Variable Selection Methods Comparison

Hongjie Liu, Jiajun Tao, Shaohan Chen

2023-02-27

Background

- Variable selection methods help to optimize models in high-dimensional settings where we need to select predictors that balance fitness and complexity.
- The presence of weak predictors is a problem that plagues traditional variable selection methods.

Statistical Methods to be Studied

Step-wise forward method

- Starts 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,

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

Automated LASSO regression

- Estimates model parameters by optimizing a penalized loss function:

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

Objectives

(1) Evaluate the effectiveness of both methods in identifying weak and strong predictors.

(2) Examine how the absence of "weak" predictors affects parameter estimations.

Types of Signals

- Strong signals

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

- Weak-but-correlated (WBC) signals

$$S_{WBC} = \{j : 0 < |\beta_j| \le c\sqrt{\log(p)/n} \text{ and } \mathrm{corr}(X_j, X_j') \neq 0 \text{ for some } j' \in S_1,\ 1 \le j \le p\}$$

- Weak-and-independent (WAI) signals

$$S_{WAI} = \{j : 0 < |\beta_j| \le c\sqrt{\log(p)/n} \text{ and } \mathrm{corr}(X_j, X_j') = 0 \text{ for all } j' \in S_1,\ 1 \le j \le p\}$$

- Null signals: $S_{null} = \{j : \beta_j = 0,\ 1 \le j \le p\}$

Types of Signals

Thus, p predictors can be partitioned as

$$\{1, \cdots, p\} = S_{strong} \cup S_{WBC} \cup S_{WAI} \cup S_{null}.$$

- We assume that $|S_{strong}| = p_S$, $|S_{WBC}| = p_{WBC}$, $|S_{WAI}| = p_{WAI}$.
- The number of true predictors $p_S + p_{WBC} + p_{WAI}$ should be less than n.

Data Generation

- Normality assumption

$$\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

- For $j \in S_{strong}$, $\beta_j = 20$; For $j \in S_{WBC} \cup S_{WAI}$, $\beta_j = 0.5$.
- We choose $c = 20$ so that $0.5 \leq c\sqrt{\log p/n} < 20$ for all scenarios to be investigated.
- $\sigma^2 = 8$.

Data Generation - Random Number Generation (Binary Outcome)

```
set.seed(20220217)
seed_vec <- runif(100000, min, max)
  for (i in 1:n) {
    set.seed(seeds[i])
    long_rnorm <- rnorm(size*3, mean = 0, sd = 1)
    long_runif <- runif(size*2)
    beta_error <- rnorm(size, mean = 0, sd = 0.25)
    L1 <- long_rnorm[1:size]
    L2 <- long_rnorm[(size + 1):(2*size)]
    L3 <- long_rnorm[(2*size + 1):(3*size)]

    comp_pA = long_runif[1:size]
    A = (prob_A > comp_pA)
    # function continues...
    }
```

Parameters of Interest

- The sample size of each dataset $n_{\text{sample}} \in \{100, 1000\}$
- The population proportion of treated individuals $\pi \in \{0.113, 0.216, 0.313\}$
- The true average treatment effect $\beta_1 \in \{0.15, 0.30\}$ for binary data; $\beta_1 \in \{-1, 1\}$ for continuous data

Other Parameters

- The number of datasets $m_{\text{sample}} = 100$
- The number of bootstrap re-samples $m_{\text{boot}} = 500$
- The sample size of bootstrap re-samples $n_{\text{simple}} = n_{\text{complex}} = n_{\text{sample}} \times \pi$
- Strength of covariate effect on treatment $\alpha_1 = \log(1.25), \alpha_2 = \log(1.75)$
- Strength of covariate effect on outcome $\beta_2 = \log(1.75), \beta_3 = \log(1.25)$

Evaluation Metrics

Define true predictors as positive and null predictors as negative

Signal Identification

- **Complexity:** Number of Selected Parameters
- **Sensitivity:** $\frac{TP}{TP+FN}$
- **Specificity:** $\frac{TN}{TN+FP}$
- **F1-Score:** $\frac{2 \cdot \text{sensitivity} \cdot \text{precision}}{\text{sensitivity} + \text{precision}}$
- **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$

Parameter Estimation

- **MSE:** $\frac{1}{p}\sum_{i=1}^{p}(\hat{\beta}_i - \beta_i)^2$

Signal Identification Performance - Complexity

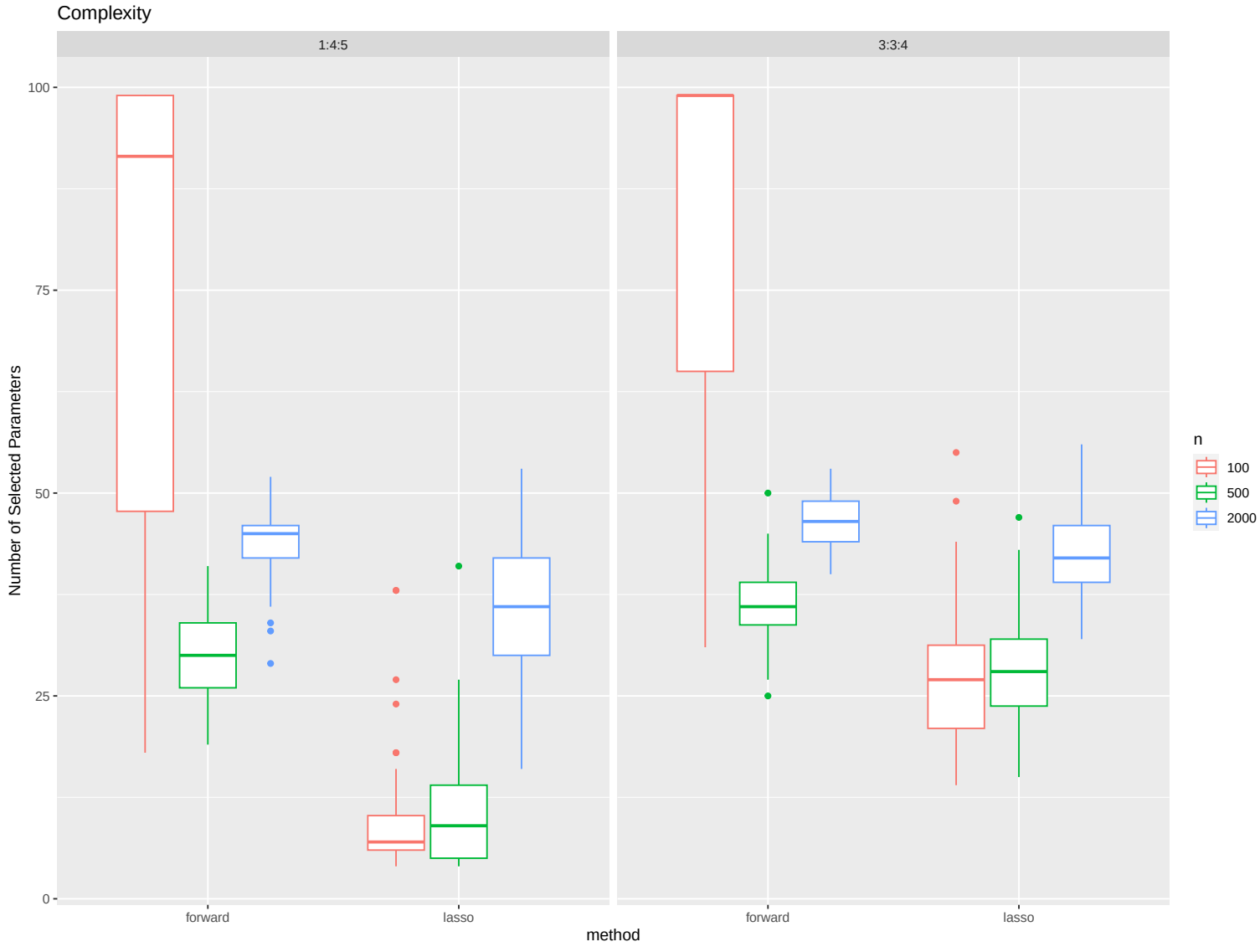

Figure 1: Model Complexity

Signal Identification Performance - Complexity Exploration

- When in high dimensional scenario, forward selection tends to select nearly all of the predictors but Lasso does not
- Lasso tends to select much fewer predictors than forward selection, and the parameters it select will increase as n increases
- As n increases, the predictors that two models select are more precise(closer to true predictor number 40)

Signal Identification Performance - Sensitivity

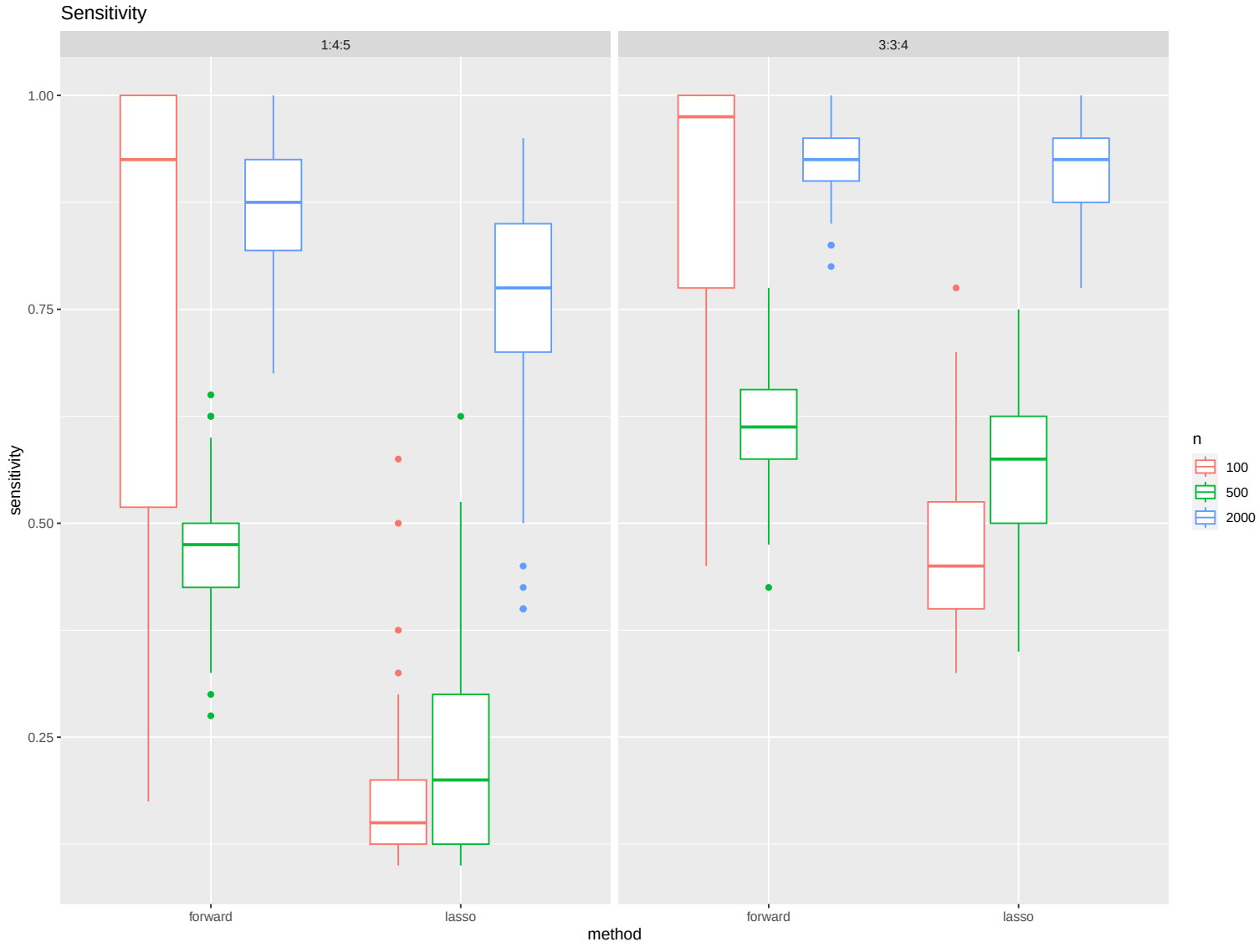

Figure 2: Sensitivity Performance

Signal Identification Performance - Sensitivity

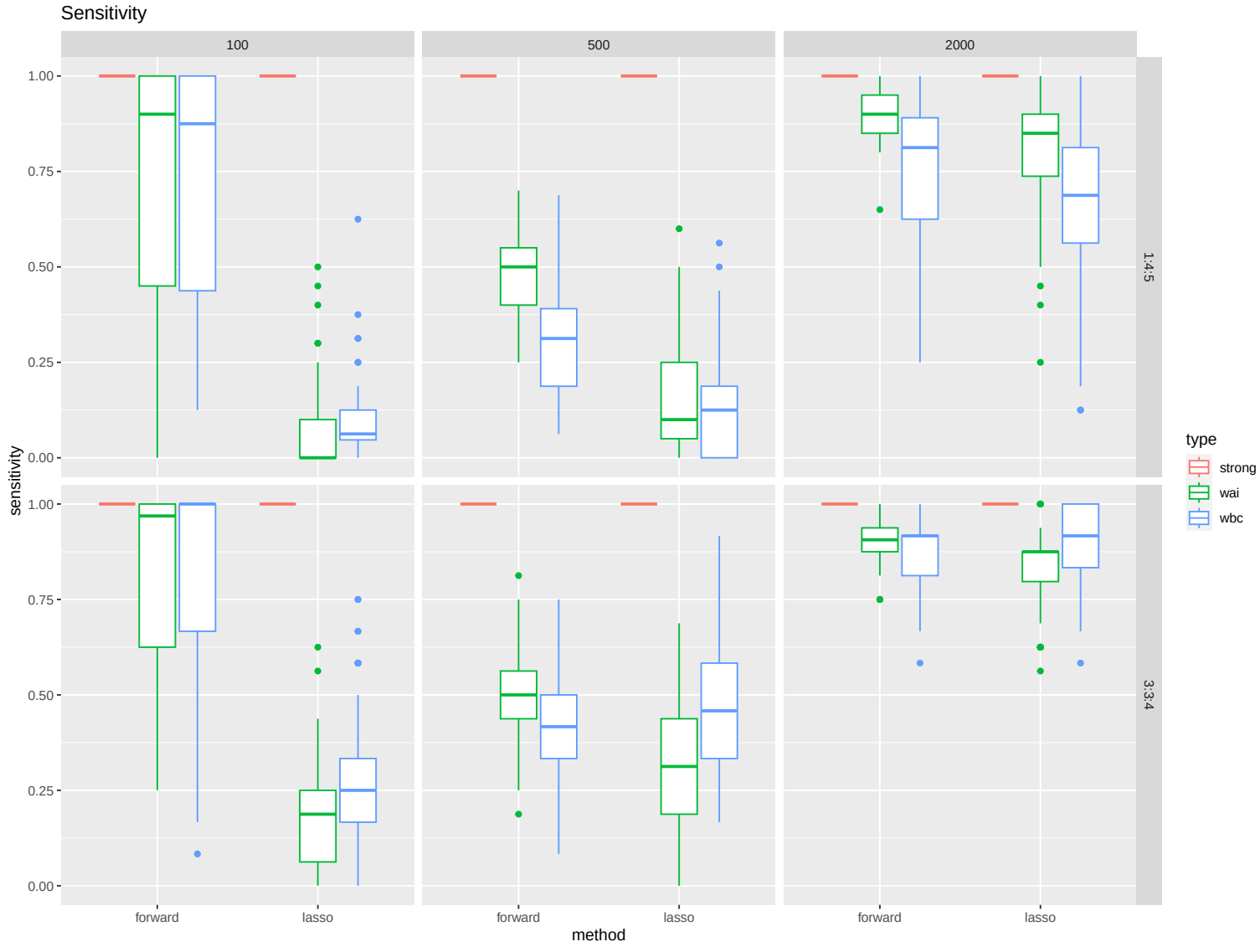

Figure 3: Sensitivity of True Signals

Signal Identification Performance - Sensitivity Exploration

- Forward selection are highly sensitive in high dimensional case(n=100) while Lasso does not
- Overall, the sensitivity of two models increases as n increases
- Both models are sensitive in selecting strong signals. But when it comes to weak predictors, Lasso is much less sensitive in high dimensional scenario than forward selection.
- When n increases, the sensitivity discrepancy of selecting weak predictors between two models becomes smaller. But still, forward selection is overall more sensitive than Lasso
- The sensitivity discrepancy between two models are smaller when the ratio of strong predictors becomes larger

Signal Identification Performance - Specificity

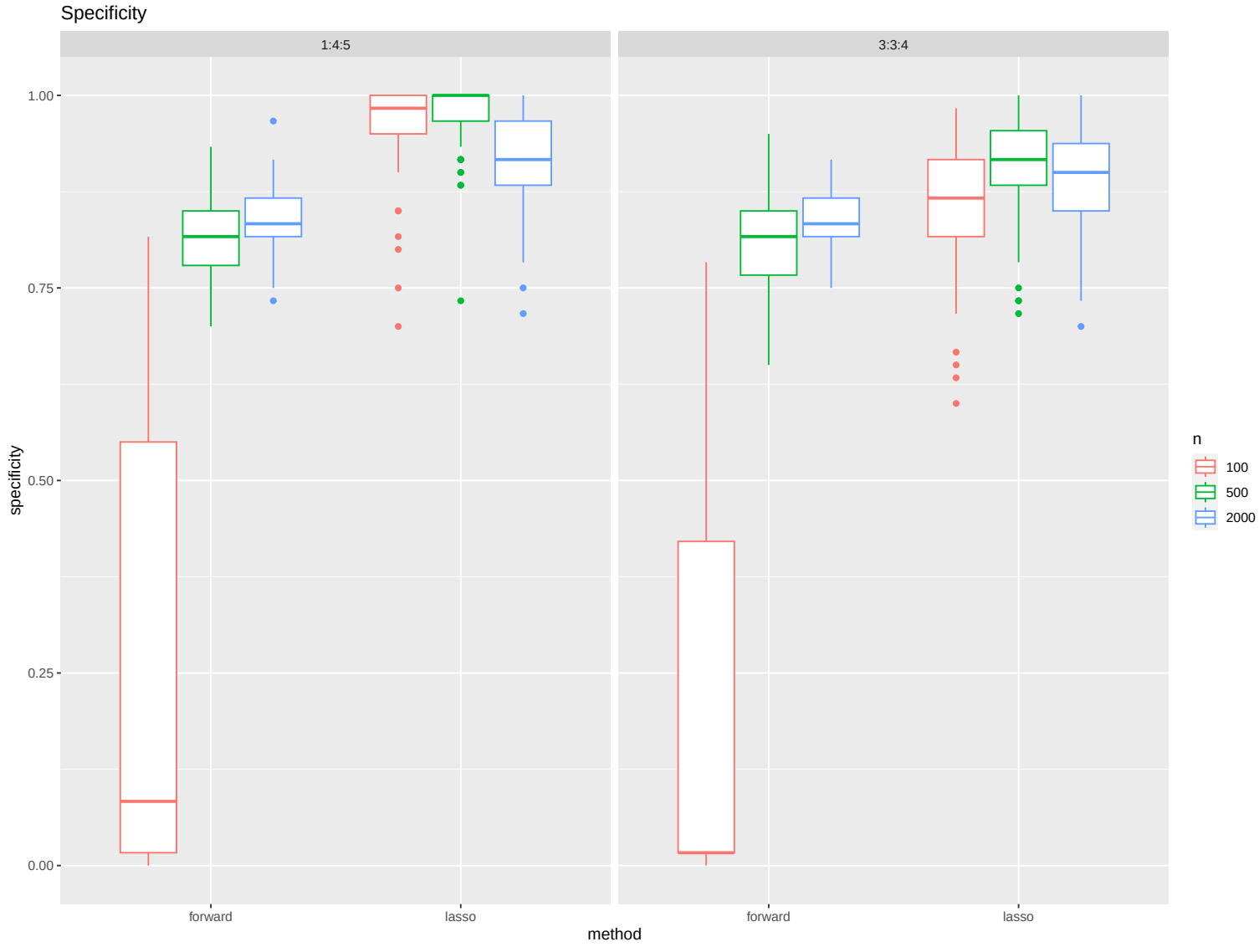

Figure 4: Specificity Performance

Signal Identification Performance - Specificity Exploration

- In high dimensional scenario, the specificity of forward selection is near 0, which means it almost does not identify any null predictor, but Lasso in turn has high specificity
- In high dimensional scenario, forward selection is very assertive and tends to identity all 100 predictors as true, leading to extremely high sensitivity but low specificity
- In high dimensional scenario, Lasso is very conservative and tends to identity most 100 predictors as null,leading to low sensitivity but high specificity
- As n increases, forward selection has higher specificity. And overall, specificity are higher when the ratio of strong predictors are lower

Signal Identification Performance - F1-Score

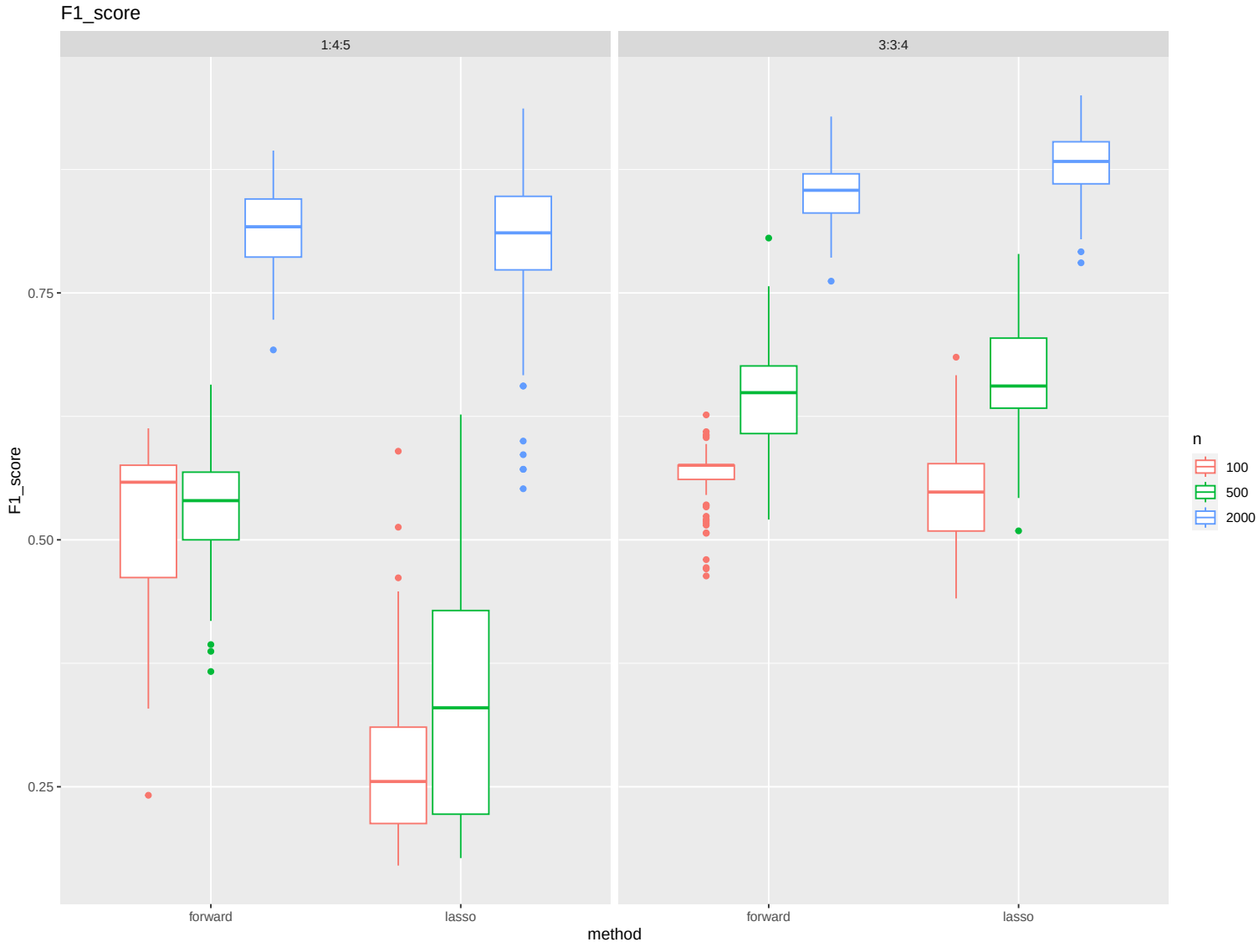

Figure 5: F1-Score Performance

Signal Identification Performance - F1-Score

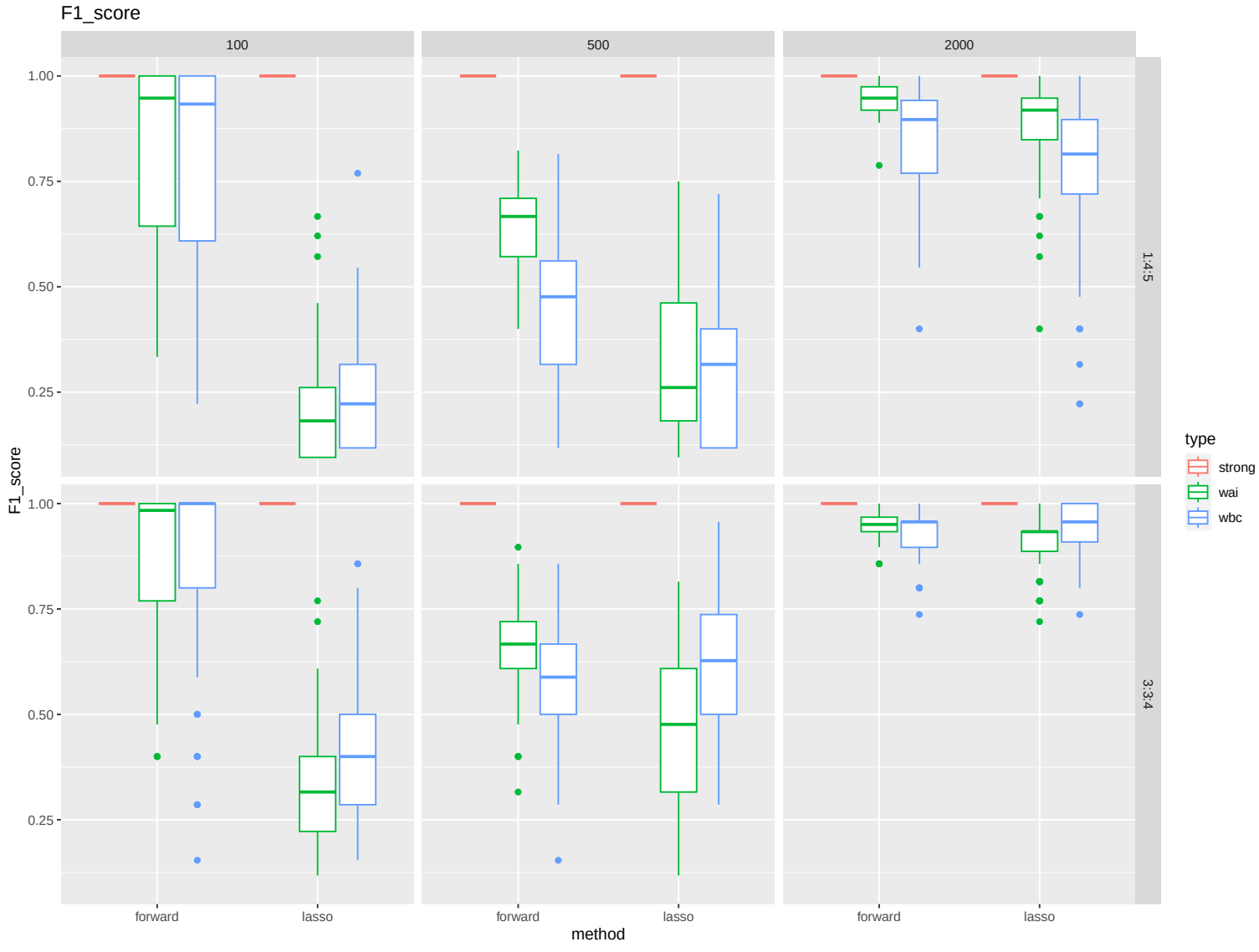

Figure 6: F1-Score of True Signals

Signal Identification Performance - F1-score Exploration

- Lasso has lower F1-Score when n is not large. When n=2000, both models have similarly high F1-score
- F1-score will increase significantly for both models, when the ratio of strong predictors are larger. And F1-score is also higher when the ratio of strong predictors are larger
- Strong predictors have F1-Score=1 for each scenario, and weak predictors have higher F1-score when n increases for Lasso

Signal Identification Performance - Accuracy

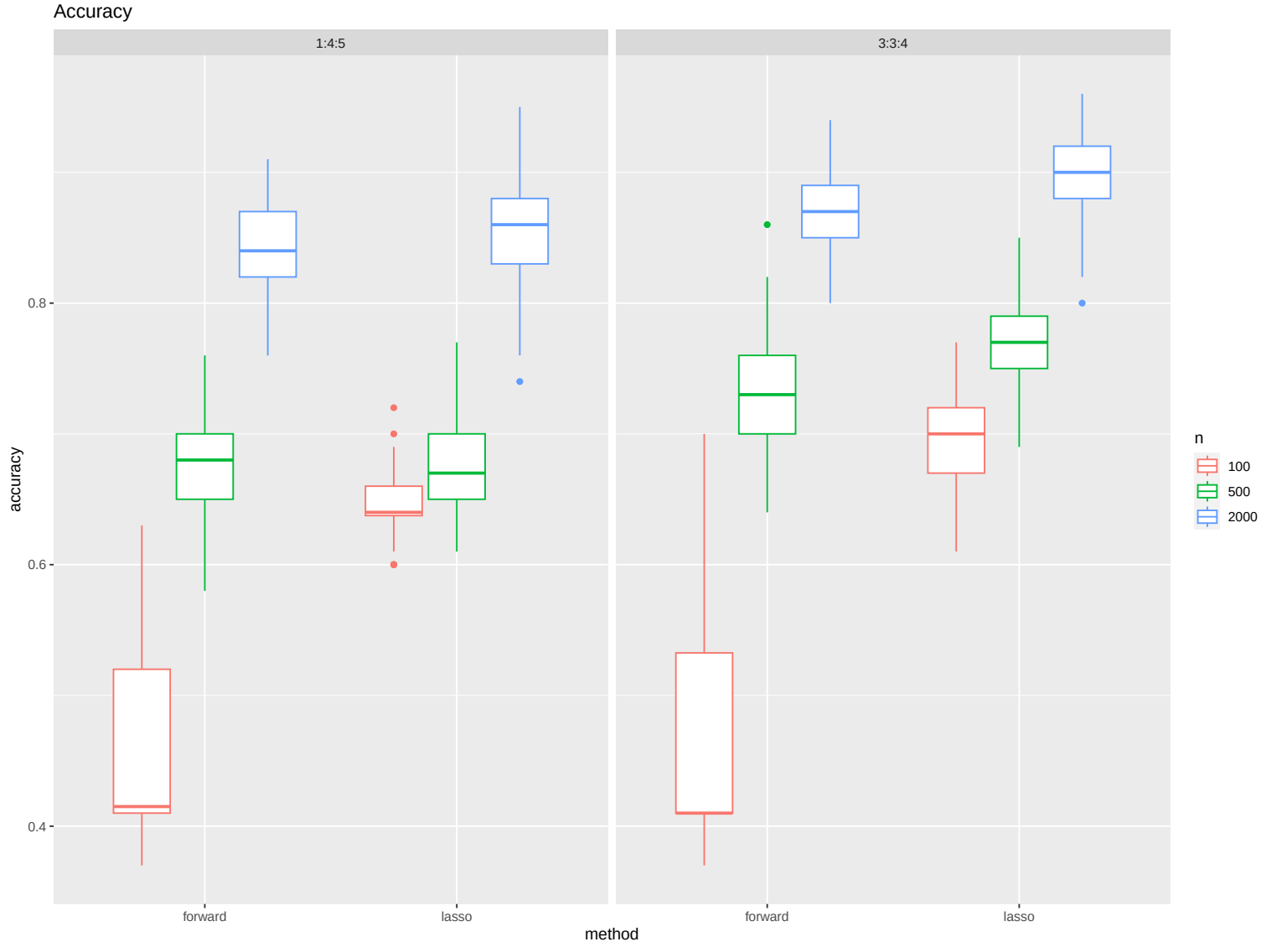

Figure 7: Accuracy Performance

Signal Identification Performance - Accuracy

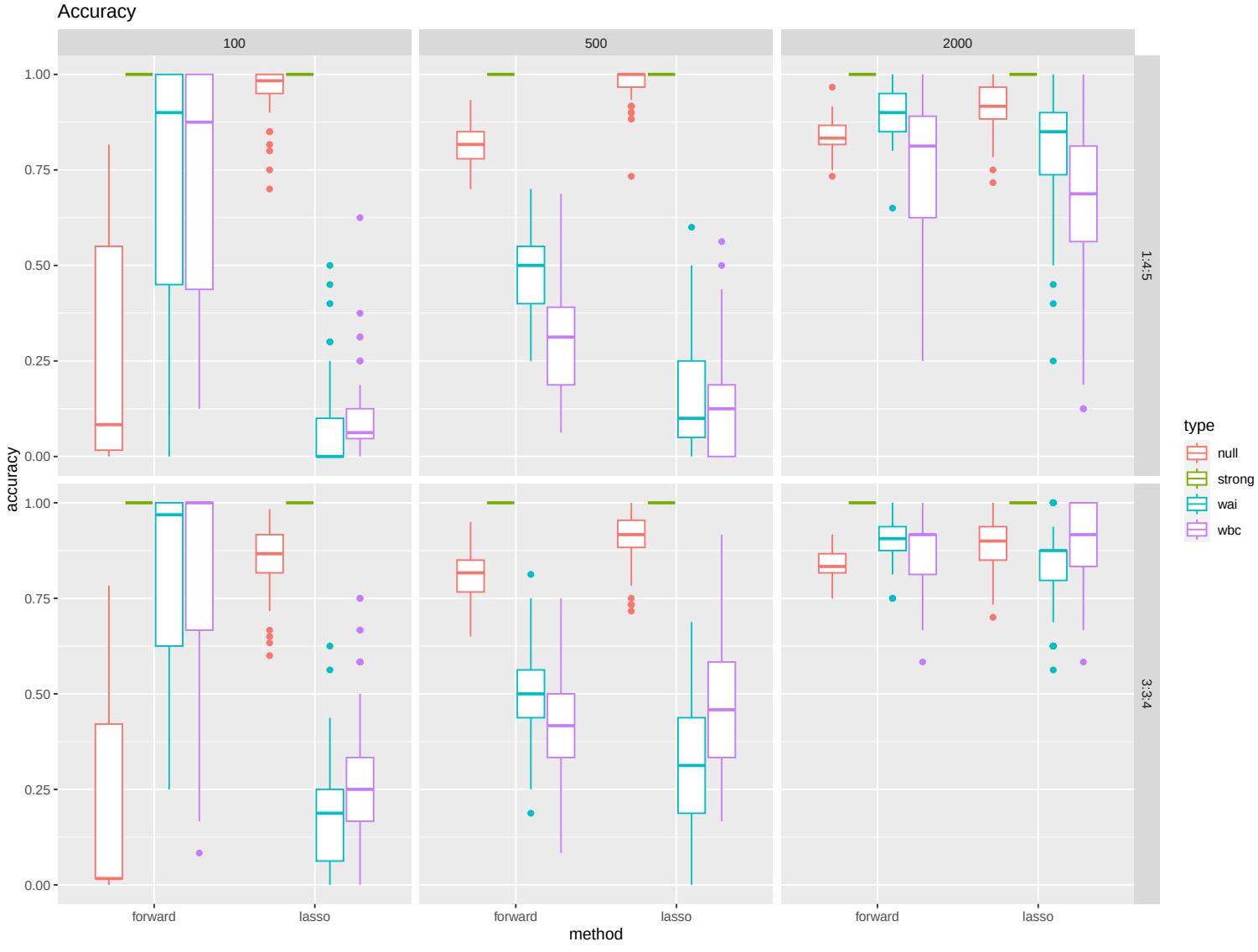

Figure 8: Accuracy of Different Signals

Signal Identification Performance - Accuracy Exploration

- Accuracy is low in high dimensional scenario, especially forward selection
- Accuracy increases for both models when n increase
- Accuracy is higher when the ratio of strong predictors is higher, and Lasso has overall higher accuracy than forward
- In high dimensional, forward has higher accuracy for weak predictors than Lasso

Parameter Estimation Performance - n=100

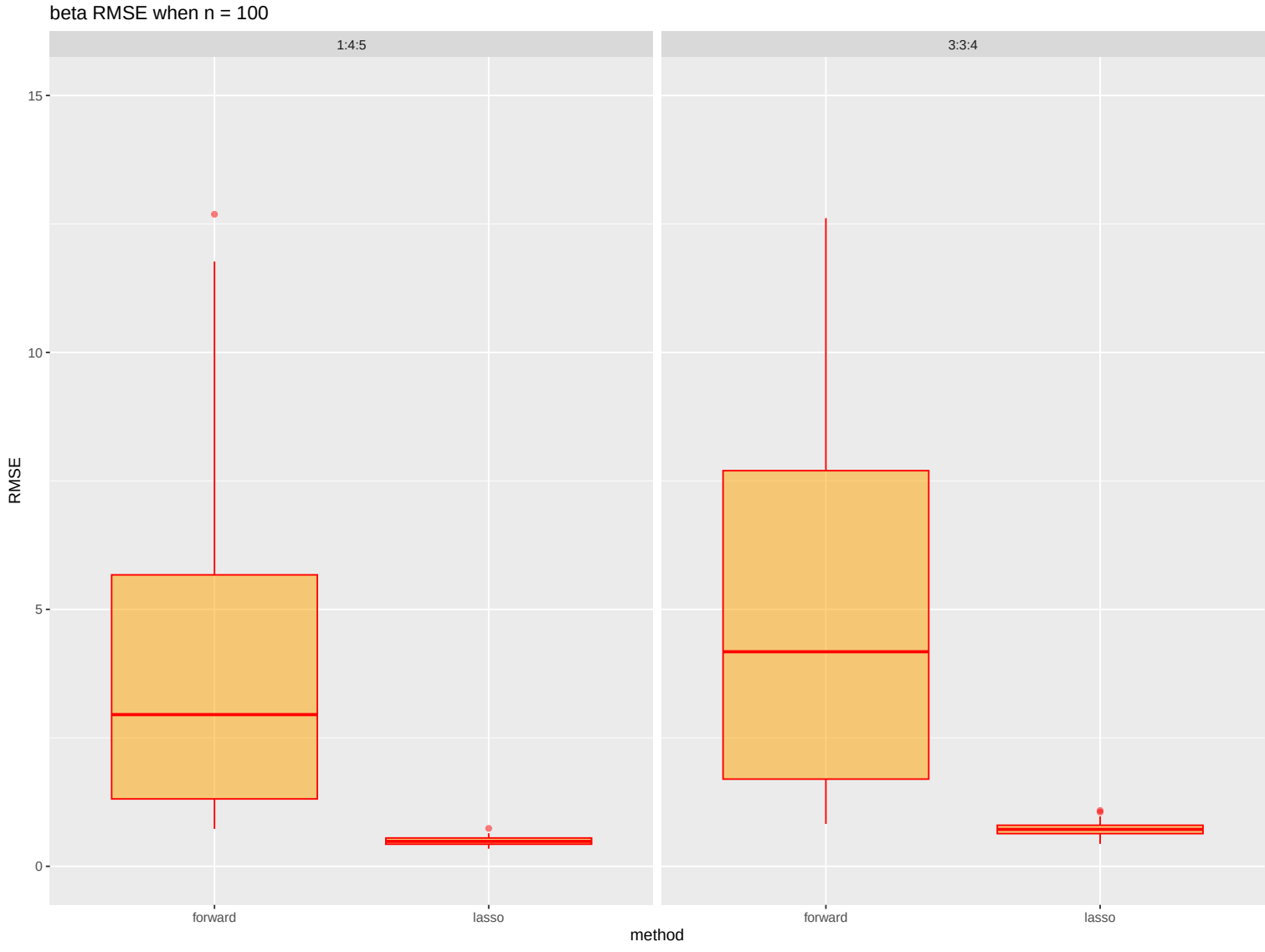

Figure 9: Beta Rmse when n=100

Parameter Estimation Performance - n=500

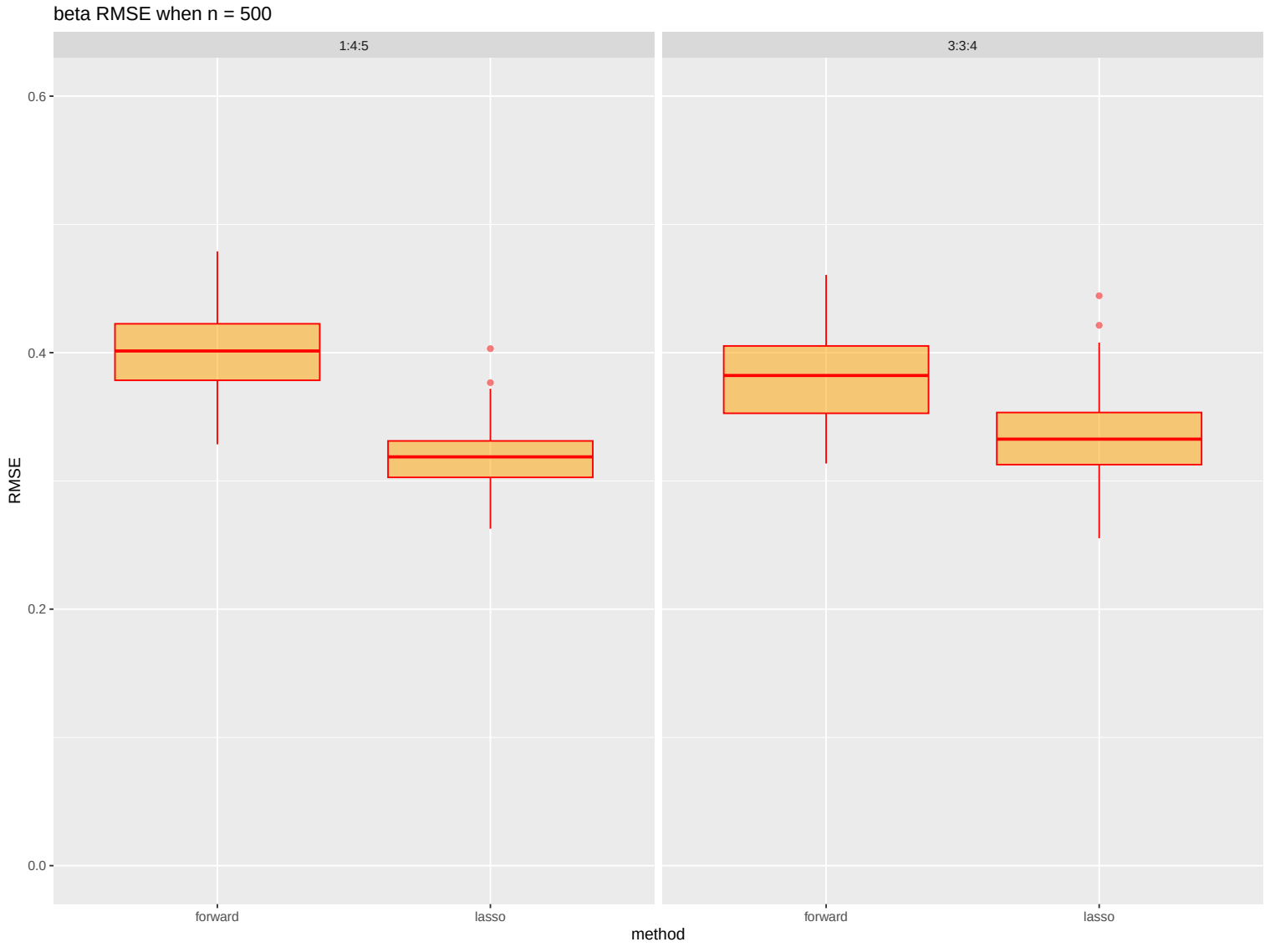

Figure 10: Beta Rmse when n=500

Parameter Estimation Performance - n=2000

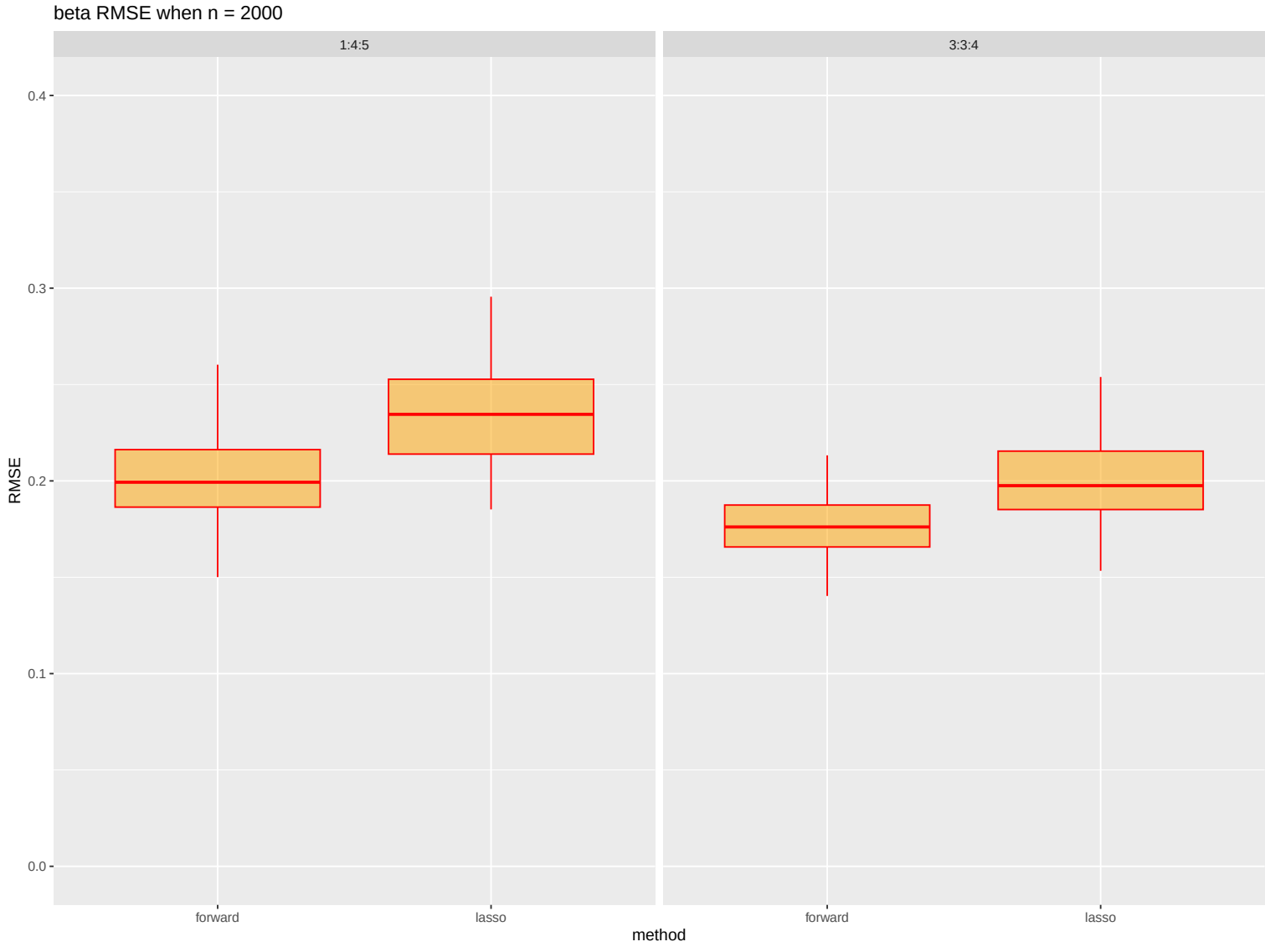

Figure 11: Beta Rmse when n=2000

Parameter Estimation Performance - Exploration

- In high-dimensional scenario, Lasso performs much better than forward
- Lasso performs better than forward when n=500, but worse than forward when n=2000

Predictors Identification Conclusions

High Dimension Scenario - Forward selection tend to be assertive in selecting true predictors while Lasso is conservative, resulting in extremely high sensitivity and more selected predictors for forward selection, and high specificity and less selected predictors for Lasso - In high-dimensional scenarios, forward selection has an overall much higher F1-score, and both methods perfectly identify the strong predictors -

Missing Weak Predictors Analysis - Introduction

How missing "weak" predictors impacts the parameter estimations

Definition: missing weak predictors = true weak predictors but estimated as null

How to value parameter estimations: RMSE

Missing Weak Predictors Analysis - Methods

How to value parameter estimations: RMSE

Most missing: simulations that have the least non-null estimations

Least missing: simulations that have the most non-null estimations

Middle: in between

Missing Weak Predictors Analysis - Result: n=100

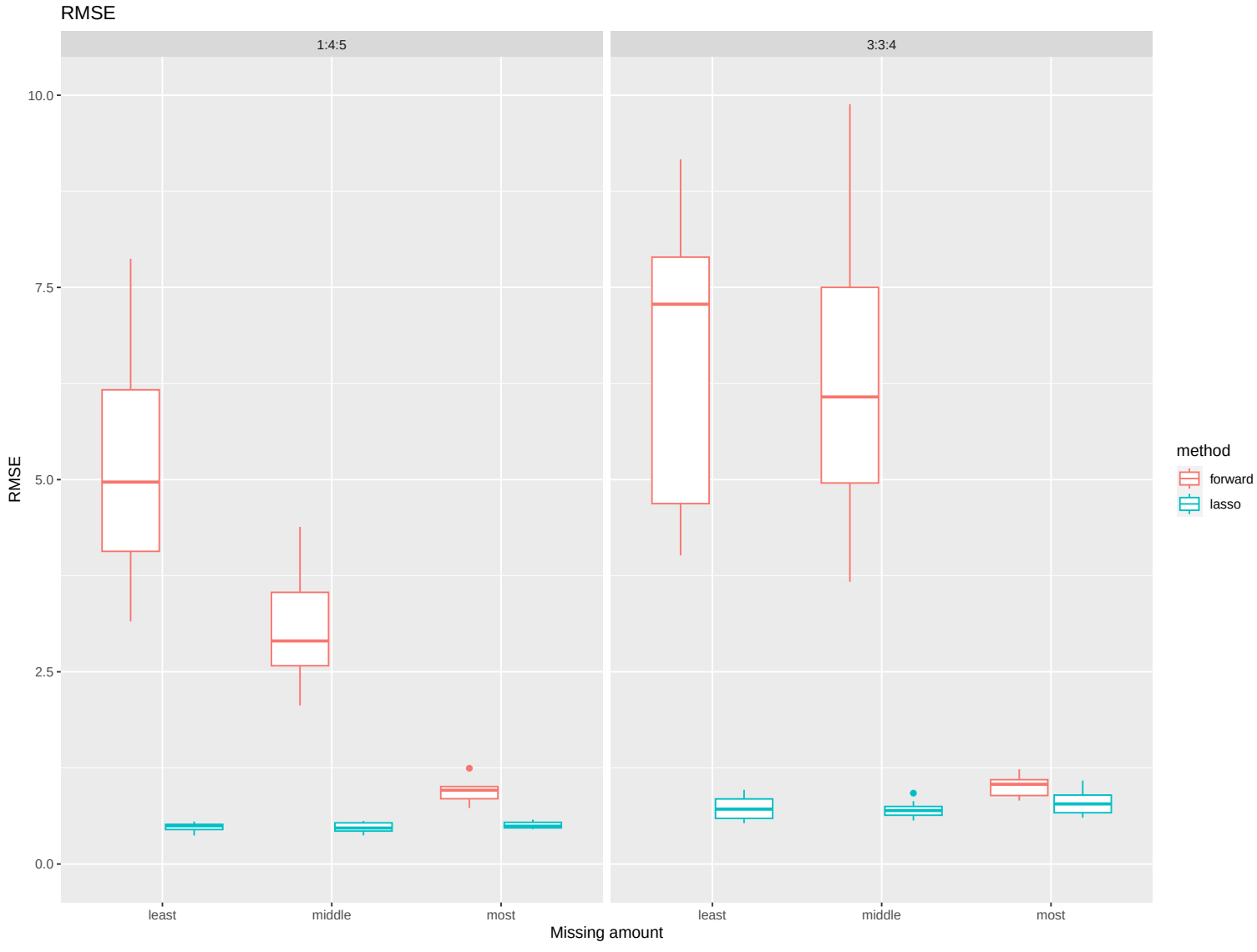

Figure 12: RMSE when n=100

Missing Weak Predictors Analysis - Result: n=500

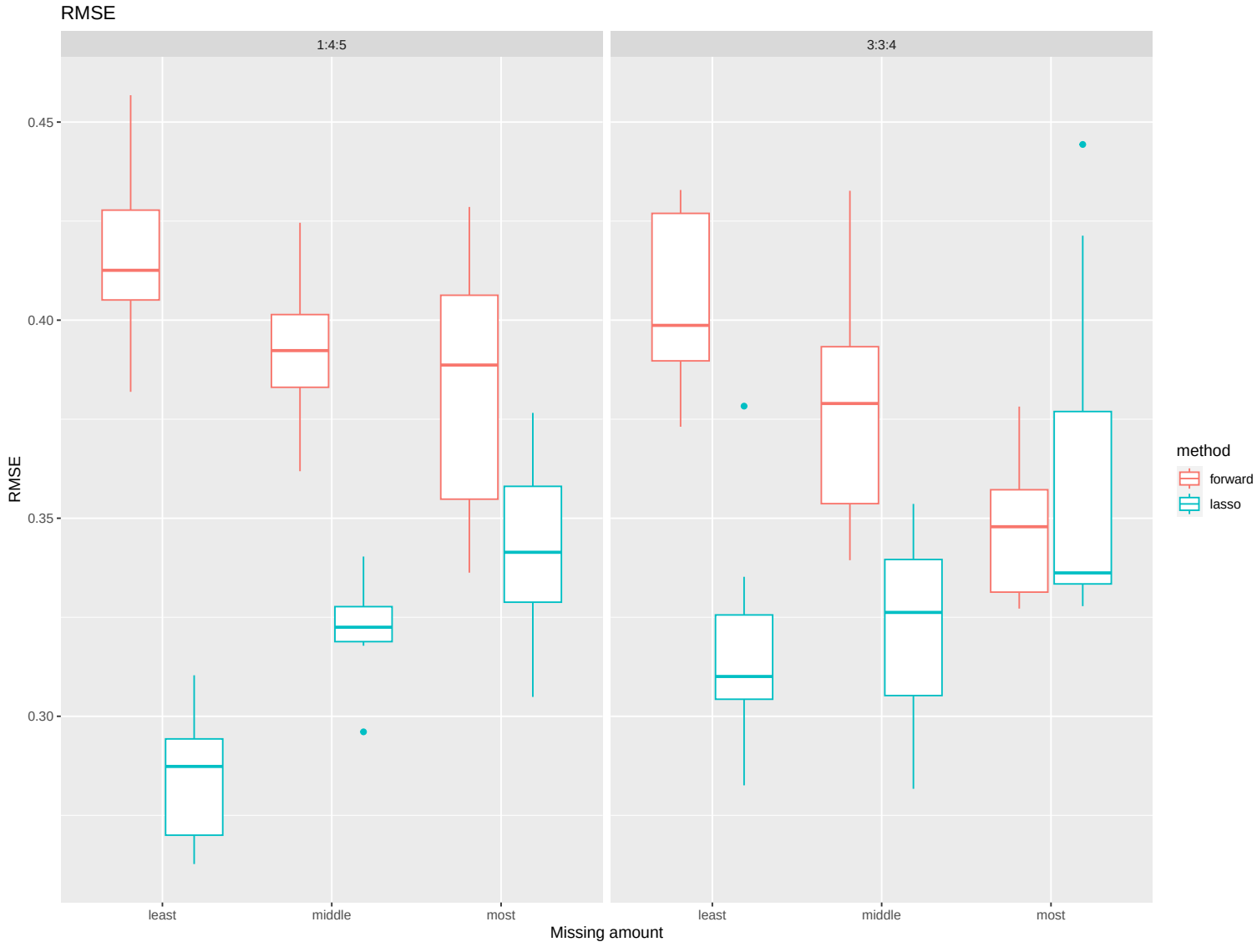

Figure 13: RMSE when n=500

Missing Weak Predictors Analysis - Result: n=2000

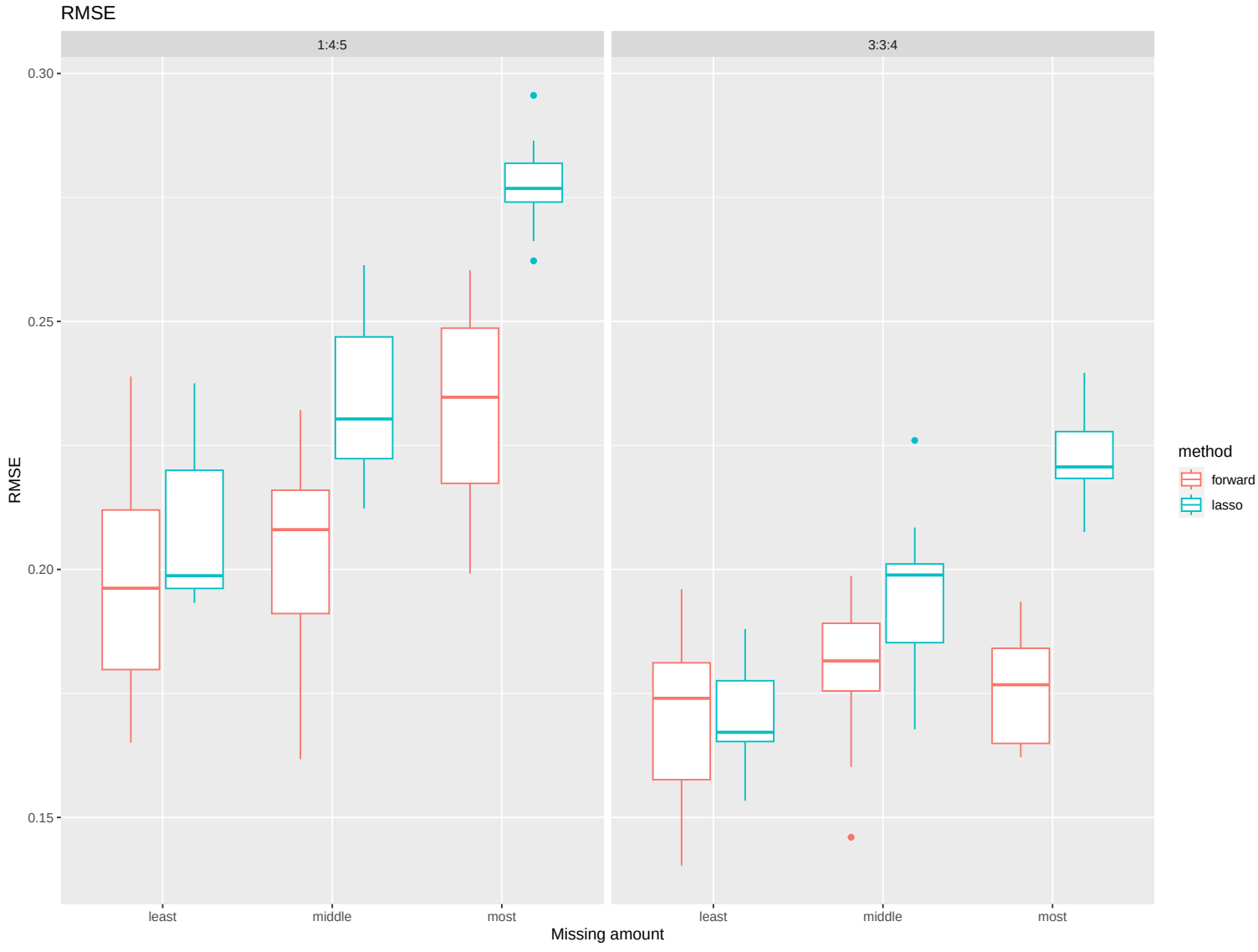

Figure 14: RMSE when n=2000

Missing Weak Predictors Analysis - Discussion

- No apparent patterns between different ratios
- In high-dimensional scenarios, Lasso performs much better than forward selection according to RMSE, no matter how much missing.
- When n is large enough, RMSE of both methods become small.
- When $n = 500$, Lasso is slightly better than forward selection, however, when $n = 2000$, just the reverse.
- In Lasso, RMSE seems to increase if the missing amount increases, but in forward selection, RMSE decreases when missing amount increases.

Summary of Results

- For binary outcomes, the simple bootstrap tended to underestimate the standard error
- Larger standard error estimates from complex bootstrap in binary and continuous settings
- Differences between simple and complex bootstrap were smaller for larger sample sizes
- Complex bootstrap not as reliable in small sample sizes

Limitations

- Sample size / treatment (or exposure) prevalence
- Small number of initial samples, limited in detecting significant differences in coverage rate

Future Work

- Larger number of initial samples, narrower coverage window
- Increased sample size, changes in bootstrap performance?
- Changes in treatment propensity model
- Non-normal distributions of covariates