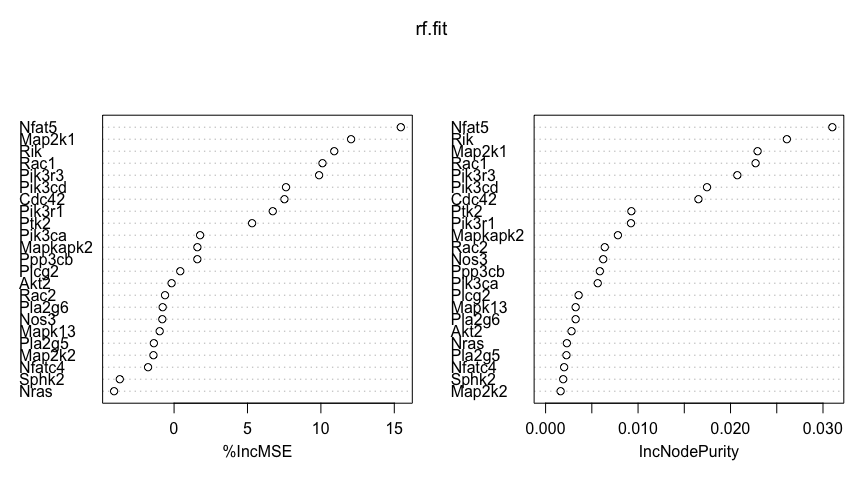
Random forest method

We also tried random forest method to create a model. The random forest builds a number of decision trees, every single time a predictor from the full set is selected to split in the tree.

We creates 100 decision trees for random forest in our estimation. In order to compare with other methods, we also use ten-fold cross validation to train our model of random forest.

We can observe the importance for each predictor to interpret the result from the random forest. The way to understand the importance of each predictor is to capture the variance base on %IncMSE and IncNodePurity.



For the %IncMSE, the larger the mse of prediction(which is estimated with out-of-bag-CV) if a result of predictor is permuted, the more importance of predictor is.

For the IncNodePurity, it interprets the importance of predictor in another way. The most important predictor, that split the data, could achieve the highest increase in node purities.

According to our results, the Nfat5, Map2k1, Rik, Rac1, Pik3r3 are the five most important predictors to the model, which is very similar to the result that our Lasso method create.

The testing error is around 0.010, which also achieve a good performance for the prediction.

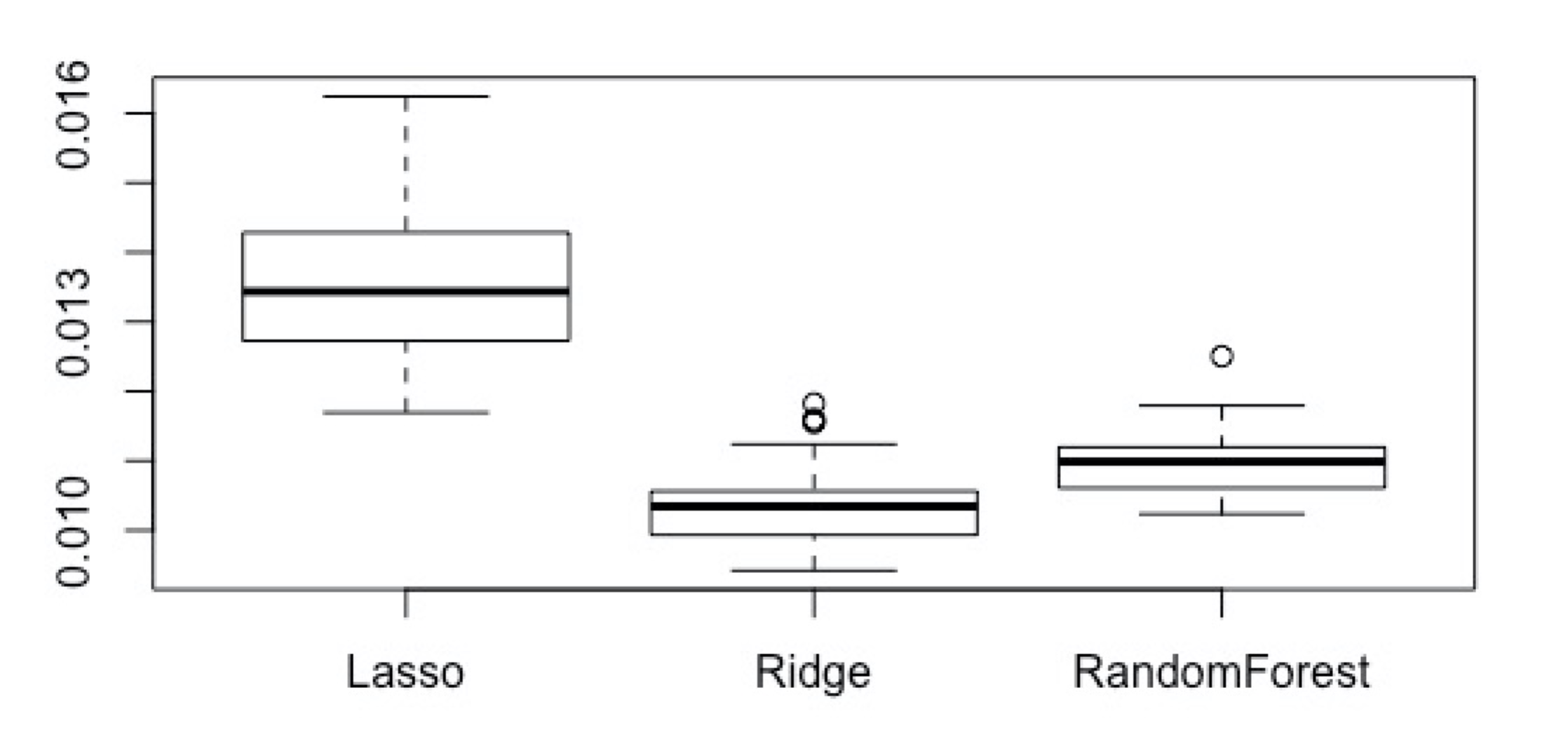
Random Forest VS the Ridge VS the Lasso

The procedure of cross-validation: Set seed from 1 to 50, and with each seed randomly shuffle dataset to perform one time cross-validation. Use boxplot to show means, variances and outliers of different methods.

Total lasso: 0.0134601980207286

Total ridge: 0.0103132458276357

Total random forest 0.010975346957304



The model of the Ridge performs the best among the models, then the random forest, then the Lasso. However, the first two models are also more complicated. The Ridge doesn’t perform the variable selection. Although the random forest prune insignificant variables, the model is still close to the full mode. In contrast, the Lasso has the feature of variable selection, so its model is more interpretable. Therefore, in order to obtain a simple and easily interpreted model, we decided to use the Lasso for this dataset.

We learned from the random forest and the Ridge that models with high computation complexity could create good prediction. Hence, if heavy computation is affordable, models which require high computation could be a good choice.