Project 1-Topic 3 Report

Hongjie Liu, Jiajun Tao, Shaohan Chen

02/27/2023

Background

When dealing with high-dimensional data

Statistical methods to be studied

Stepwise Forward Selection

Step-wise forward selection

LASSO

LASSO regression

Objective

Several parameters

Simulation

Data is generated

Experiment Settings and Scenario

Model Evaluation

Evaluation Metrics

In this project, we define the true predictors as positive and null predictors as negative.

For signal identification, we use the following five metrics to compare the two models:

- Complexity: The number of selected predictors in the model
- Sensitivity: $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FP}$
- F1-score: $\frac{2 \cdot sensitivity \cdot precision}{sensitivity+precision}$
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

For parameter estimation, we use the following to metrics to compare the two models:

- RMSE: $\sqrt{\frac{1}{p}\sum_{i=1}^{p}(\hat{\beta}_i - \beta_i)^2}$
- Variance: $\sqrt{\frac{1}{p}\sum_{i=1}^{p}(\hat{\beta}_i - \bar{\beta}_i)^2}$

Signal Indentification Performance

Complexity of the models is indicated by the number of selected predictors. We can see that in high dimensional scenario(when n=100), Forward selection model tends to select lots of predictors and Lasso tends to select very few. One interesting thing is that, if we increase the ratio of strong predictors(i.e. more strong predictors), Lasso also tends to select more predictors too. When it comes to normal scenario, Forward selection still tends to select more predictors than Lasso, but the discrepancy is smaller than high dimensional case, and will be further narrowed down with n or the ratio of strong predictors increasing. And as n increases, the number of selected parameters of both models are closer to the true number 40.

As for overall classification performance, if in high dimensional scenario, Forward selection tends to be very assertive and much better at identifying weak signals, leading to an extremely high sensitivity but low specificity. Lasso in turn tends to be very conservative and much better at identifying null signals, leading to an extremely high specificity but low sensitivity. Like high dimensional case, Lasso becomes more sensitive and not that assertive when ratio of strong predictors increases. Based on the above plot, we can conclude that both models does not perform too well based on F1-score and accuracy, because they are very radical and tend to identify most of the predictors either as positive or negative, but far away from the truth.

For normal scenario, both models become less radical under normal scenarios, but Forward selection is still more sensitive and assertive than Lasso, while Lasso has higher specificity and more conservative. Both models perform better on those metrics with n increasing. Overall, Lasso and Forward selection has similar F1-score and accuracy performance. But when there are more strong predictors, Lasso performs obviously better than Forward selection.

About the classification performance of different signals. Under all n values, both models perfectly identify the strong signals. In high dimensional scenario, Forward selection performs much better on identifying weak predictors while Lasso performs much better on identifying null predictors, that' s why we see the high sensitivity of forward selection and high specificity of Lasso in previous section. When there are more strong predictors, Lasso also performs better on selecting weak predictors. When it comes to normal scenario, Forward selection is still better at selecting null and Lasso is better at selecting weak predictors. But the discrepancy is smaller compared with high-dimensional data, and will continue be smaller as n increases. When there are more strong predictors, Lasso performs much better at selecting weak-but-correlated signals.

Parameter Estimatio Performance

In high dimensional scenario, Lasso performs much better than forward selection, with obvious much lower and centered RMSE and also lower variance. When it comes to normal scenario. Though when n=500, Lasso outperforms forward selection, but with n increasing, forward selection starts to outperform Lasso model on RMSE and variance. And overall, Lasso tends to perform better when there are more strong signals. If there are more strong predictors, the variance is also larger.

Effect of Missing Weak Predictors

Here we define missing weak predictors as the true weak predictors but we estimated them as null predictors, and we used RMSE to evaluate. In each scenario, we picked up three kinds of missing situations i.e. most missing, least missing, and middle missing. As the name suggests, the most missing are the ones that have the least non-null estimations. Foe each kind of situation, we picked 10% to draw the plot. For example, we fixed 100 parameters and 100 simulation times. In each simulation, we can get the number of non-null parameters. After arranging them, we can pick the top 10 simulation times that have the most non-null parameters. In the same way, we can pick the last 10 and the middle 10 which ranks 45 to 55. It is worth mentioning that, Lasso only picks which parameters to use and the coefficients of Lasso can not be used directly. In order to compare RMSE, we need to refit the linear regression model using the parameters that Lasso picks.

When in high dimensional scenario(when n=100), Lasso has a very small RMSE however forward selection's RMSE is big. But when the missing amount increases, the RMSE of forward selection drops dramatically. When the number of missing parameters increases, the RMSEs of both methods decrease.

When under normal scenario(n = 500 or 2000), the RMSEs of both methods are small. It's hard to tell which method is better since their differences are small. There seems to be no apparent patterns between different ratios as well.

Since we care more about high dimensional scenario, the conclusion should be Lasso performs better than forward selection when in high dimensional scenario according to RMSE, and the more missing parameters, the better the RMSE.

Discussions

Limitation

There is much freedom when designing the simulations. In our algorithm, we have 5 parameters, number of observations, number of parameters, the ratio of strong and weak signals, the definition of strong and weak signals and the correlation between strong and weak signals. However, even more parameters can be adjusted such as the correlation between WBC and WAI, or between null and strong, etc. We generated many versions of data and found that many things can affect the result. Here we only fixed p and c and the results and conclusions may not be comprehensive.

Future Work

For the future work, we could adjust other parameters to investigate this problem further. What's more, we reproduced the high dimensional scenario and faced the stuggle of choosing covariates. We still could not have a clear solution to deal with this difficulty. It would be hard to tackle the problem, but it can be a direction of effort.

Reference

1. 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

Appendix

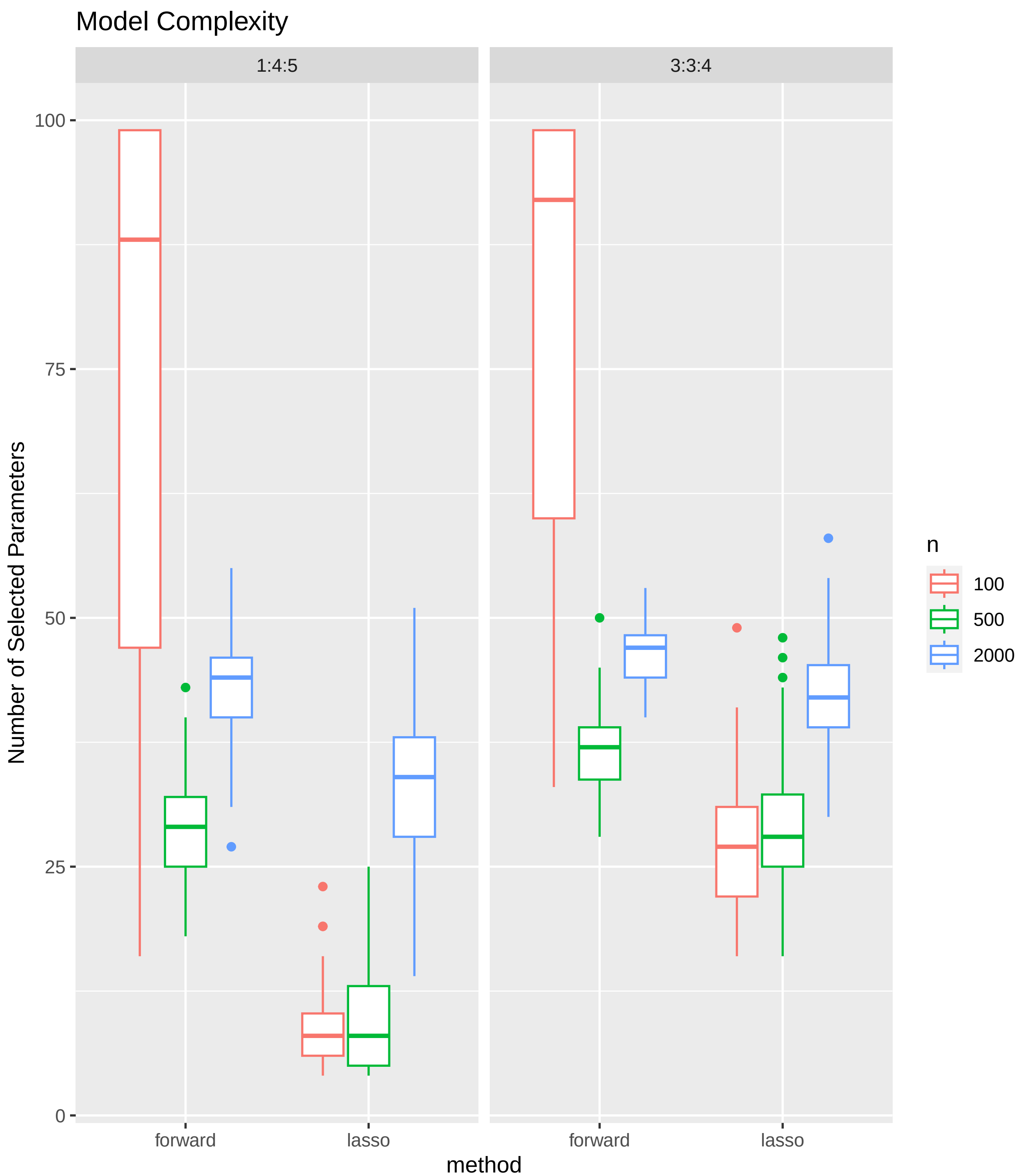

Figure 1: Model Complexity

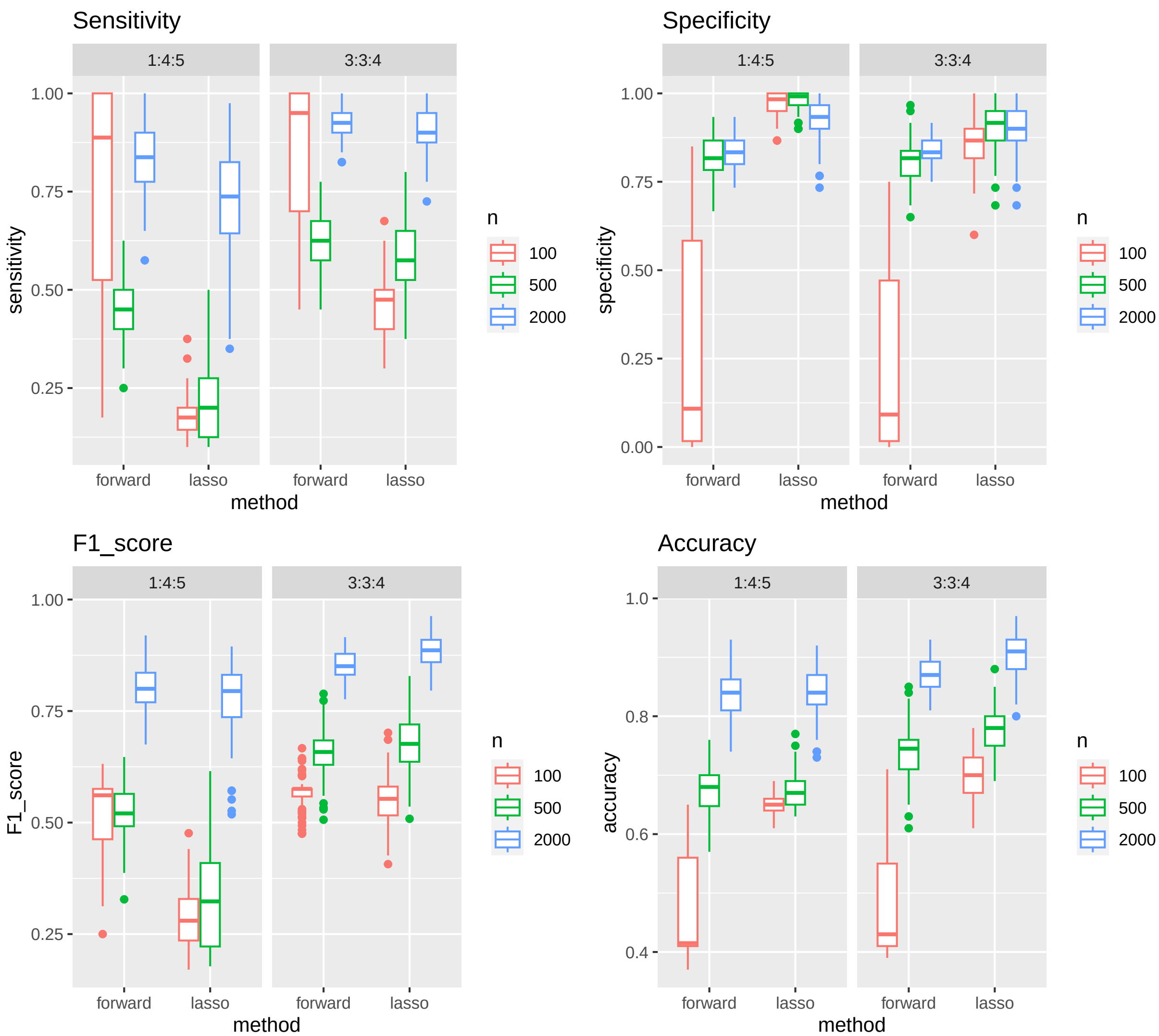

Figure 2: Overall Classification Performance

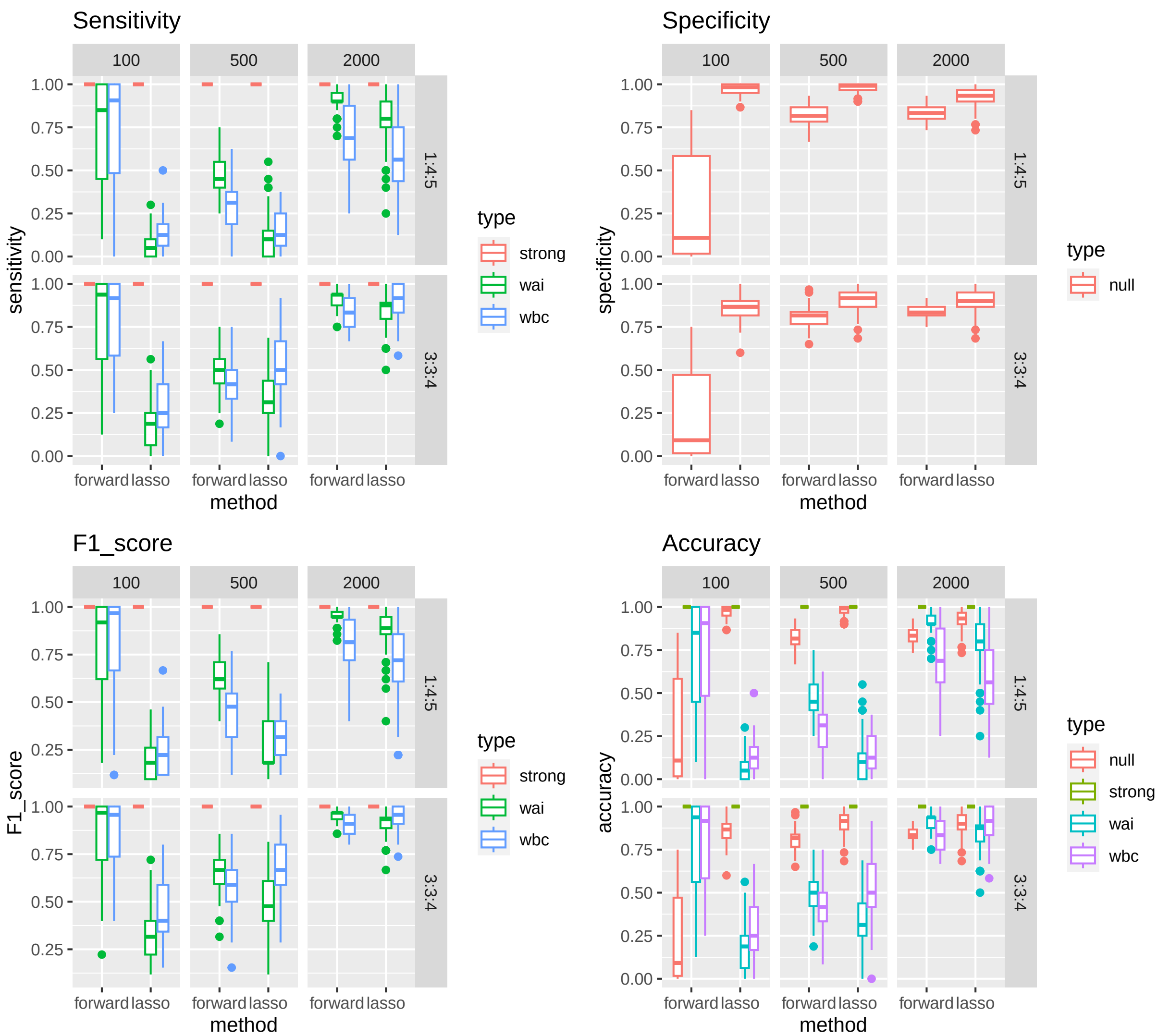

Figure 3: Classification Performance by Signals

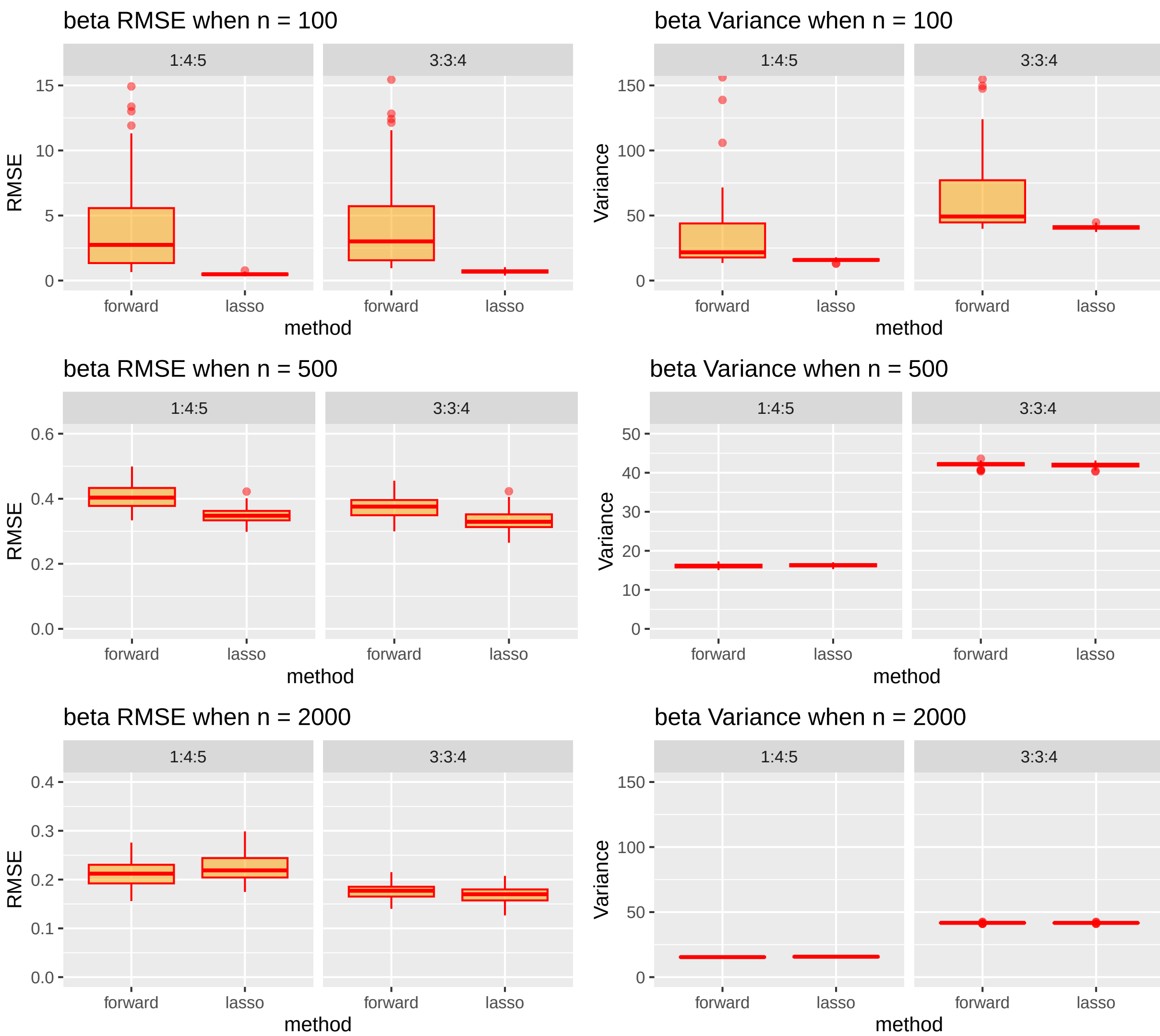

Figure 4: Parameter Estimation Performance

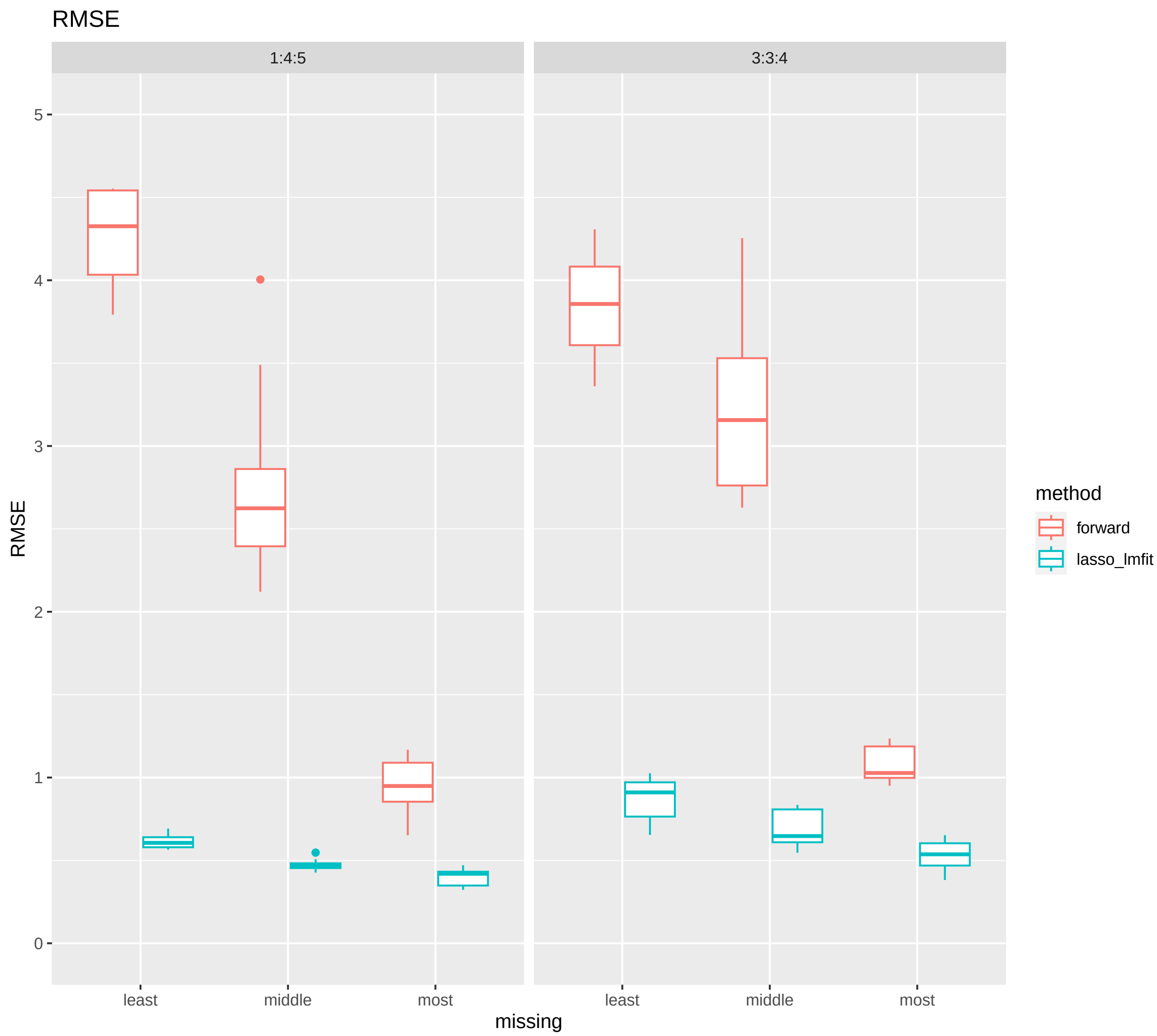

Figure 5: RMSE comparison when n = 100

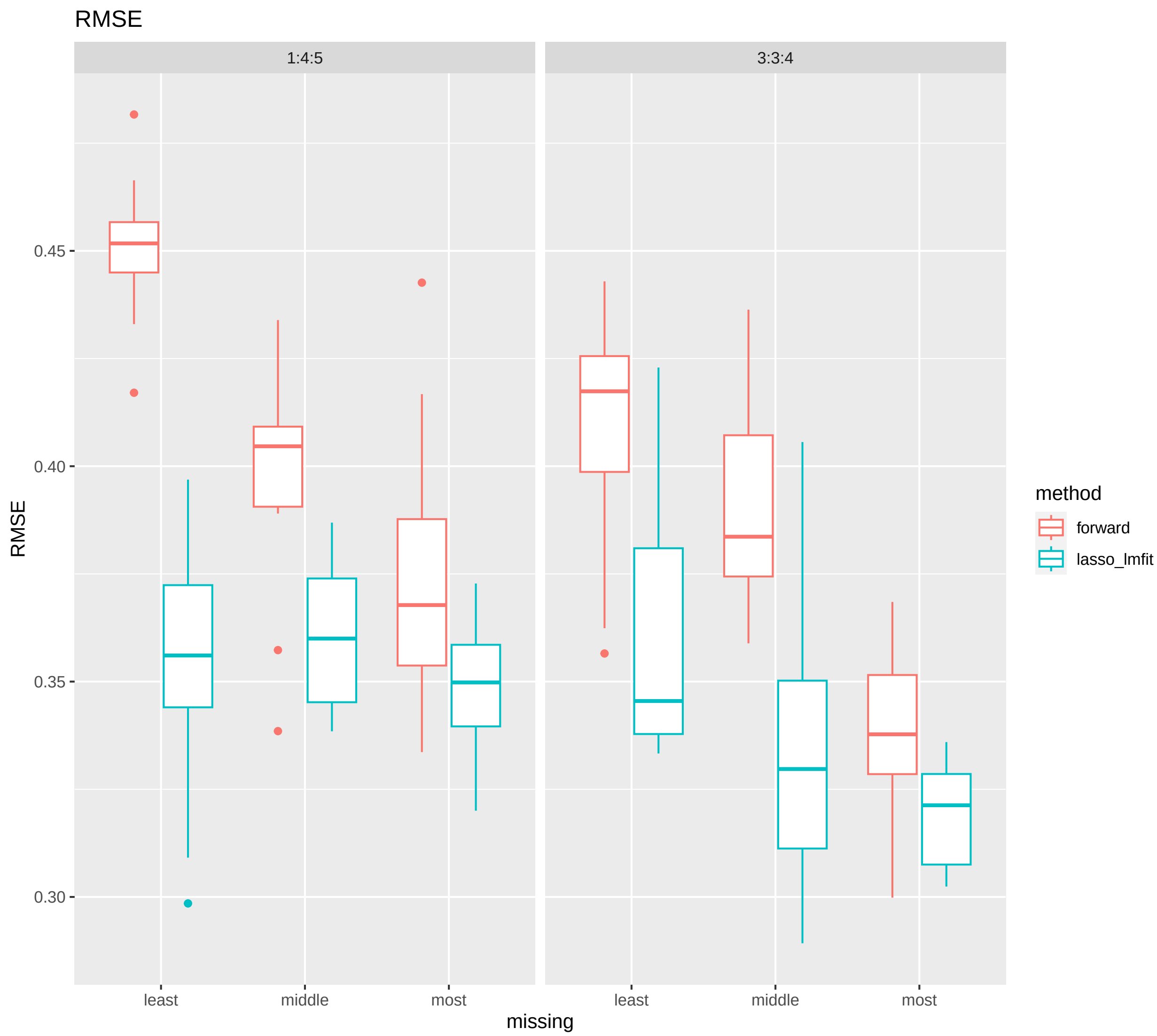

Figure 6: RMSE comparison when n = 500

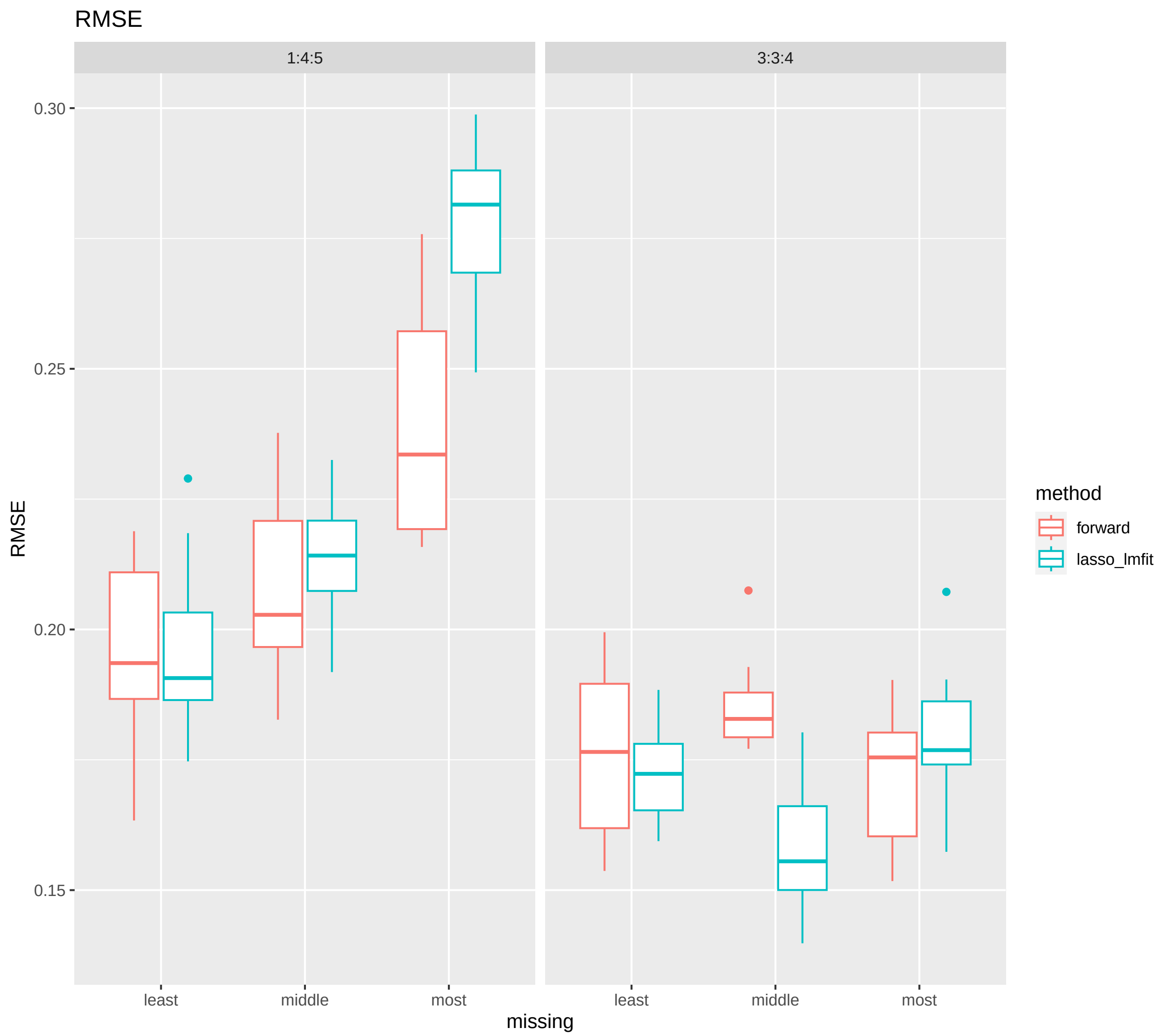

Figure 7: RMSE comparison when n = 2000