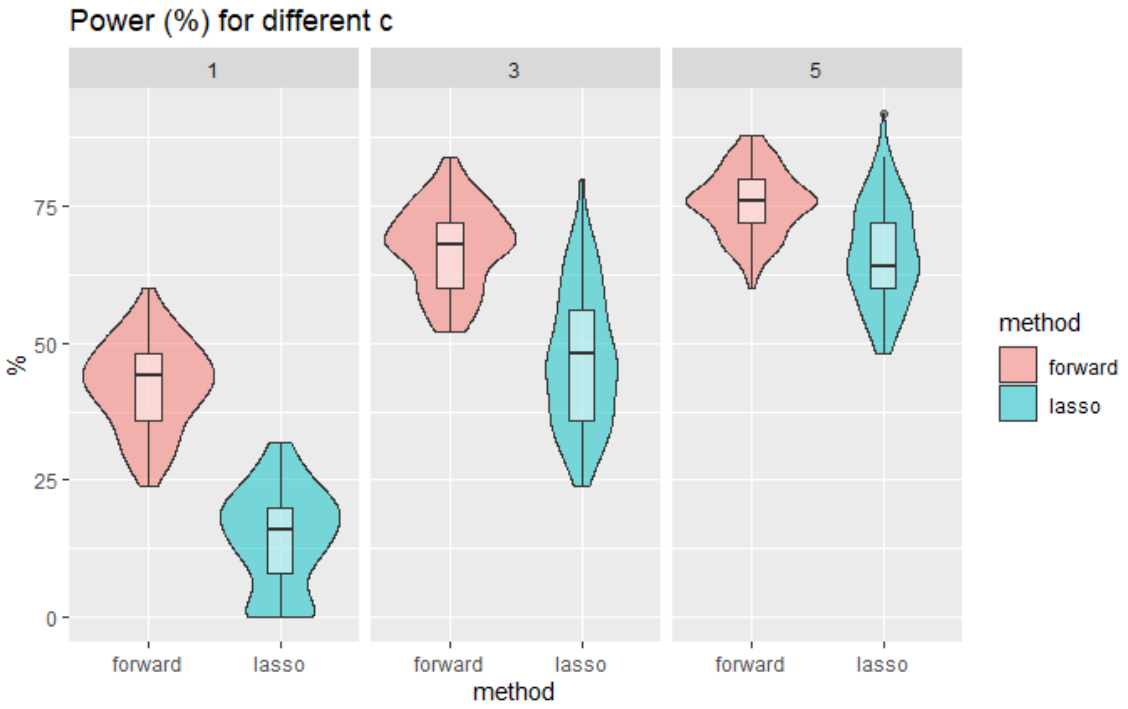

Figure 13: Power for different c

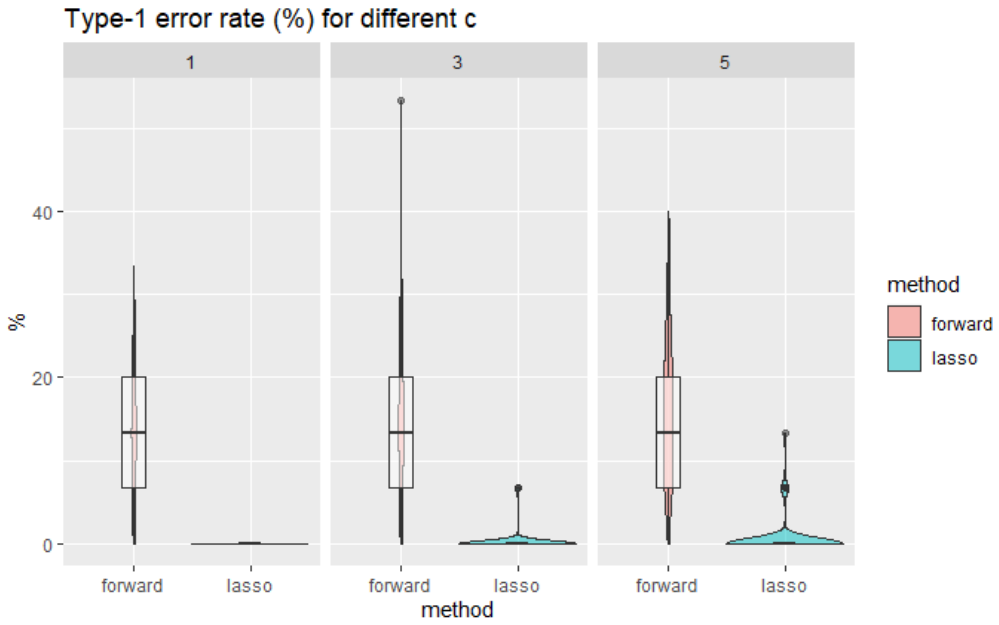

Figure 14: Type-1 error rate for different c

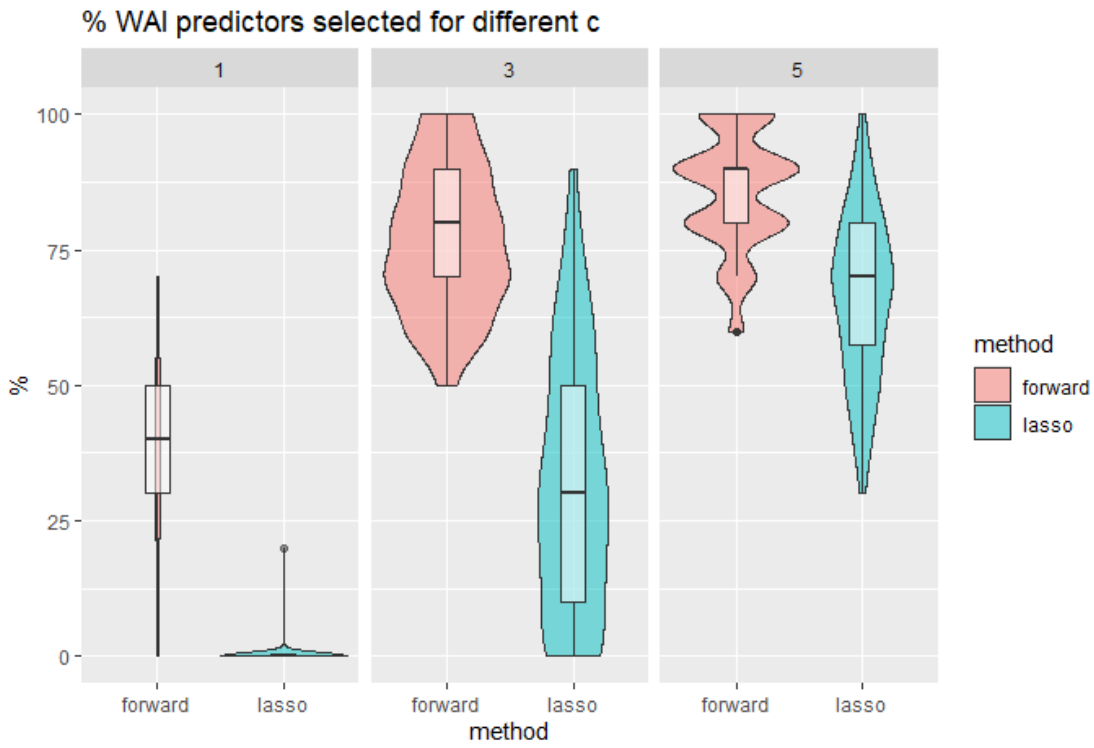

Figure 15: % WAI predictors selected for different c

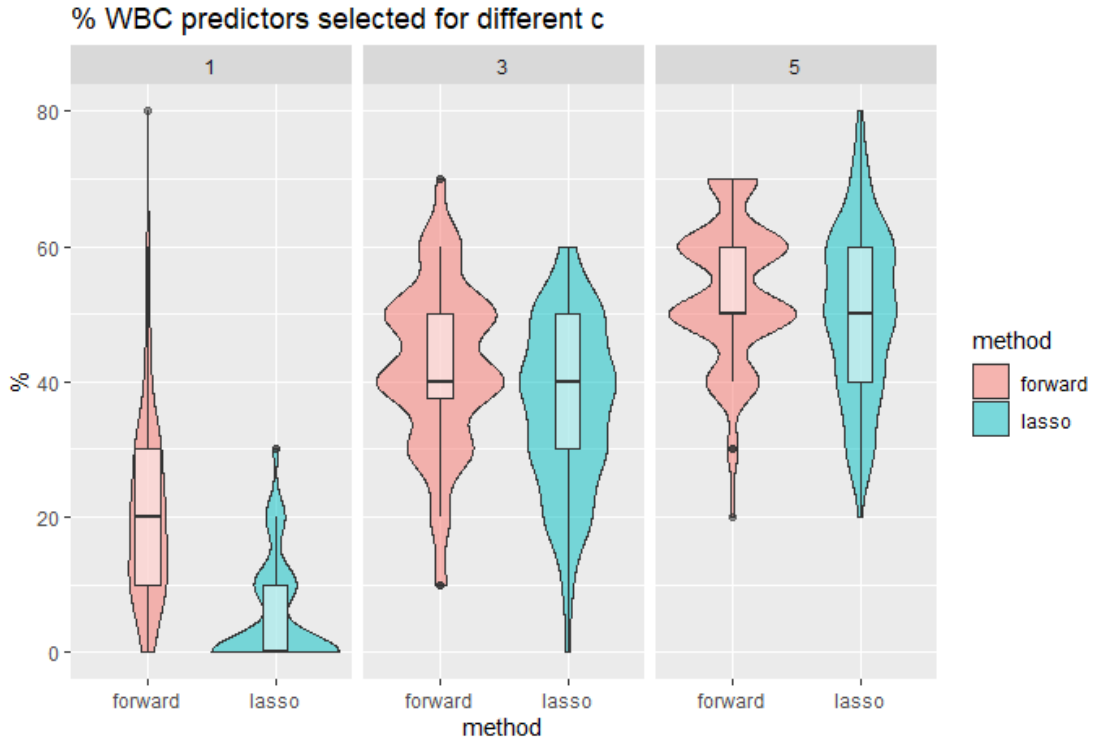

Figure 16: % WBC predictors selected for different c

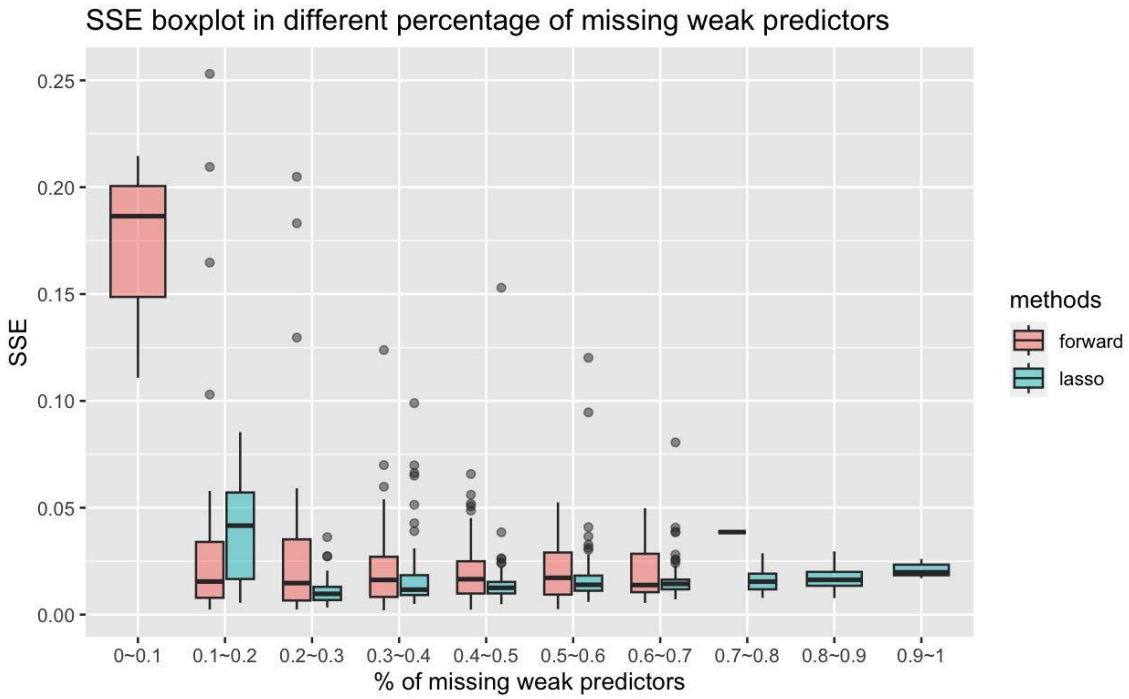

Figure 17: SSE for different % of missing weak predictors

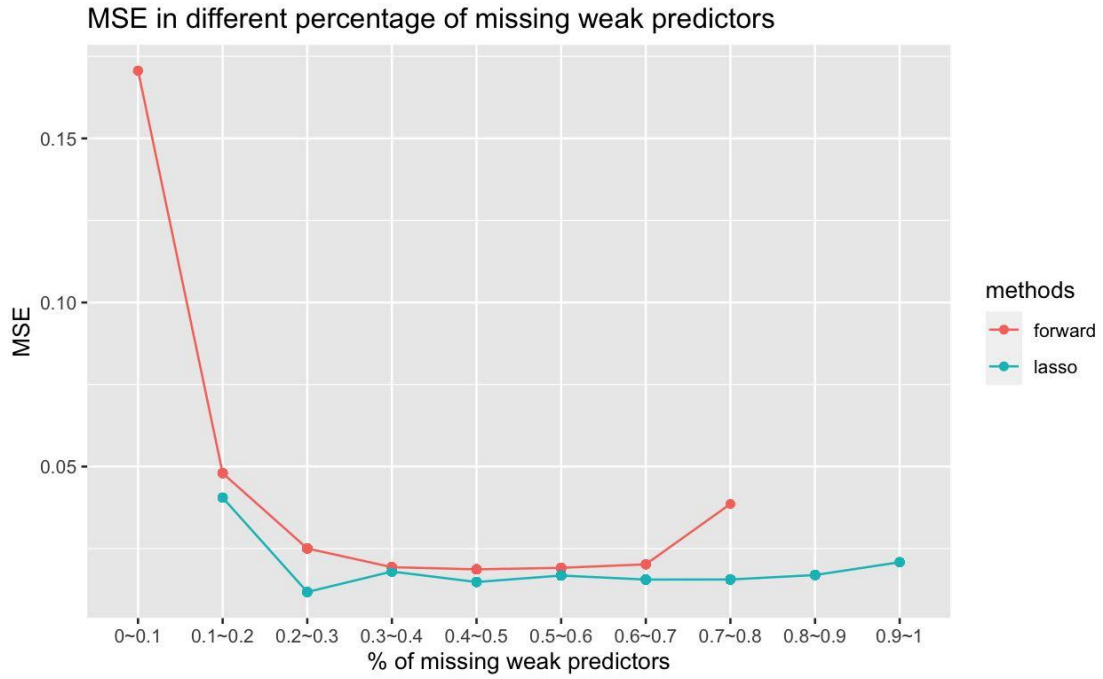

Figure 18: MSE for different % of missing weak predictors

Reference

Burton, A. , Altman, D. G., Royston, P. and Holder, R. L. (2006), The design of simulation studies in medical statistics. Statist. Med. 25: 4279-4292. doi:10.1002/sim.2673

Li Y, Hong HG, Ahmed SE, Li Y. Weak signals in high-dimension regression: detection, estimation and prediction. Appl Stoch Models Bus Ind. 2019 Mar-Apr;35(2):283-298. doi: 10.1002/asmb.2340. Epub 2018 May 25. PMID: 31666801; PMCID: PMC6821396.

Tibshirani, R. (1996), Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58: 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x