Final project

By Ming han, Lei Kang, and Yanyu Long

In this project, we used best subset, forward stepwise and backward stepwise method to generate models. For each method, we used Cp, BIC and cross-validation to determine the best models. In addition, we used Lasso method with cross-validation to determine the best lambda and then the best model. When cross-validation is used, the effect of number of folds is briefly evaluated. With more folds the results have smaller variation and the same best model is always selected. Then the 6 models obtained were compared with cross-validation, which gave the best model among them with the lowest test MSE, i.e. **Mapk1 ~ Akt2 + Rik + Pik3r3 + Pik3r1 + Rac1** (\*).

We then checked whether the selected model has collinearity. It is found that Rik and Pik3r3 are highly correlated, so either one of these variables can be dropped to give a simpler model. However, further examination using cross-validation show that the performance of the (\*) model is better than the models with either Rik and Pik3r3 being dropped. We concluded that (\*) is the simple yet best model. In the end, we also checked the linear assumptions of the model and compared it with the results generated by random forest.

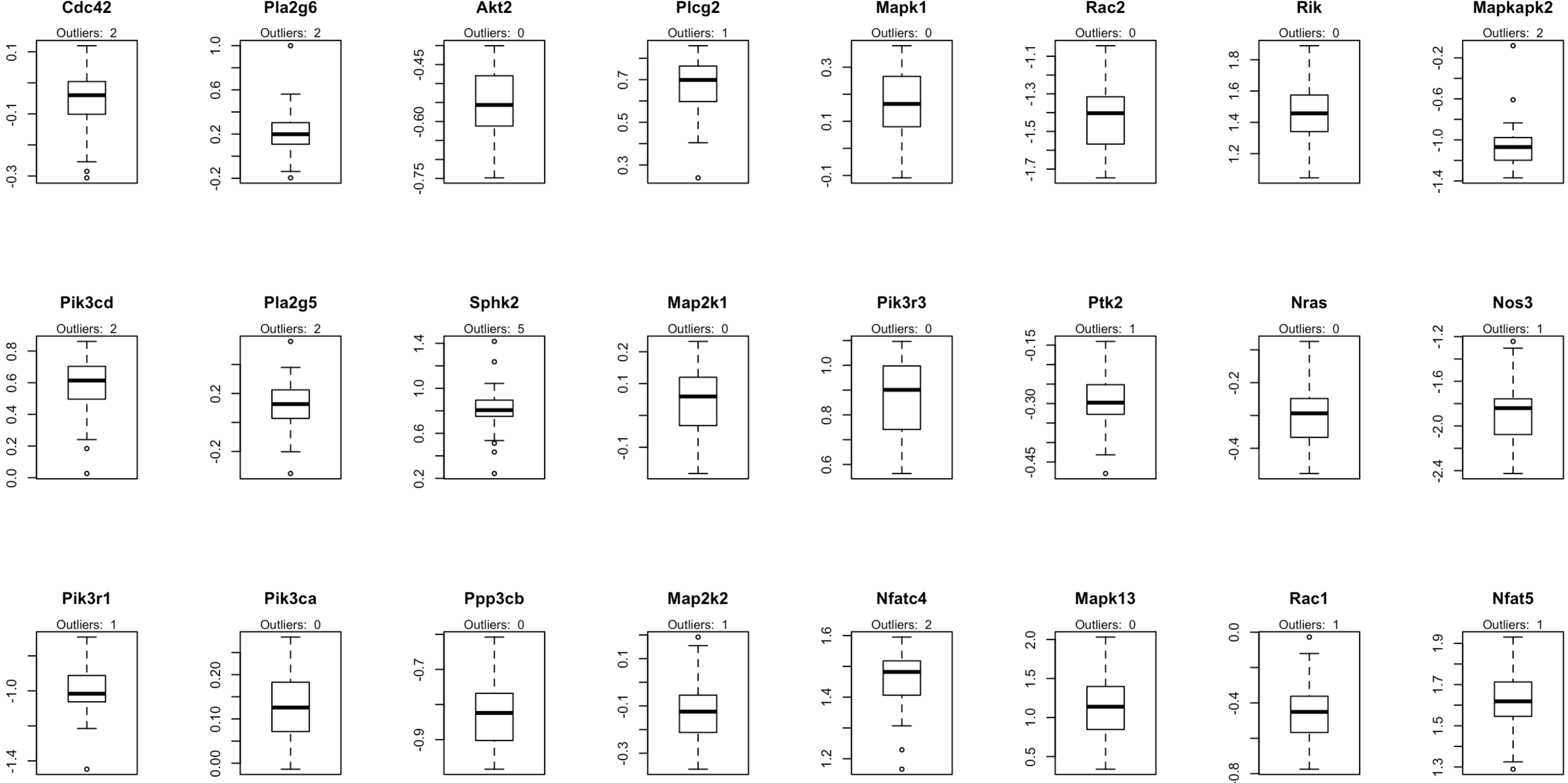
1. Summary of the data

The variables in this dataset:

"Cdc42", "Pla2g6", "Akt2", "Plcg2", "Mapk1", "Rac2", "Rik", "Mapkapk2", "Pik3cd", "Pla2g5", "Sphk2", "Map2k1", "Pik3r3", "Ptk2", "Nras", "Nos3", "Pik3r1", "Pik3ca", "Ppp3cb", "Map2k2", "Nfatc4", "Mapk13", "Rac1", "Nfat5"

For the sake of simplicity, the variables are assigned number #1-#24.

Check the distribution of all the data:



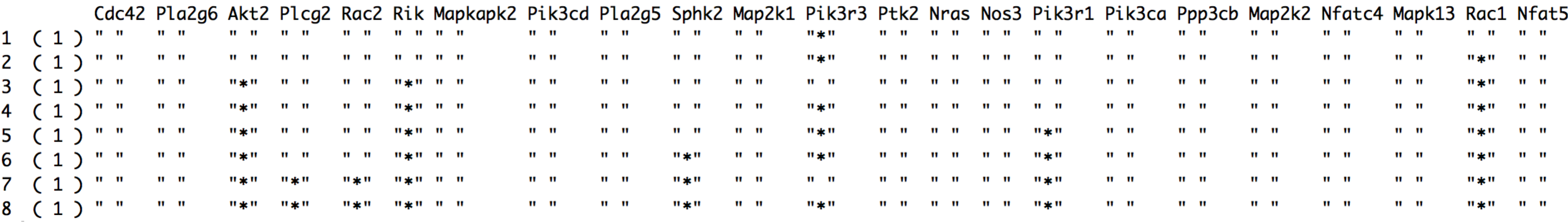
*Figure 1. boxplots of all the variables*

There are some outliers in some variables (1.5 IQR). In the later analysis we did not exclude them, because the sample size is small (n = 40); if all the outliers are excluded, the sample pool will be reduced to about 25.

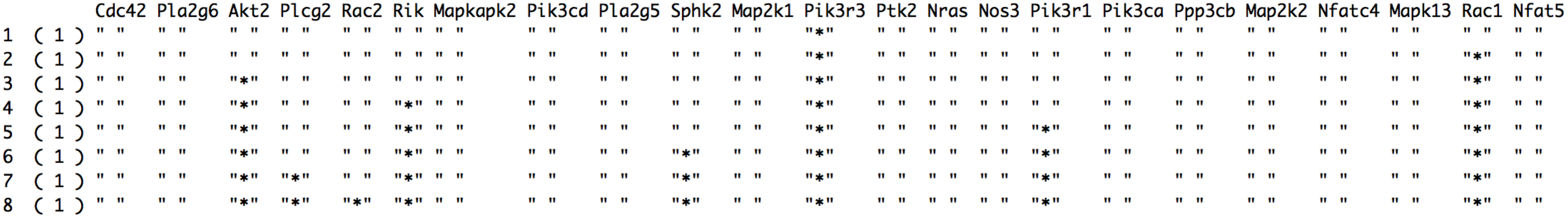
1. Best subset, forward and backward models

* Results of the three approaches (in red squares are models selected by Cp and BIC)

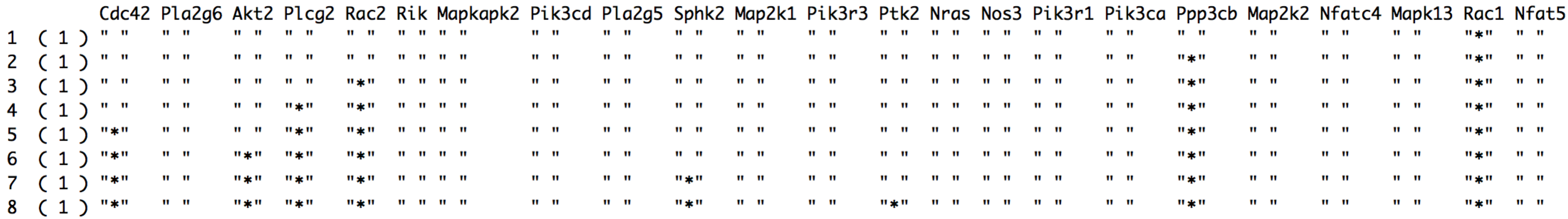
1. Best subset:



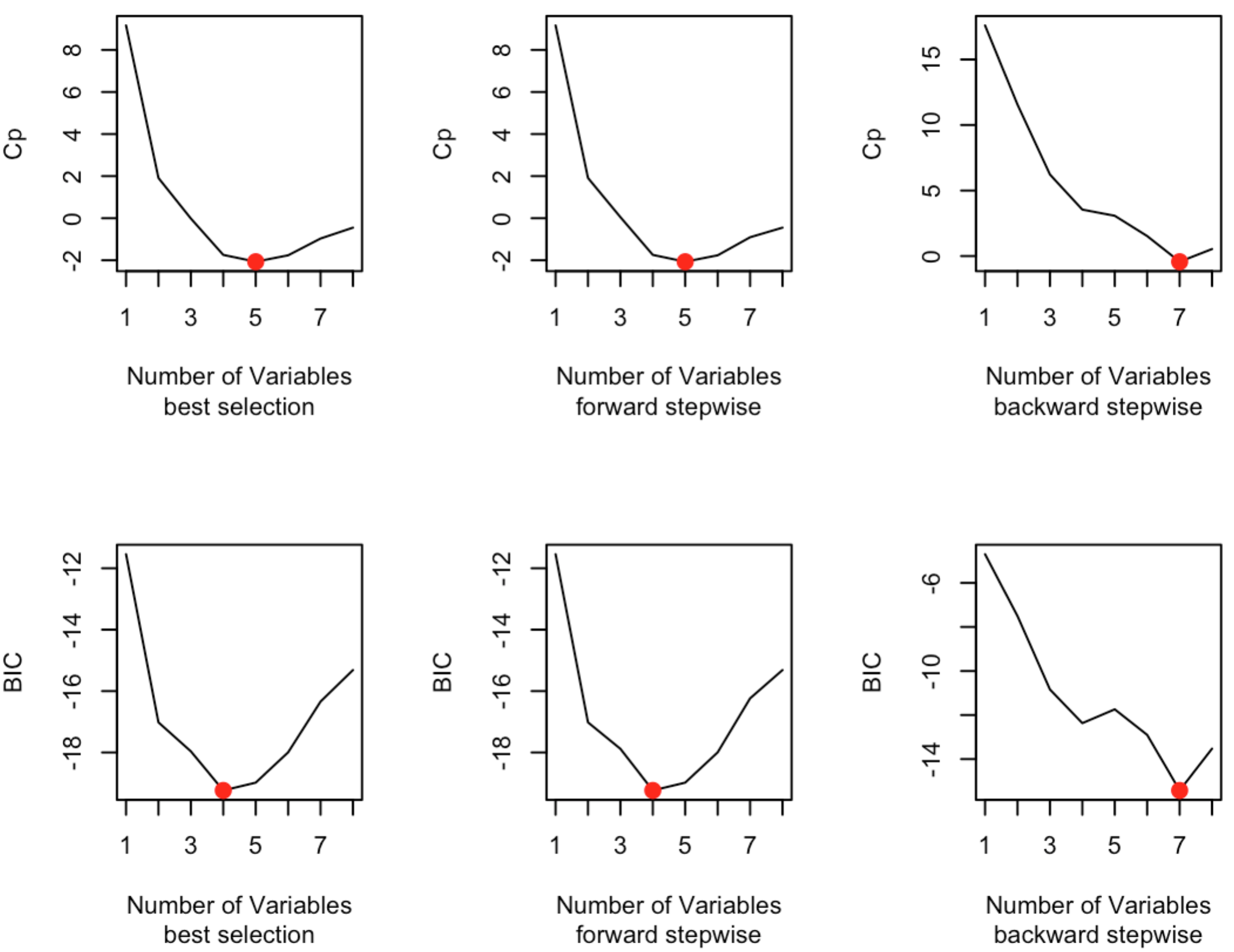
1. Forward (very similar to the models selected by best subset):



1. Backward (quite different from the results of the best subset and forward stepwise):



* Evaluation of the three models using Cp and BIC



*Figure 2. Cp and BIC for models selected by the three methods*

* Best subset and forward method:

The best model selected by Cp: Akt2, Rik, Pik3r3, Pik3r1, Rac1

(Intercept) Akt2 Rik Pik3r3 Pik3r1 Rac1

-0.6184661 -0.3965544 0.2496411 0.2235856 -0.1619541 0.3302738

The best model selected by BIC: Akt2, Rik, Pik3r3, Rac1

(Intercept) Akt2 Rik Pik3r3 Rac1

-0.4446112 -0.4054803 0.2207164 0.2439920 0.3171645

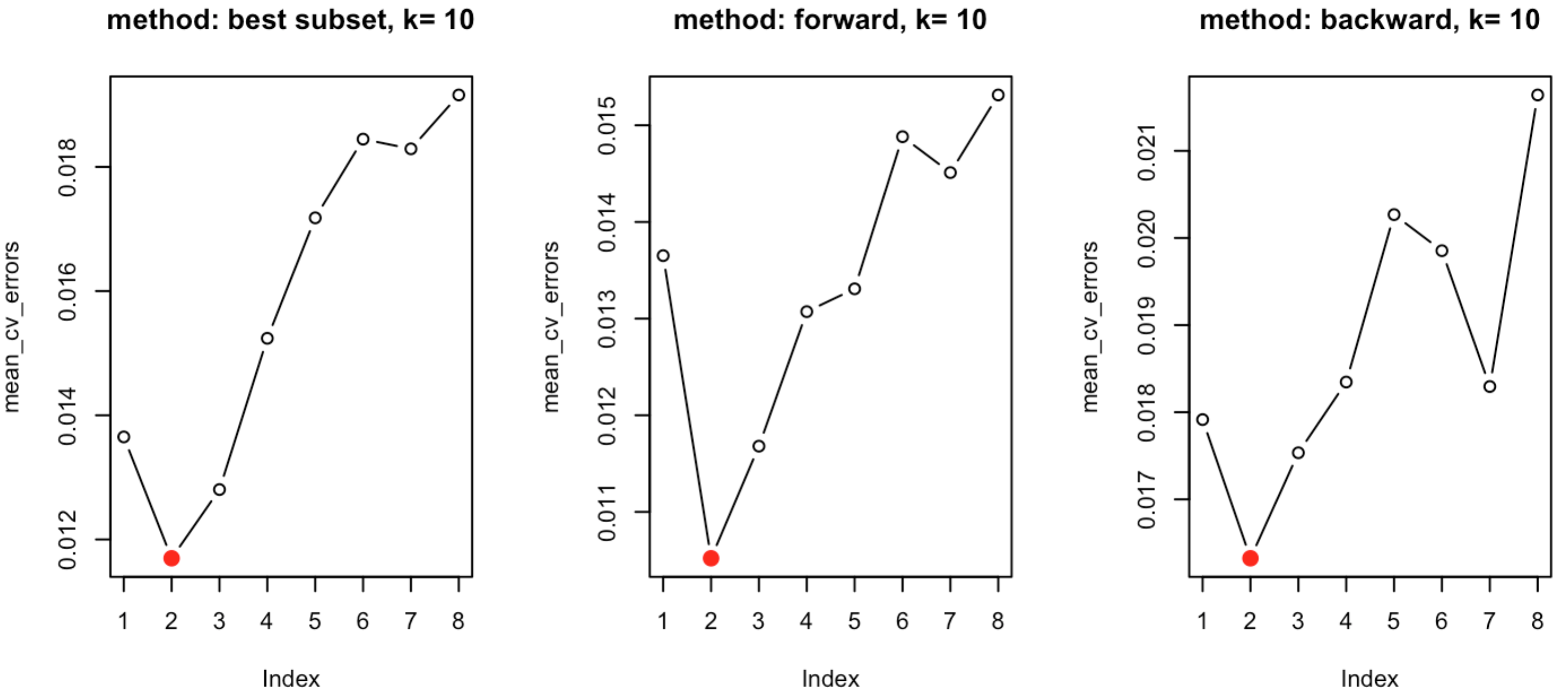
* Backward method:

Both Cp and BIC select Cdc42, Akt2, Plcg2, Rac2, Sphk2, Ppp3cb, Rac1

(Intercept) Cdc42 Akt2 Plcg2 Rac2 Sphk2 Ppp3cb Rac1

0.07666044 0.37749256 -0.38665341 0.28977515 -0.25753671 -0.17559530 0.43222257 0.35938288

* Evaluation of the three models using 10-fold cross-validation[[1]](#footnote-1)



*Figure 3. cross-validation test MSE for models selected by the three methods*

Best subset and foward: Pik3r3, Rac1

(Intercept) Pik3r3 Rac1

-0.05642834 0.40410424 0.27832221

Backward: Ppp3cb, Rac1

(Intercept) Ppp3cb Rac1

0.7430295 0.4716582 0.4022791

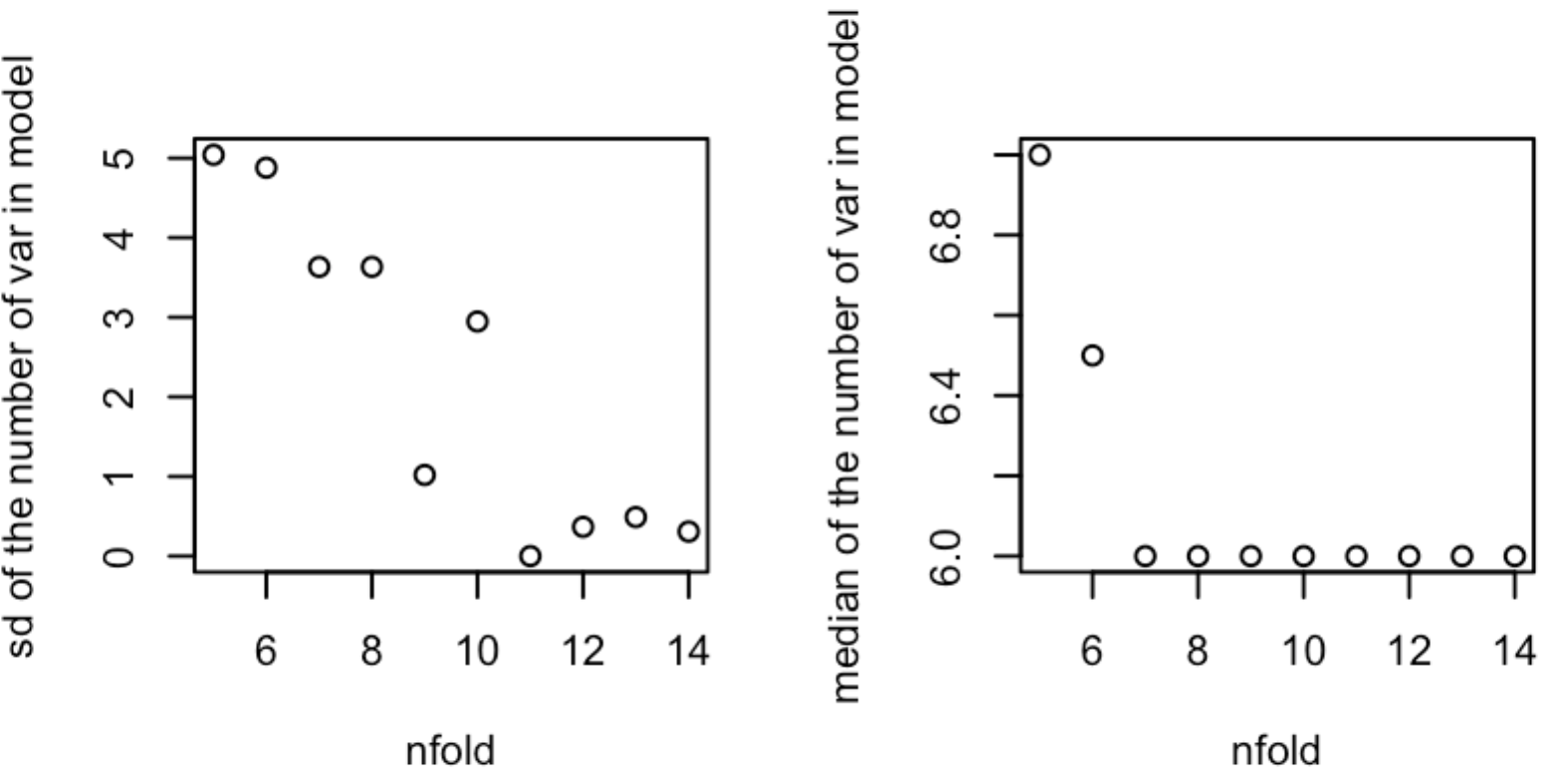
Compared to using Cp and BIC, cross-validation approach yields significantly smaller models.

1. Lasso method:

We used cross-validation to obtain the best lambda that gives the lowest test MSE, and then used the best lambda to obtain the coefficients that are not zero. This gave the final model with a small number of selected variables.

The final model changes depending on the grid used[[2]](#footnote-2) and the number of folds in the cross validation, and varies when set.seed is set different (variation).

To obtain a stable model using this method, we tested how the number of CV folds (nfold) can influence the selected variables in the final model. We used the number of selected variable (*n*) as a simple indicator of whether the final models are the same or different. We obtained the standard deviation and median of *n* by setting set.seed different (so every time the training/test set are different). The results are present below:



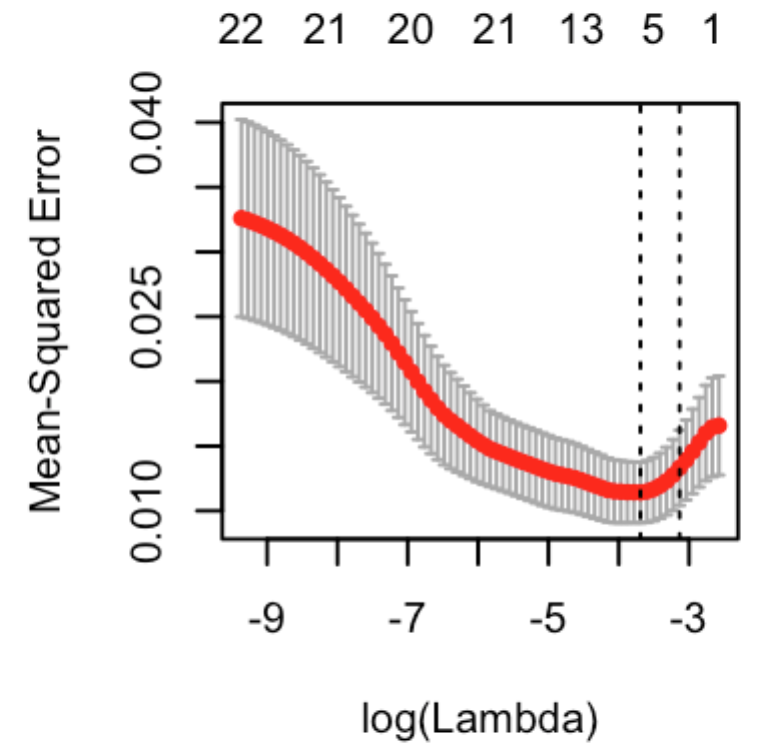
*Figure 4. sd and median of the number of variables in the best models using different set.seed(n)*

As nfold increases, sd decreases (less variable); also the median is stable at 6 from nfold = 7. Therefore when nfold is not too small, the best models given by this method should always include 6 variables, which are listed below: Rik, Pik3cd, Pik3r3, Rac1, Nfat5

(Intercept) Rik Pik3cd Pik3r3 Rac1 Nfat5

-0.13205963 0.10555116 -0.03199517 0.16367227 0.13200524 0.05115234

When nfold = 10, the test MSE as a function of log(lambda) is provided below. The lambda is 0.02493913.



1. Compare models:

The above methods gave 6 models:

(1). Mapk1 ~ Pik3r3 + Rac1 (Best/Forward CV)

(2). Mapk1 ~ Ppp3cb + Rac1 (Backward CV)

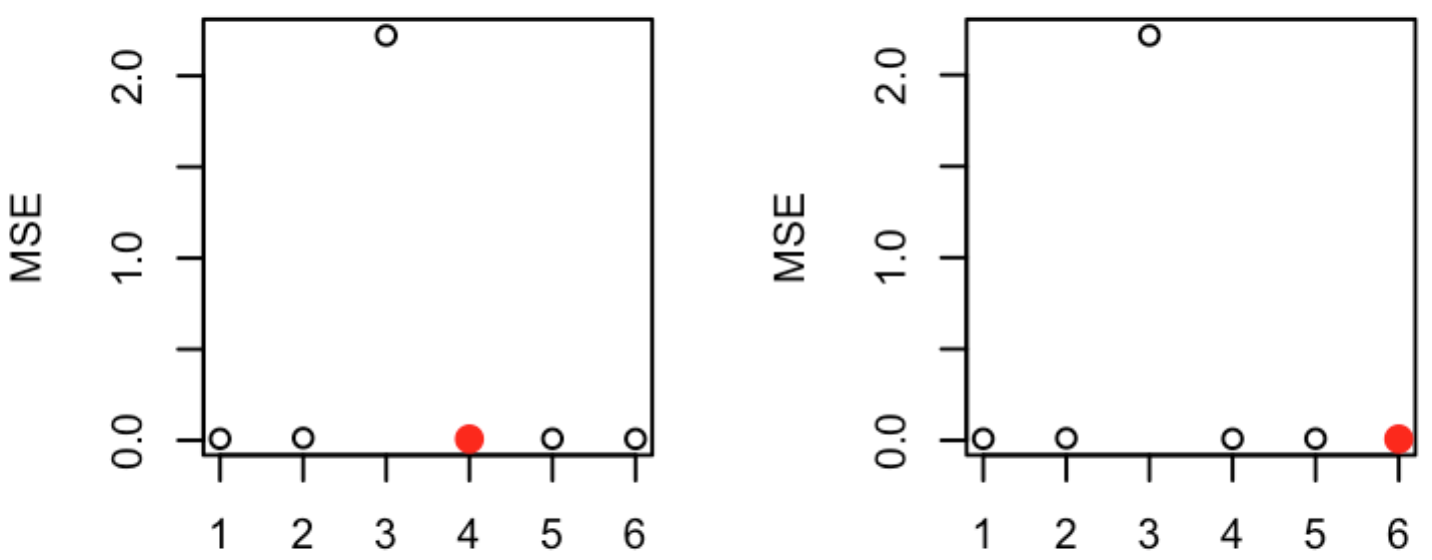
(3). Mapk1 ~ Akt2 + Rik + Pik3r3 + Rac1 (Best/Forward BIC)

**(4). Mapk1 ~ Akt2 + Rik + Pik3r3 + Pik3r1 + Rac1 (Best/Forward Cp)**

(5). Mapk1 ~ Rik + Pik3cd + Pik3r3 + Rac1 + Nfat5 (Lasso)

(6). Mapk1 ~ Cdc42 + Akt2 + Plcg2 + Rac2 + Sphk2 + Ppp3cb + Rac1 (Backward BIC/Cp)

We used 10-fold[[3]](#footnote-3) CV to compare these models. When set.seed(n) differ, the selected model differ. Below 2 such scenarios are provided:



*Figure 5. the CV test MSE of the 6 models (left panel: set.seed(1); right panel: set.seed(2))*

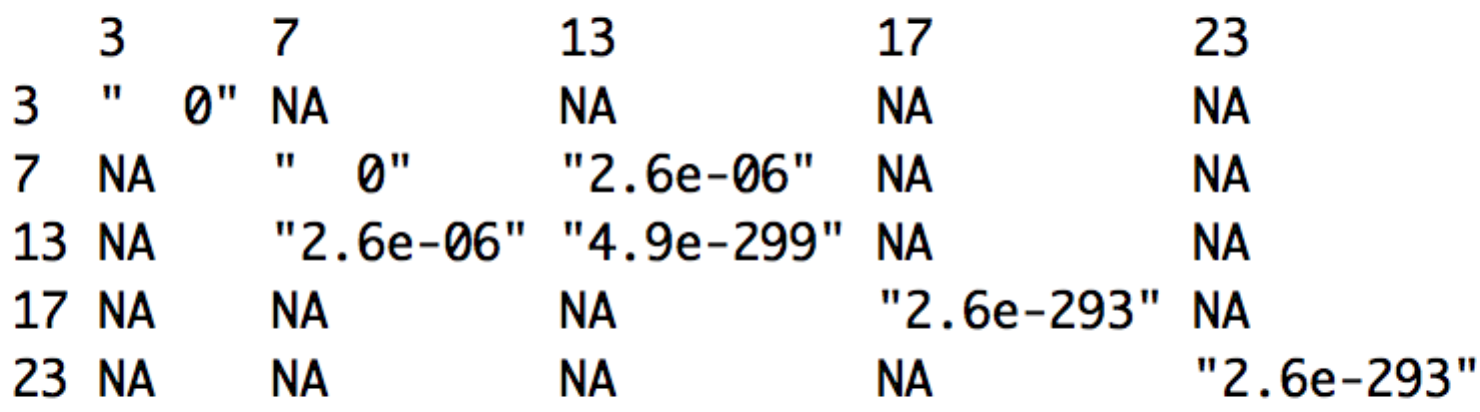
The x-axis is the indices of each model as listed above, where (4) indicates the model generated by the best subset/forward method with Cp as the evaluation method. The selected variables are: Akt2, Rik, Pik3r3, Pik3r1 and Rac1; (6) indicates the model generated by backward method with Cp/BIC as the evaluation method, which contains 7 variables. When set.seed(n) differ, most of time the selected model is (4); since it also contains fewer variables than (6), this should be the best simple model. Therefore, the final model is:

**Mapk1 = -0.61847 - 0.39655 \* Akt2 + 0.24964 \* Rik + 0.22359 \* Pik3r3 - 0.16195 \* Pik3r1 + 0.33027 \* Rac1**

1. Check the model:

1) Collinearity

We then checked the correlation between each variable in the selected model #4, since high correlation between variables may cause collinearity, and a simpler can be obtained if some correlated variables are dropped. The p-values of Pearson correlation of each pair of variables in #4 are given below (NA indicate p > 0.05)



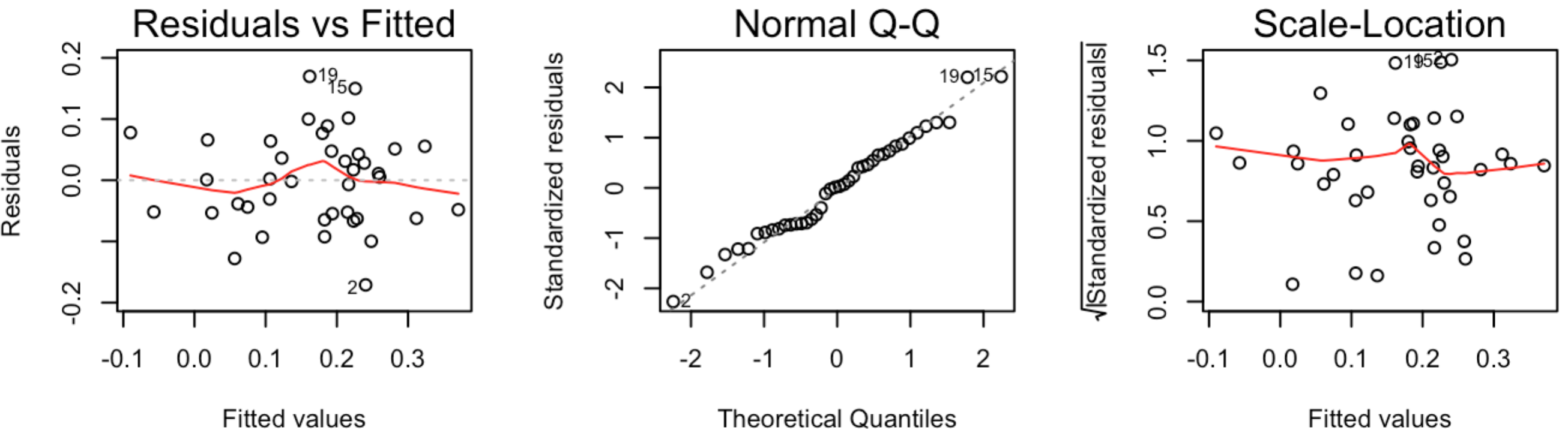
It can be seen that Rik (#7) and Pik3r3 (#13) is highly correlated. In the case of the collinearity, we can choose/drop either variable in the model. We then have two more models:

(7). Mapk1 ~ Akt2 + Pik3r3 + Pik3r1 + Rac1

(8). Mapk1 ~ Akt2 + Rik + Pik3r1 + Rac1

We performed another 10-fold cross-validation to evaluate whether (7) and (8) will have better performance (lower test MSE). The results show that the model with all the five variables (model 4) always has better performance (we leave out the graph here).

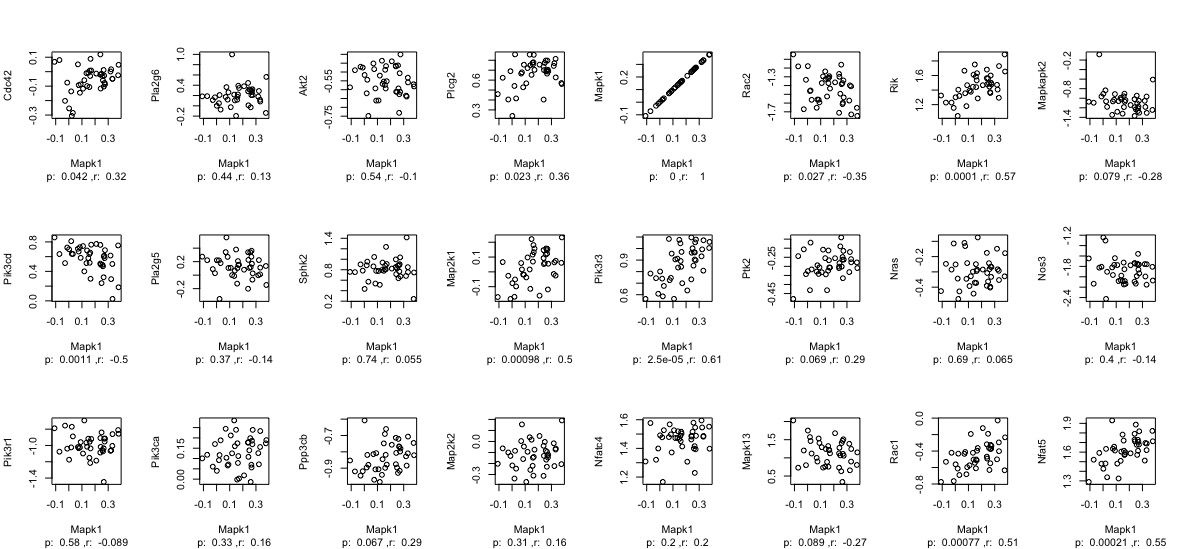
Besides collinearity, we also checked some assumptions of linear models, such as whether the relationship between the responses and predictors is linear and whether there are outliers that may distort the results (some outliers are shown in the boxplots). By plotting the model, we can see that 1) there is no apparent pattern in residuals vs. fitted values; 2) the QQ plot of standardized residuals suggests a normal distribution; 3) there is no obvious outliers (no standardized residual exceeds 3). Therefore the model satisfied the linear assumptions.



*Figure 6. Check the assumptions of linear models*

2) Further check

We did a brief examination on the correlation between all the variables and the responses variable to see whether individual variables included in the model correlate with the response variable (may not need).



*Figure 6. correlation between Mapk1 and other variables:*

Significant correlation (possible predictors): Cdc42\*, Plcg2\*, Rac2\*, Rik \*\*\*, Pik3cd \*\*, Map2k1\*\*\*, Pik3r3\*\*\*, Rac1\*\*\*, Nfat5\*\*\*

\* p < 0.05; \*\* p < 0.01; \*\*\* p < 0.001

Compared with our final model: **Mapk1 ~ Akt2 + Rik + Pik3r3 + Pik3r1 + Rac1,** Rik, Pik3r3 and Rac1, which are highly correlated with the response variable, are included in the final model; while Akt2 and Pik3r1 are two variables that are not correlated with the response variables. This suggests that in a simple linear model (correlation), the contribution of a predictor on the response variable may be covered and only can be shown when some other variables are controlled, i.e. as in the multiple regression.

Compare the final model with the results of random forest:

1. We did not use validation approach set because the sample size is small, the training and test set would be 20 if the validation set approach were used, which would yield results (the number of variables) with large variation (tested by using different set.seed(n)). The number of folds is set to be 10 instead of 5, because when k=5, the results also have larger variation, while the variation is very small when k = 10: always 2 predictors for forward and best subset method, and around 2 for backward method (larger variation than the other two). [↑](#footnote-ref-1)
2. We used the default grid values given in glmnet/cv.glmnet. We also tried varying n1, n2 and l in grid=10^seq(n1,n2,length= l): when n1 is no less than -1 but not very big, n2 is no bigger than -1 but not very small, and l is big enough, the grid is fine-grained enough to yield stable results. One example set of these values are n1 = 0, n2 = -4 and l = 200. Using these values gave the same results as using the default grid values. Therefore we do not go into details of how different grid can influence the results. [↑](#footnote-ref-2)
3. Again, as nfold gets bigger, the best model selected by CV is more stable. Later we will stick to 10-fold and only report the most common results. [↑](#footnote-ref-3)