

p8160-project1

1 Objectives

Design a simulation study to investigate and illustrate how well each of the two methods in identifying weak and strong predictors; how missing “weak” predictors impacts the parameter estimations. To do so, you need to simulate data with a combination of strong”,weak-but-correlated” and “weak-and- independent” predictors.

2 Statistical methods to be studied

2.1 Step-wise forward method

Starting with the empty model, and iteratively adds the variables that best improves the model fit. That is often done by sequentially adding predictors with the largest reduction in AIC. For linear models,

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

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

2.2 Automated LASSO regression LASSO

another popular method for variable selection. 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 tuning parameter. We use Cross-validation (CV) to select λ for LASSO.

3 Scenarios to be investigated

- weak-to-strong predictor ratio; In this study, we want to simulate the situation where the predictors are having underlying correlation, we define certain scenarios to compute the robustness of 2 variable selection methods.

3.2 tuning parameter Sample size: 1) $n = 200$ 2) $n = 1000$ 3) $n = 5000$ Strength predictors ratio: The total number of predictors are 200, where 100 are noise predictors. Here we consider 2 strength predictors ratio (r: strength: weak_but_correlated: weak_and_independent) 1) $r = 5:1:4$ 2) $r = 5:3:2$ 6 combinations of above parameters are the scenario to be studied

4 Methods for generating data

the simulated population needs to meet following characteristics Firstly, the expectation of outcome variable is the linear combination of predictors with an constant-variance error term; Secondly, the predictors are

mutually correlated, Thirdly, the parameters are distinctly correlated with outcome, which is classified by a critical value. The detailed definitions of data generating function parameters are as follows:

4.1 Distribution of population

True distribution of outcome variable is defined as

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

Where X is the predictor matrix, β is the parameters vector and σ^2 is the constant variance in normal distribution.

4.2 Predictor strength

The strength of parameters are defined by following criterias:

Definition of strong predictors:

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

Definition of weak but correlated predictors:

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

Definition of weak and independent predictors:

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

Definition of noise:

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

Parameters can be generated by 2 ways. Firstly, given beta by artificial setting, and secondly, by defining constant c , parameters are generated uniformly with in above ranges.

4.3 Parameter correlation

The data of predictors are generated by R function `mvrnorm`. In order to change the correlation between different predictors, a pre-defined covariance matrix is passed to `mvrnorm` function. Here we consider a certain scenario: the strong variables and a certain ratio of weak variables are indepent to other variables, and other weak predictors are correlated with 1 specific strong variable. To ensure the positive definite attribute of covariance matrix, we restrict one strong predictor can only be correlated with one weak predictor.

5 performance measures

5.1 Predictor identification performance

In order to compare the identification performance for these 2 methods, we regard the identification as a classification problem, where the signal predictors are defined as positive and null predictors are defined as negative. 3 indicators have been established:

$$\text{recall} = \text{sensitivity} = \left(\frac{\text{True positive}}{\text{True positive} + \text{False negative}} \right)$$

$$specificity = \left(\frac{True\ negative}{True\ negative + False\ positive} \right)$$

$$accuracy = \frac{True\ selection}{total\ number\ of\ predictors}$$

$$F_1 = \left(\frac{2}{recall^{-1} + precision^{-1}} \right)$$

Where precision is defined as above, and recall is defined as

$$precision = \left(\frac{True\ positive}{True\ positive + False\ positive} \right)$$

5.2 estimation performance

To compare the parameter estimation performance, 3 indicators were calculated to evaluate the estimation: bias is calculated as the mean difference between true parameter and estimated parameters, variance is defined as the variance of the estimated parameter among simulation, and MSE

By tuning parameters differently, robustness of the variable selection methods were also evaluated through the above indicators.

6 simulation results.

6.1 identification performance

F1 score (see Fig 1.) is used as an indicator for an overall performance assessment for identification signals. There is a clear trend for both methods that when the ratio of predictors to sample size is decreasing, the overall performance enhances, and there is no significant difference between these two methods when sample size is large enough comparing to predictor numbers, while in high dimensional occasion (where $n = p$), stepwise has a poor performance.

In terms of variable selection size (see Fig 2.) , lasso tends to choose more predictors while stepwise tends to choose less overall, in the settings of this report, where there are 100 true signals and 100 noises, lasso selects more predictors than true occasion in all scenarios, while stepwise tends to choose less. When comparing in terms of sample size, lasso performs stably in both high dimensional (where $n = p$) and normal scenario, but stepwise tends to lose its select ability when the ratio between number of parameters and sample size is large.

Sensitivity (see Fig 3.) is an indicator for how many predictors can be captured by the methods. For lasso, strong predictors can be perfectly captured in all scenarios with stable performance, while for weak predictors, difference occurs when correlated-to-independent ratios change, for less correlated scenario, lasso tends to perform better to capture weak and independent predictors compared to weak but correlated predictors, while in more correlated case, there is no clear difference between these two types of predictors, and performance for capturing weak but correlated predictors is relatively stable. For stepwise methods, except for the same stability for capturing weak but correlated predictors in more correlated scenario, there is no clear difference in different settings. When comparing the sensitivity of these 2 methods, lasso performs better overall but with a comparatively higher variance, and stepwise has a relatively poor performance in aspect of capturing weak predictors in all settings, but the performance stabilises at 0.1 to 0.25.

Correspondingly, the specificity (see Fig 4.) for stepwise is relatively higher since there is an overall trend for over-screening, but the specificity performance for lasso is also acceptable especially when the predictor-to-sample-size ratio decreases.

Finally, comparing the accuracy of these two methods (see Fig 5.) , lasso performs better in every category of predictors in 6 settings. Above all, in terms of variable identification performance, lasso tends to over-select predictors and performs stably in high dimensional scenarios, while stepwise tends to under-select predictors

with a lower overall performance and less accuracy. As compare to the correlation between predictors, there is also a clear trend that a lower correlation ratio of predictors tends to decrease the overall variance of identification performance, this may due to the high correlation between predictors may reduce the overall variance of the sample space and leads a more stable selection results.

6.2 Parameter Estimation

- bias

$$\frac{1}{p} \sum_{i=1}^p (\hat{\beta} - \beta)$$

- variance

$$\frac{1}{p} \sum_{i=1}^p (\hat{\beta} - \bar{\beta})^2$$

- MSE

$$\frac{1}{p} \sum_{i=1}^p (\hat{\beta} - \beta)^2$$

1. bias
2. rmse
3. comparision of missing vs no missing

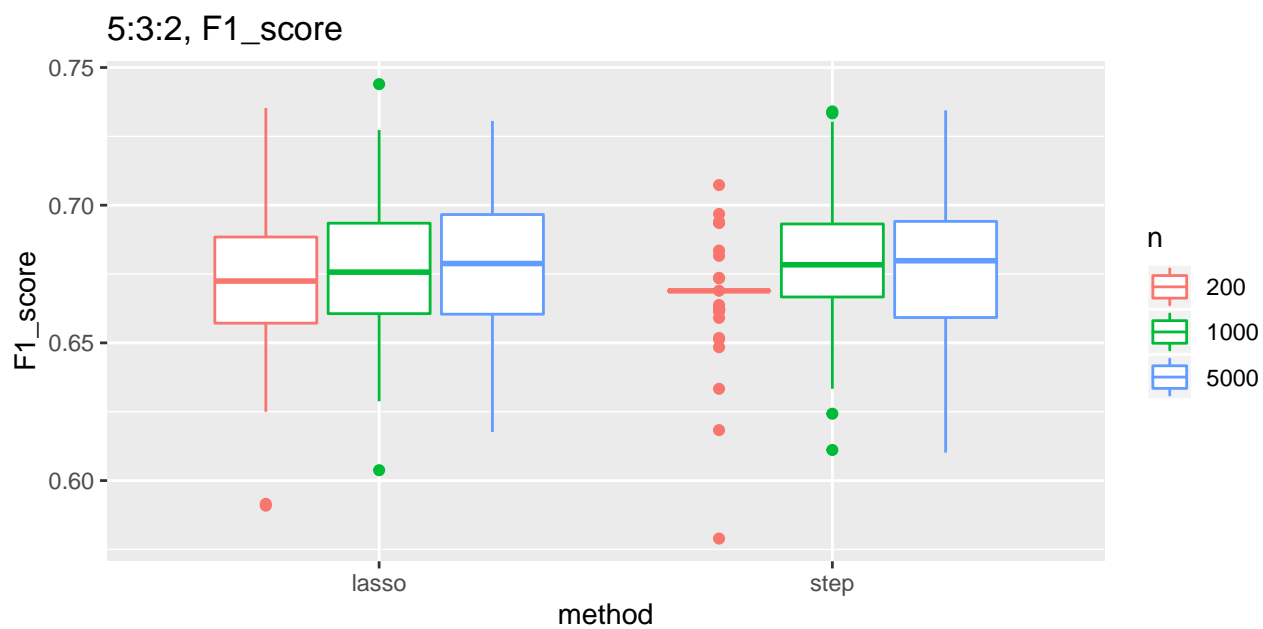
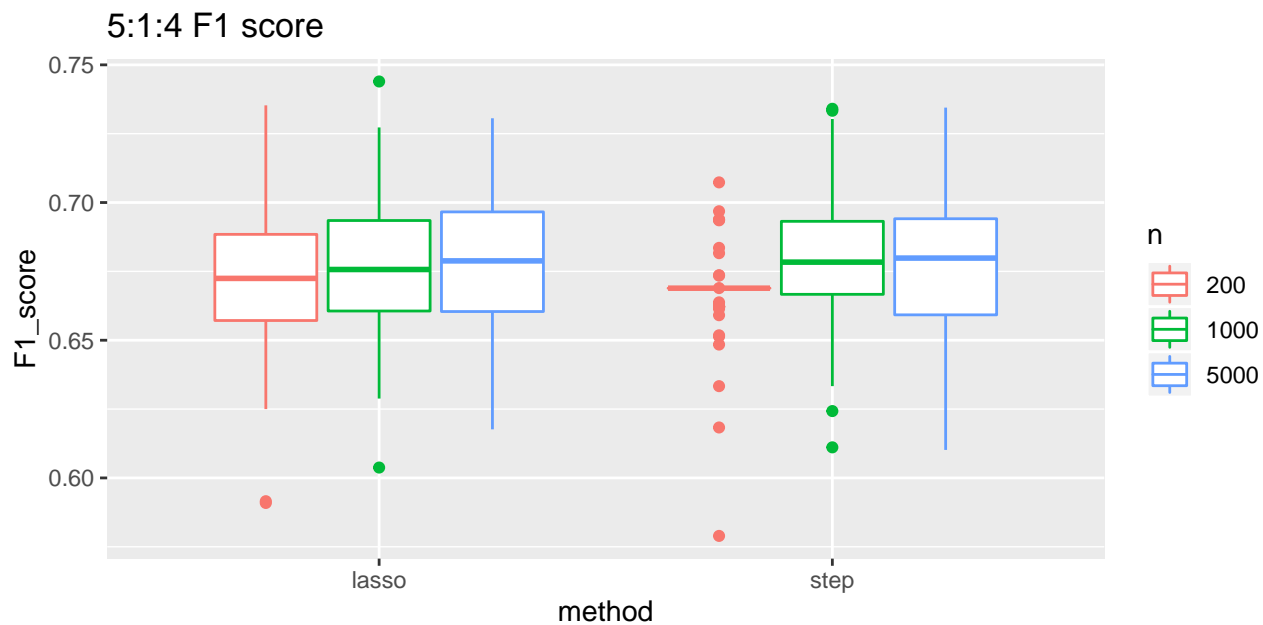


Figure 1: F1 score for 2 variable selection methods in 6 settings

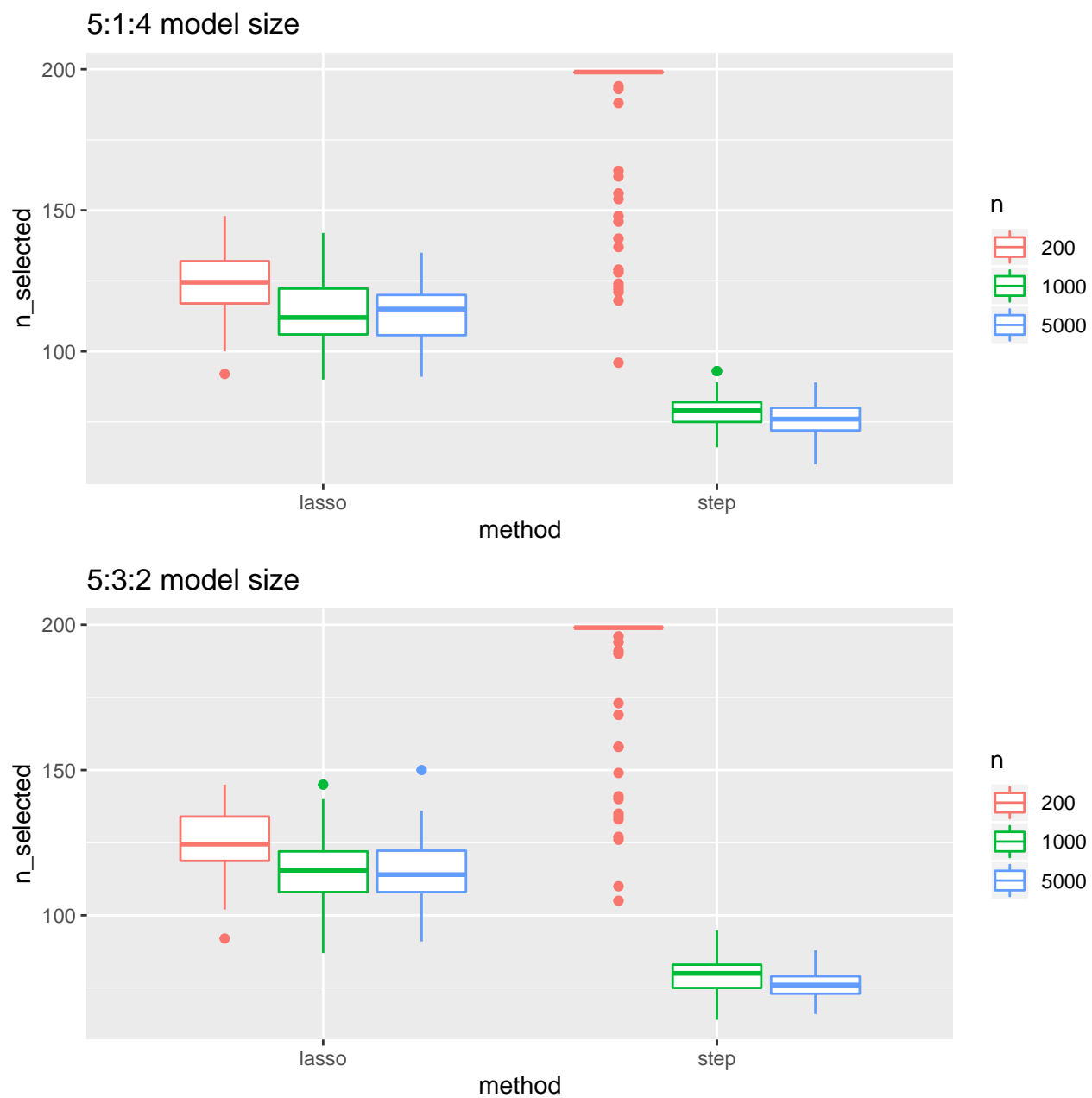


Figure 2: Model size for 2 variable selection methods in 6 settings

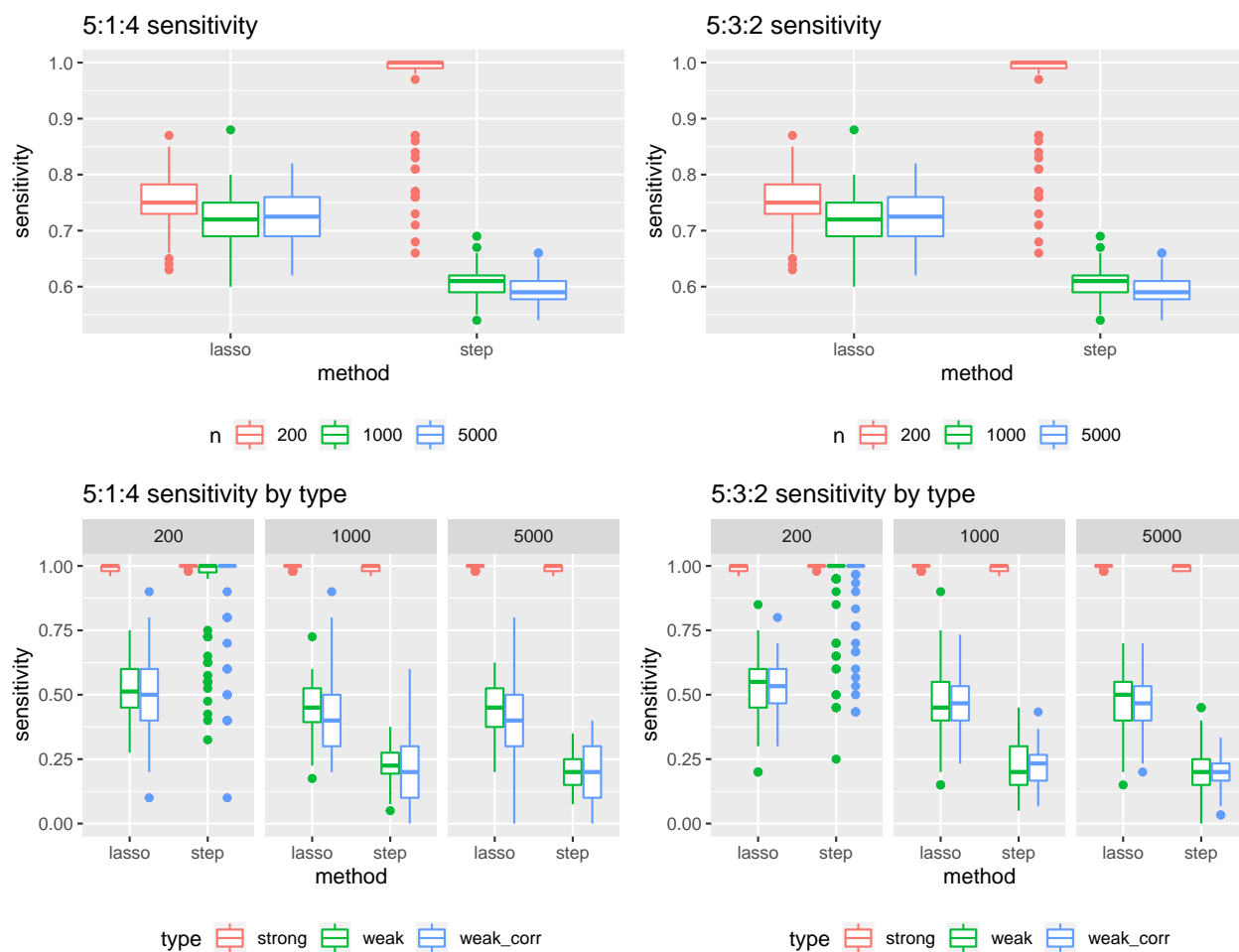


Figure 3: Sensitivity for 2 variable selection methods in 6 settings

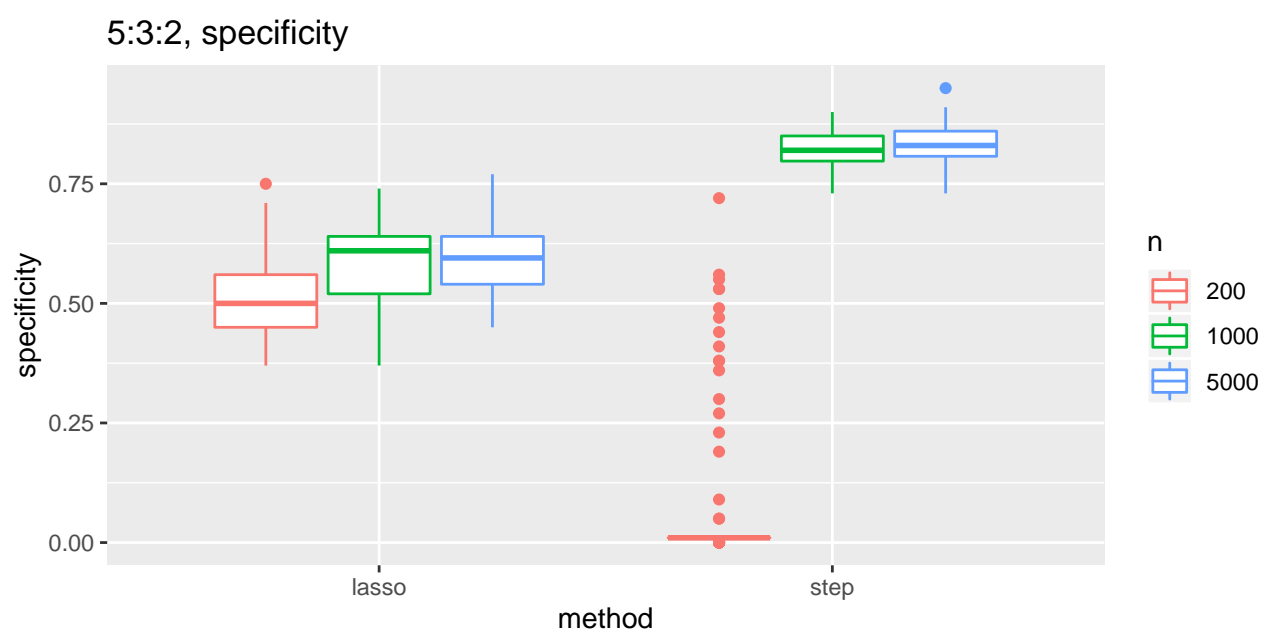
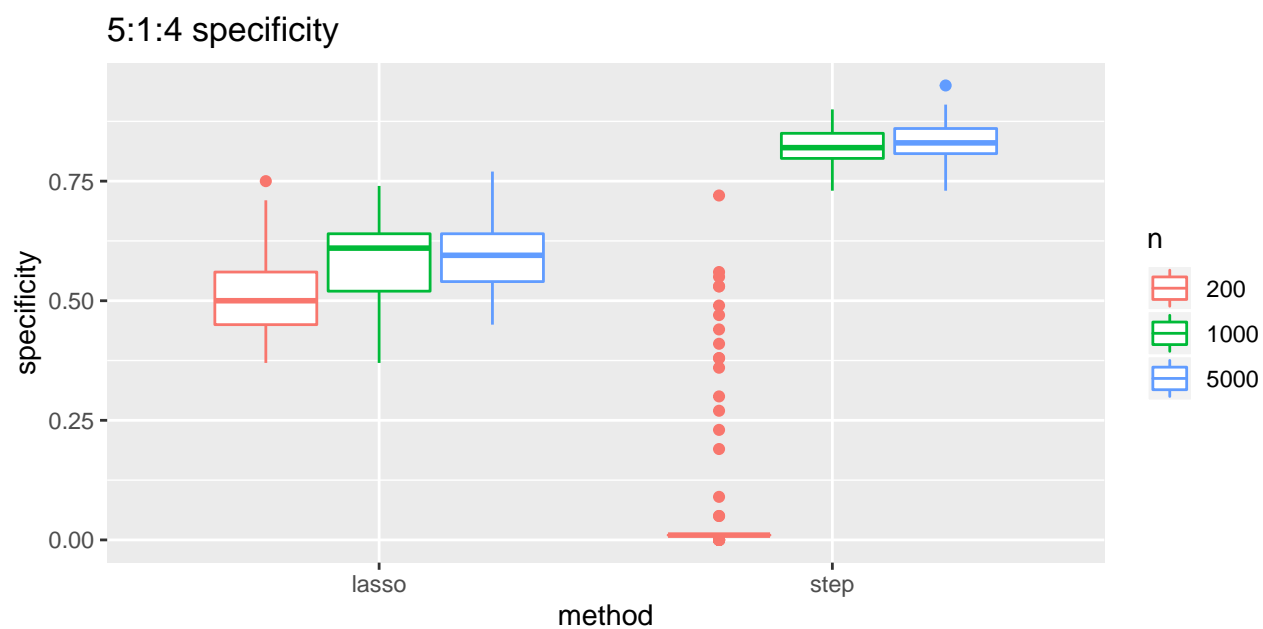


Figure 4: Specificity for 2 variable selection methods in 6 settings

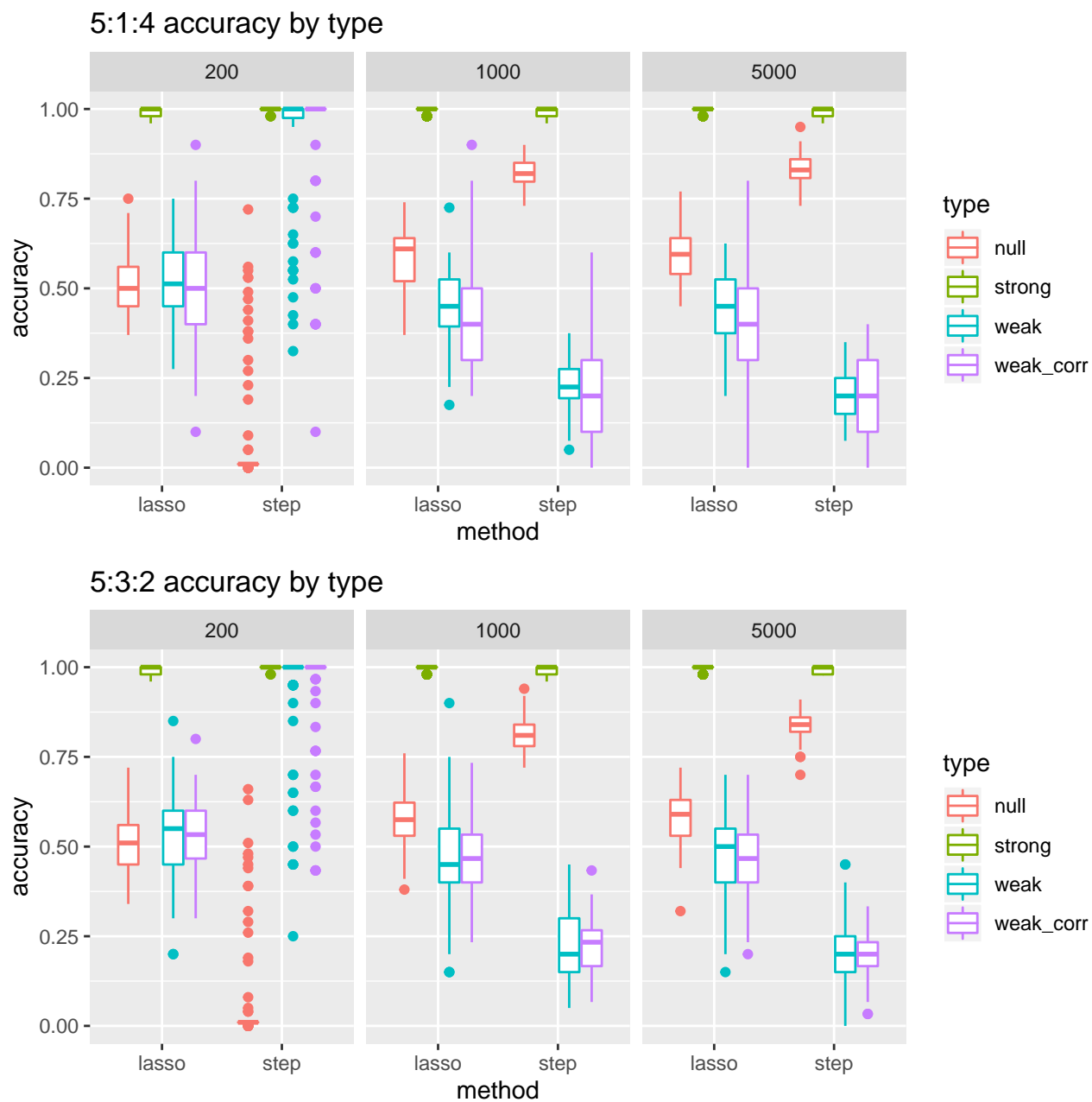


Figure 5: Specificity for 2 variable selection methods in 6 settings