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,

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

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).

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 tunning parameter. Cross-validation (CV) is the most common selection criteria 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 funcion 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 strenghth

The strength of parameters are defined by following criterias:

Definition of strong predictors:

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

Definition of weak but correlated predictors:

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

Definition of weak and independent predictors:

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

Definition of noise:

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

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:

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

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

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

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 paramater 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 is used as an indicater for an overall performance assessment for identification signals. There is no significant difference between this two methods when sample size is large. While in high dimensional occasion, stepwise has a poor performance.

In terms of variable selection size, overal, lasso tend to choose more predictors while stepwise tend to choose less overall, especially in the settings in this report, where there are 100 true signals and 100 noises, lasso tend to select more predictors than true occasion, while stepwise tends to choose less. When comparing in terms of sample size, lasso performs stable in both high dimensional and normal scenario(where n = p), but stepwise tend to break when the ratio between number of parameters and sample size is large.

```
## Warning: Removed 600 rows containing non-finite values (stat_boxplot).

## Warning: Removed 600 rows containing non-finite values (stat_boxplot).

## Warning: Removed 600 rows containing non-finite values (stat_boxplot).
```

When comparing the specificity of these 2 methods, lasso performs better overall since more predictors are chosen, when observing sensitivity in terms of predictors strength, both methods perform well for strong predictors when sample size is large enough, while for weak predictors, lasso still has a better performance

6.2 Parameter Estimation * bias

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

- variance

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

- MSE

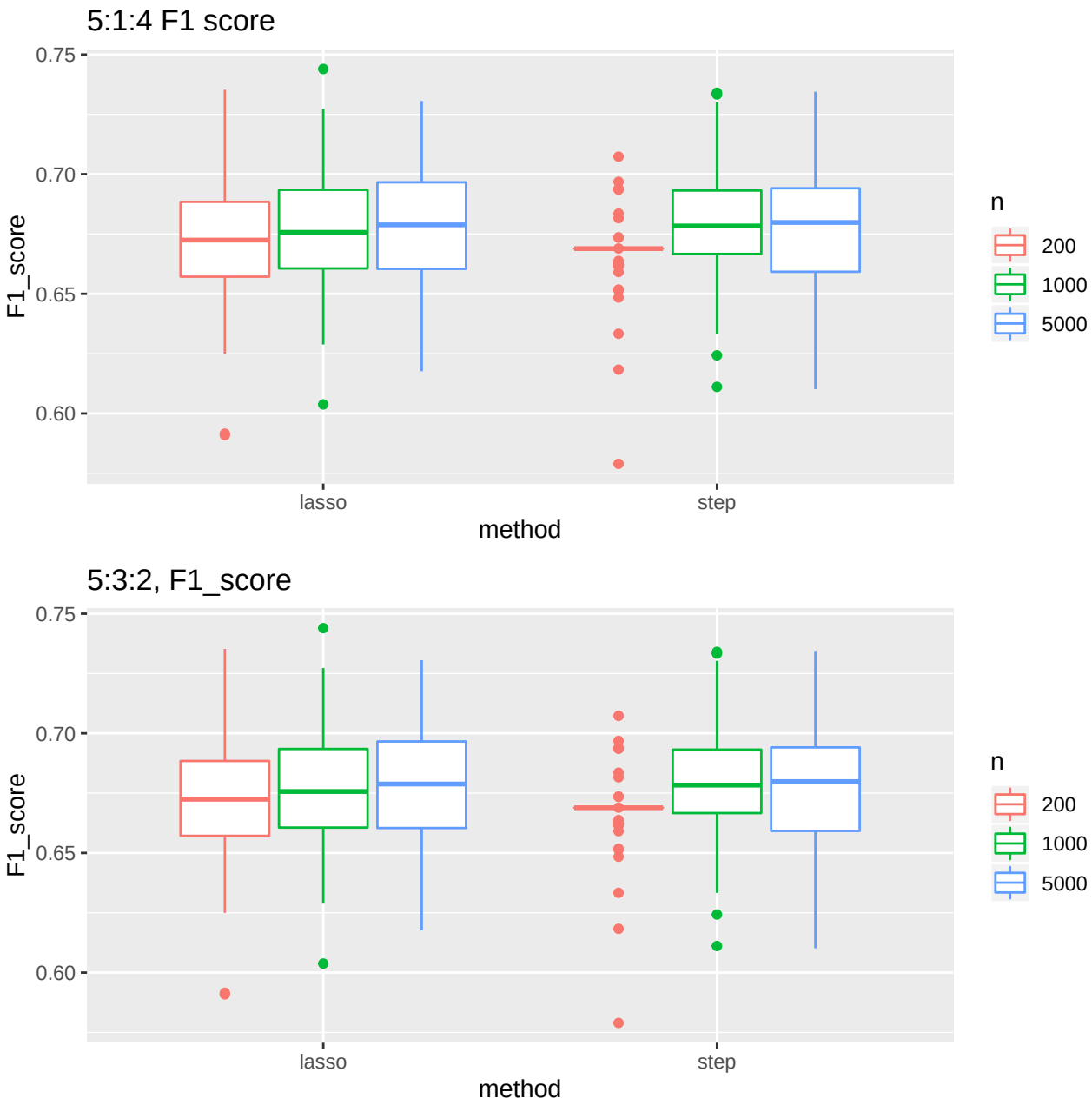

Figure 1: F1 score for 2 variable selection methods in 6 settings

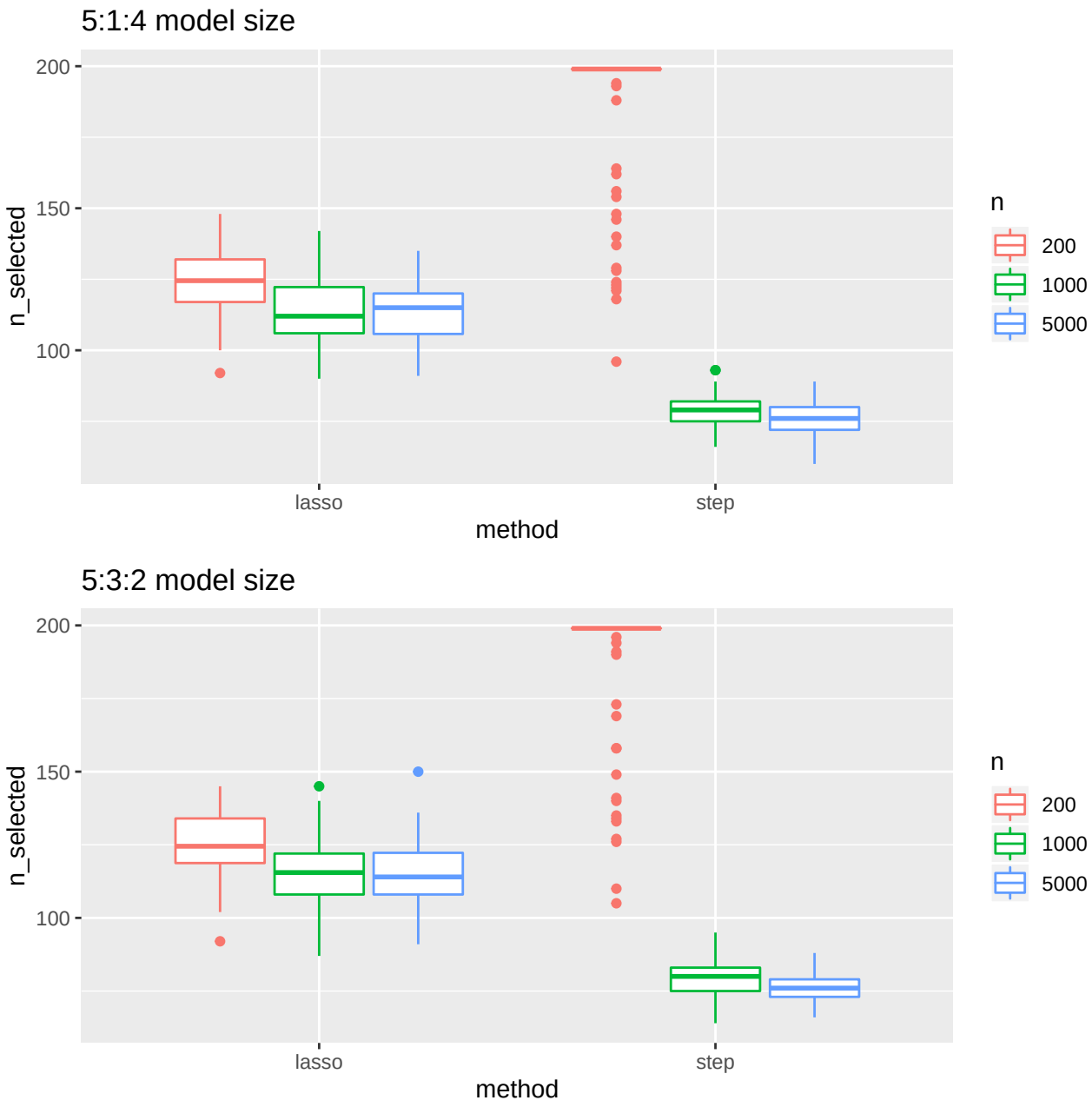

Figure 2: Model size for 2 variable selection methods in 6 settings

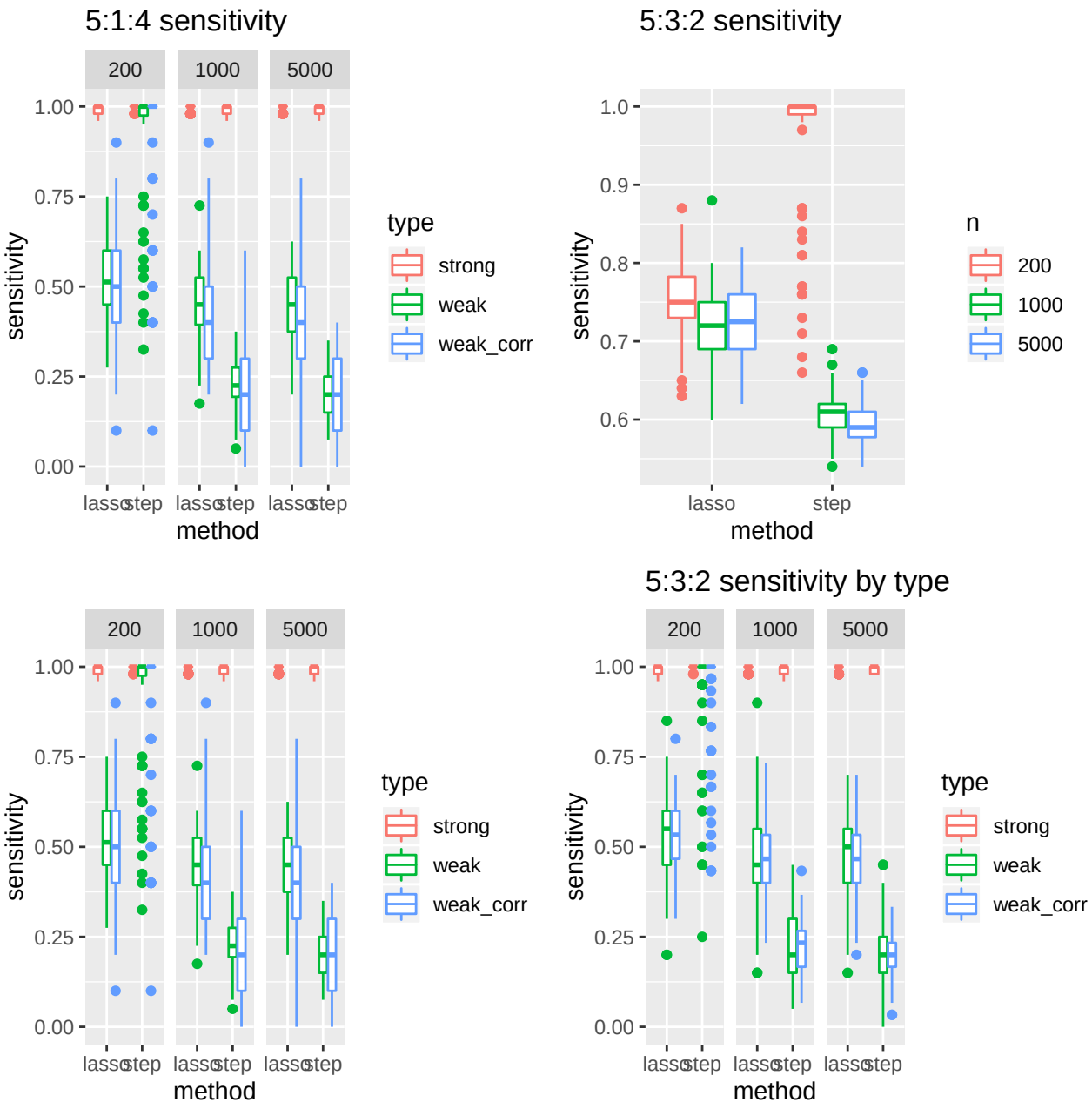

Figure 3: Sensitivity for 2 variable selection methods in 6 settings

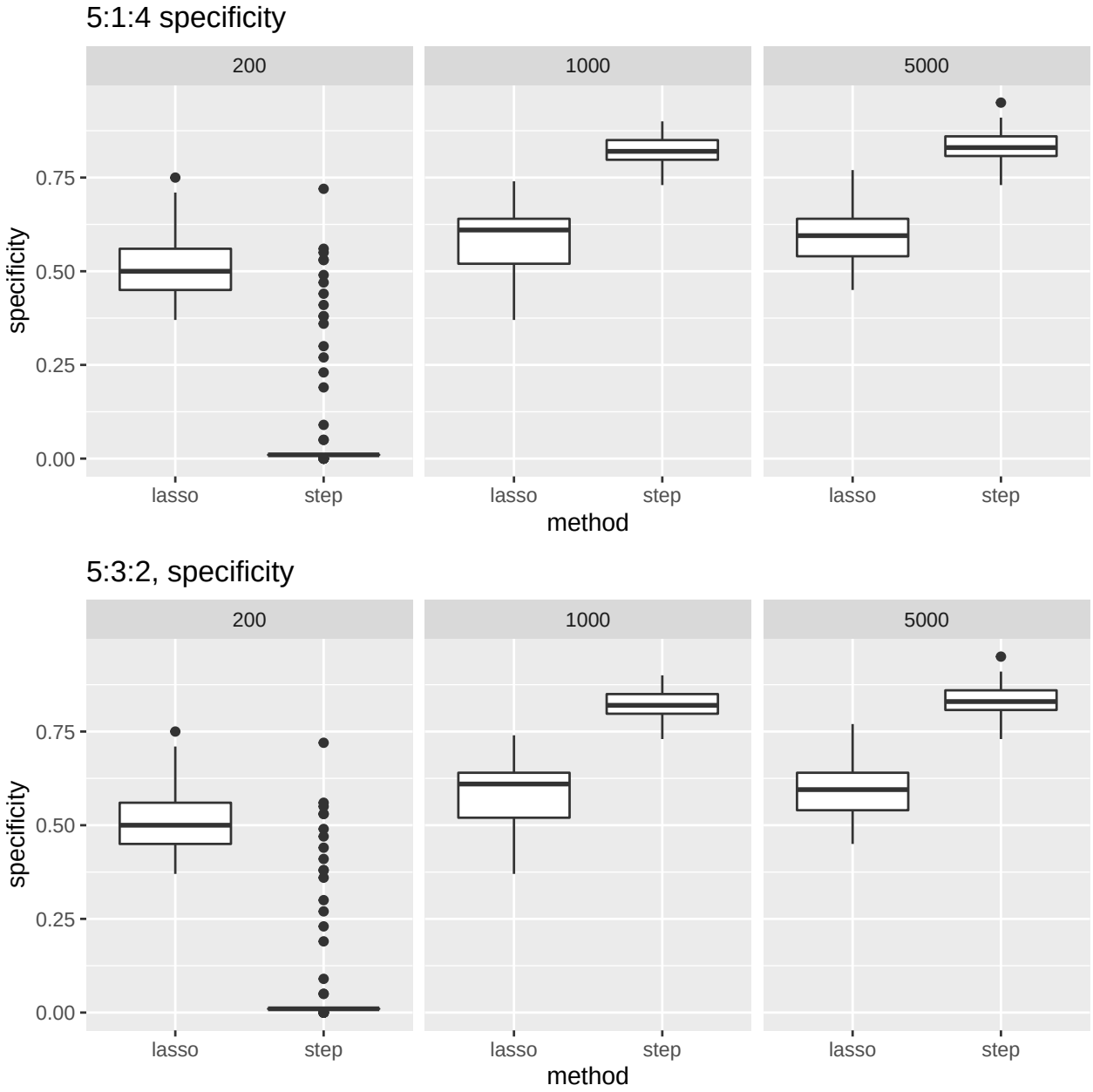

Figure 4: Specificity for 2 variable selection methods in 6 settings

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

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