STSCI 4740 Final Project

Binglin Wang and Nivedita Vatsa

December 2, 2016

Abstract

In this report, we consider several statistical methods to determine the most appropriate one to predict the $Mapk1$ protein in a gene expression dataset. We compare the best subset selection method, the lasso method, the ridge method, and the principal component regression (PCR) method. There are two main randomization processes. First, the data are randomly split into a training set (n_1=32) and a test set (n_2=8). We call this randomization process "RP-1." Second, the use of 10-fold cross-validation to estimate errors has randomization in the way that the ten folds are selected. We refer to this randomization process as "RP-2." Our analysis addresses both these elements of randomness in our analysis to ensure that the selected model is the strongest. The models are compared on the basis of of the test mean squared error (MSE) from a single test validation set, the 10-fold cross-validation (CV) errors from multiple iterations of RP-2, and test MSE from 100 iterations of RP-1. We find that PCR performs the best in terms of prediction error, though the ridge model is comparable and may be preferred for its simpler interpretation.

1 Data

The dataset contains 40 observations for 23 gene expressions other than $Mapk1$. The mean of $Mapk1$ is 0.17 with a standard deviation of 0.13. The mean and standard deviation of each variable is given in table 1 and table 2 in the Appendix. We also explored the correlation among different variable, and the correlation plot is given in the appendix. This report will explore this data with an aim to specify an appropriate model, both in terms of its simplicity and prediction power.

We first divide the data into a training (n_1=32) and test set (n_2=8). In the sections below, we describe our findings using four different types of methods, namely, the best subset selection, lasso, ridge, and PCR. The general steps of our analysis are presented in figure 1 below.

2 Best Subset Selection

We start our analysis by considering the least squares method. This is a good starting point for exploring the associations among different variables. We first do the best subset selection on the training set, from which RSS, adjusted R2, Mallow Cp, AIC, and BIC are obtained. The Mallow Cp, AIC, and BIC approximate to the test error by adding penalties to the degree of freedom. The RSS, adjusted R2, Mallow Cp, AIC,

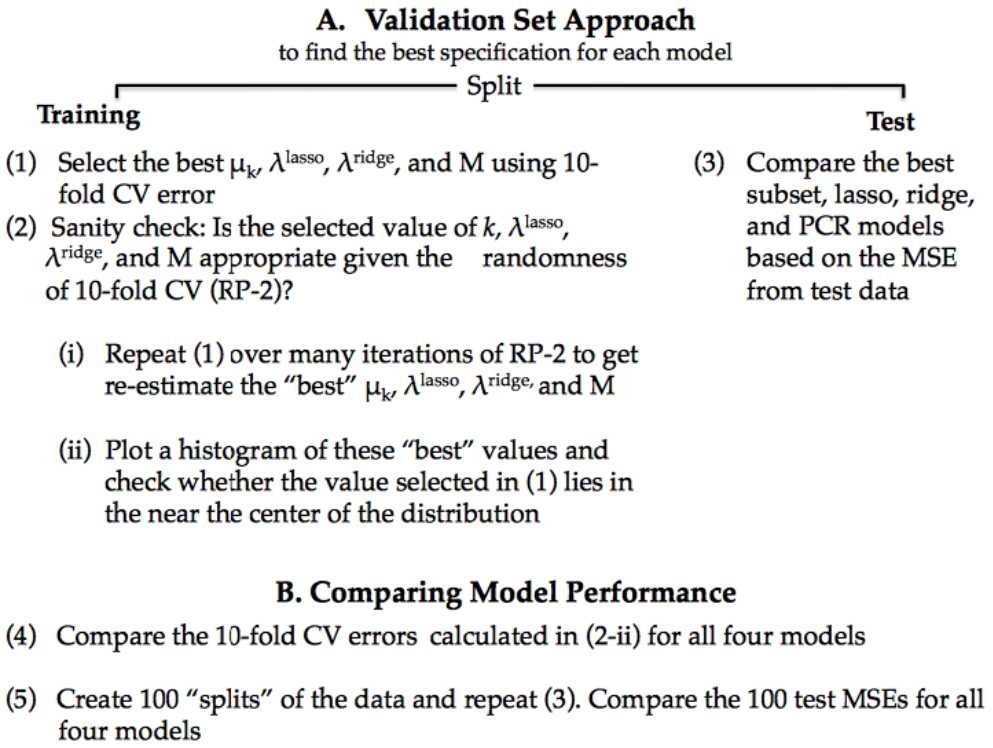


Figure 1: General steps for analysis in this project

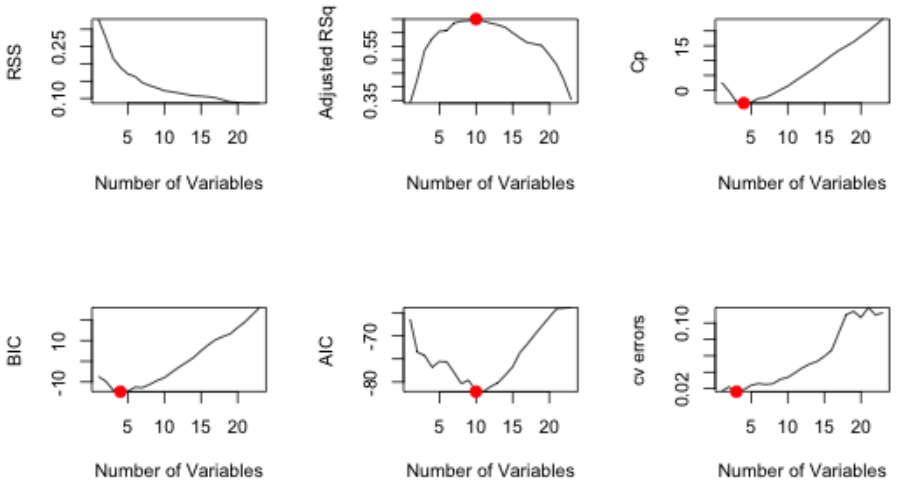
and BIC are plotted in figure 2.

Since Mallow Cp, AIC, and BIC may not actually reflect the test errors, selecting the best model based only on these criteria may therefore not be appropriate. We then perform the 10-fold cross validation to select the best model. As there is no built-in package in R to do the 10-fold cross validation, we did that procedure manually. First, we randomly divide the training set into 10 folds. Then, we run the *regsubset()* function on each fold to get the best model for each number of variables. After that, we use the cross validation sets for calculating the CV errors for each fold. Note that there is also no built-in R function to do the prediction for subset selection and we manually defined a function called *predict.regsubsets()*. In the end, we take the average across the different folds for each μ_k with $k = 1, \ldots, 23$. The 10-fold CV errors are plotted together with RSS, adjusted R2, Mallow Cp, AIC, and BIC for the sake of comparison in figure 2.

As the 10-fold cross validation has randomness in creating the folds (RP-2), we conducted a form of a "sanity check" to be certain that the best number of variables is appropriate. we reiterated the calculation of the 10-fold CV error 100 times with a different randomization setting (in the R software, we set a new "seed"). It turns out that the best number of variables is always around 2 after 100 times simulation, which is consistent with our original analysis.
Based on the best model selected by 10-fold cross validation, we compute the test MSE using the test set. The resulting model has 2 variables $Pik3r3$ and $Rac1$ (excluding the intercept).

$$Mapk1 = 0.5258 + 0.3142Pik3r3 + 0.577Ppp3cb + 0.3084Rac1 + \varepsilon$$

Finally, we compute the MSE on the test data and arrive at a value of 0.0865. We will later compare this test MSE to those of the other models.

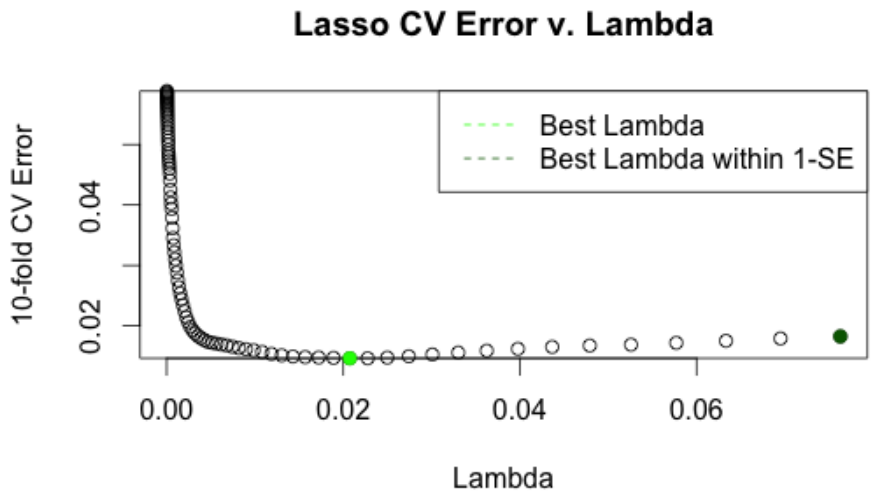


plots.png

Figure 2: RSS, AdjR2, Mallow Cp, AIC, BIC, CV errors

3 Lasso Regression

Next, we consider a lasso regression because, similar to the best subset selection method, it would help with variable selection by setting the coefficients of certain parameters to 0. A key aspect of defining a lasso regression is determining a suitable tuning parameter, λ^{lasso}. Estimating such a model on the training data yields a tuning parameter of 0.0207. Figure 3 below compares 10-fold CV errors across different values of lambda on the training set.

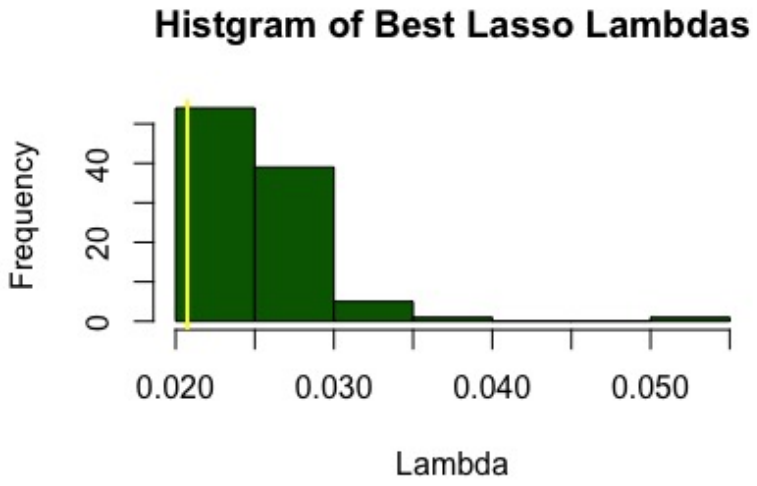


CV-Lambda.png

Figure 3: Comparing Lambda Values by CV Errors

The lasso method selects the best λ^{lasso} on the basis of the lowest 10-fold CV error. As discussed before, the division of 10 folds is randomized and therefore, there is variation in the CV error (referred to as RP-2). So we conduct a "sanity check" check similar to that in the previous section to see whether the λ^{lasso} value of 0.0207 is appropriate. To check this, we reiterated the calculation of the 10-fold CV error 100 times with

a different randomization setting (in the R software, we set a new seed). The "best" λ^{lasso} value for each iteration is plotted as histogram is figure 4.



sanity check.jpeg

Figure 4: "Sanity Check" for value of lambda

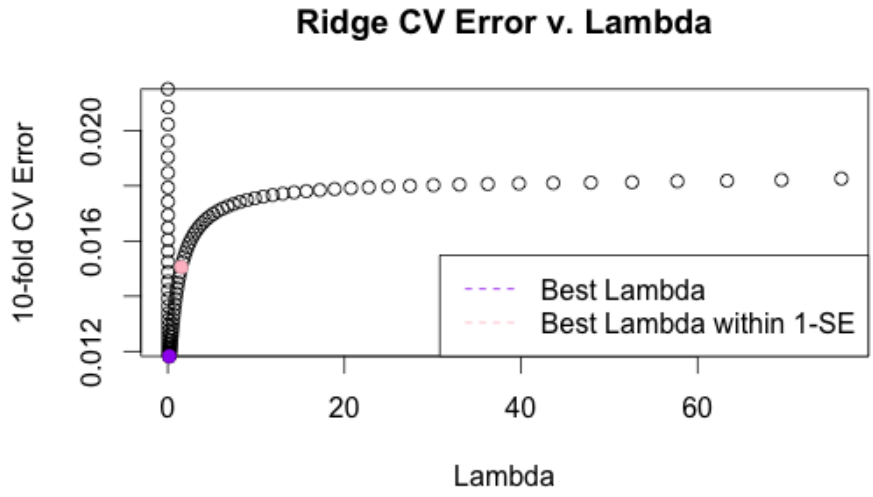
The vertical line in the figure represents the λ^{lasso} value that was initially indicated. Since it lies near the center of the distribution, we can conclude that this value is appropriate. In the process of this analysis, we considered computing the best λ^{lasso} by finding the mean of this histogram distribution. However, we choose to use this simply as a check because in regular practice, it is not feasible to determine the best tuning parameter in this way. We have the privilege of doing this check because our sample size is small (n=40) and therefore, the reiterated calculation is not computationally expensive.

It must be noted that for this analysis, we decided that the λ^{lasso} associated with the lowest CV error would be "best," as opposed to the tuning parameter associated with the lowest CV error *within 1 standard error* of the actual lowest CV error. One of the reasons for this is that the former tuning parameter already imposes a fairly large penalty. As a result, we find that for many iterations of RP-2, the best models include only 1 parameter, which is the intercept. Therefore, imposing an even larger penalty using the 1-SE rule would not result in a much simpler model.

Based on the selected value of λ^{lasso} , we compute the MSE on the test set. The resulting model has only 1 parameter, which is clearly a very sparse model due to the harsh penalty lasso typically imposes. We therefore turn our attention to less harsh models.
The test MSE is 0.0081, which is considerably lower than that of the best subset selection method.

4 Ridge Regression

We now turn our attention to the ridge regression. We acknowledge that such a method does not help us perform variable selection. In addition, it will also likely suggest a more complex model than those previously selected. Despite these limitations, the ridge regression model may be suitable if we find that many of the other genes are, at least to some extent, associated with the $Mapk1$ outcome. In other words, it is possible that this method may perform better than those previously discussed because it can account for more relationships between the genes. We therefore believe that our analysis would be remiss without considering this method.



CV-Lambda.png

Figure 5: Comparing Lambda Values by CV Errors

Following the same steps described in previous sections, we determine a λ^{ridge} tuning parameter based on the training data. Once again, we do not apply the 1-SE rule as the difference between the two tuning parameters within 1 standard error is very small (see figure 5). The selected value is 0.1644. Over 100 iterations of RP-2, we find that this value lies close to the center of the distribution of possible "best" tuning parameters (see Appendix). Therefore, we accept this value of λ^{ridge} and proceed with further analysis. We find that the model selected had non-zero coefficients for all 23 predictors, albeit the magnitude of these values is small. The are presented in the table 4 of the Appendix.

Based on the selected model with λ^{ridge}, we find that the MSE on the test validation set is 0.0057. At this point, we can see that the ridge and lasso methods perform better than the best subset selection method.

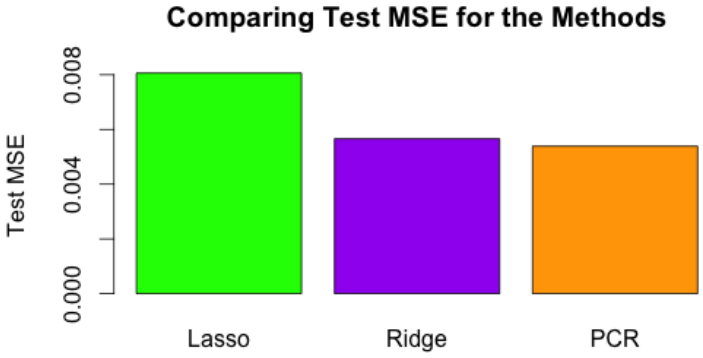
5 Principal Component Regression

The final method that we evaluate is the PCR. As with ridge regression, this method does not perform variable selection. However, it does perform dimension reduction, which is may be a powerful tool in this context. Since there are many genes that may be strongly associated with the outcome, both statistically and conceptually (from a biological perspective), we think it is important to test the performance of a model that incorporates linear combinations of many gene expressions. In the previous section, we saw that the ridge regression, which uses all the gene predictors, albeit with some penalty, performed fairly well. This lends support to the idea that PCR could also perform well because it would allow for a relatively simple model, while also making use of many predictors to explain the variance in $Mapk1$.

Our estimates based on the training data indicates that a model with 4 principal components gives the lowest 10-fold CV error. As done in the previous sections, we reiterate this process 200 times and select too best M components (see figure 10 in Appendix). Once again, we find that a value of M=4 lies near the center of the distribution. We therefore choose this value of M and compute the error on the test validation set to be 0.00539. Note that we do not standardize the data in this PCR model because all the variables are already measured in the same way as gene expressions.

In the figure below, we can see that the lasso, ridge, and PCR methods are associated with far lower test

MSEs than the best subset selection. When we compare only these three methods (as done in figure 6), we find that the PCR method has the lowest test MSE.



lasso,ridge,pcr-1split.png

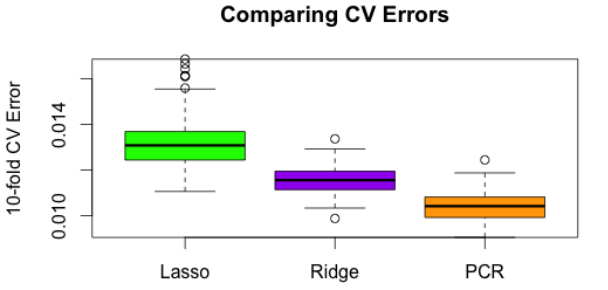
Figure 6: Comparing TMSEs for a single split of data

Although test set MSEs are useful and indicate that ridge and PCR models may be the most suitable, we recognize that this comparison is not enough to make a confident claim. In our understanding, the validation set approach has considerable variability as it depends on the "split" of the data into training and test sets (RP-1). To address this limitation, we present further model comparisons in the section below.

6 Comparison of the Methods

6.1 Comparing 10-fold CV error on many iterations of differently randomized folds

As discussed throughout this report, we compute many iterations of the "best" values of k in $\mu_k.\lambda^{lasso}$ and λ^{ridge}. These values are plotted in figures 3 and 5 above. In figure 7 below, we present boxplots of the 10-fold CV corresponding to each of the "best" parameters mentioned above. We consider this analysis necessary because the with the division of folds in is variable (RP-2). The figure shows that PCR has the lowest 10-fold CV error.



CV errors.png for 100 randomization processes for lasso/ridge and 200 for PCR

Figure 7: Comparing CV Errors over Iterations of RP-2

6.2 Comparing the test MSE on 100 iterations of differently randomized training-test splits

Although the cross-validated error is a powerful measure of prediction error, it is based solely on the training data and is therefore, not the ideal final criterion for model comparison. We consider it necessary to compare test prediction error to have a complete understanding of how the models compare to one another. In figure 8 , we compare the MSE on the test set from the *initial* splitting of the data. We now consider what it means for this initial splitting process to be random (RP-1). Because the data set is fairly small (n=40), we are feasibly able to reiterate the training-test split 100 times. With the "best" values of k in μ_k, λ^{lasso},λ^{ridge}, and M selected from the previous analysis, we proceed to estimate the test MSE 100 times for each model.

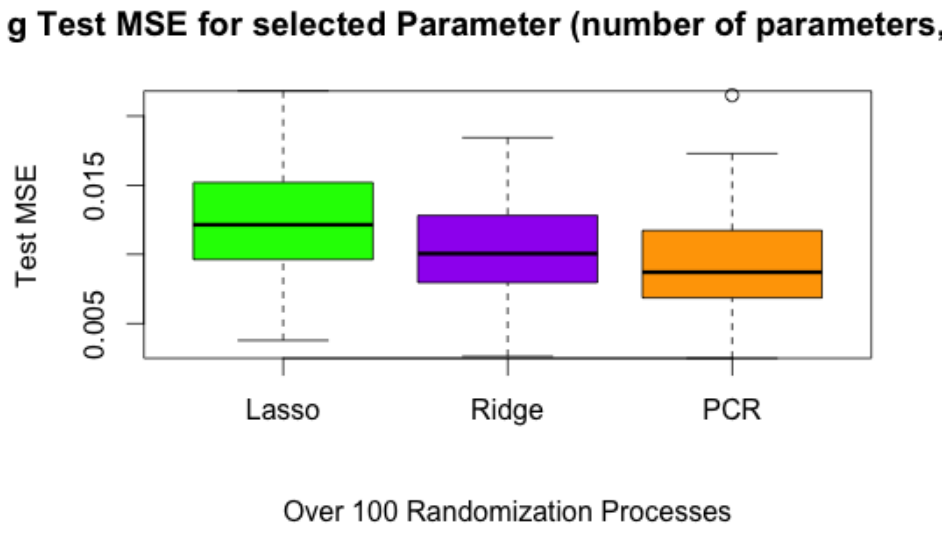


Figure 8: Comparing TMSEs over Iterations of RP-1

In the figure 8, we find that PCR has the lowest test MSE, but ridge has a comparable test MSE.

7 Conclusion

We find that the PCR is the best in terms of the test MSE. As the PCR performs the dimension reduction. Therefore, it is able to incorporate multiple variables while simultaneously reducing the variance. The correlation plots (see Appendix) show a high association among the different genes, suggesting that they are interlinked. So it makes sense that PCR, which uses many variables to explain the variation in $Mapk1$ has the most superior performance in terms of the prediction power.

Despite the good PCR performance, we also note that the ridge model has a comparable and fairly low 10-fold CV and test MSE. This is noteworthy because the ridge model as considerable strengths over the PCR model in terms of its interpretability. The coefficients of ridge model provide a penalized direct estimate of the *ceteris paribus* association between the genes.
If we would want to make use of associated genes to make predictions on the target gene, the PCR model should be more powerful in prediction performance. If we focus more on interpreting the associations among the target gene and other genes, the ridge model can provide a better interpretation.

Appendix

Table 1: Variable Means

Cdc42	Pla2g6	Akt2	Plcg2	Mapk1	Rac2	Rik	Mapkapk2
-0.05	0.21	-0.55	0.67	0.17	-1.43	1.45	-1.07
Pik3cd	Pla2g5	Sphk2	Map2k1	Pik3r3	Ptk2	Nras	Nos3
0.57	0.13	0.81	0.03	0.87	-0.29	-0.30	-1.88
Pik3r1	Pik3ca	Ppp3cb	Map2k2	Nfatc4	Mapk13	Rac1	Nfat5
-0.99	0.13	-0.83	-0.12	1.46	1.14	-0.45	1.62

Table 2: Variable Standard Deviations

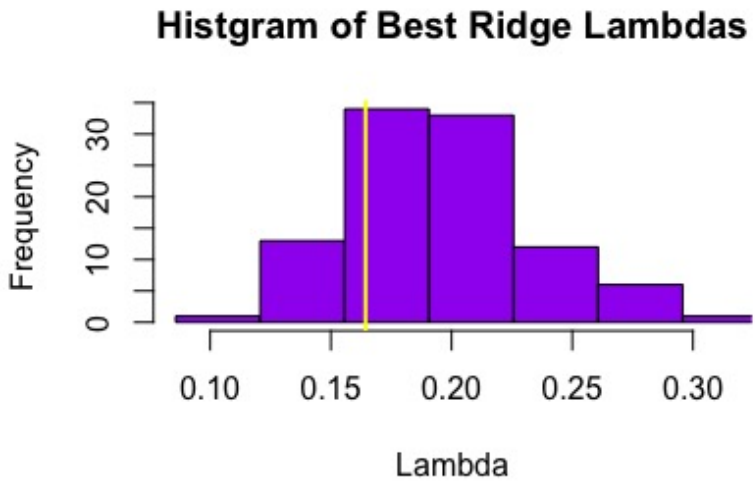
Cdc42	Pla2g6	Akt2	Plcg2	Mapk1	Rac2	Rik	Mapkapk2
0.09	0.20	0.08	0.14	0.13	0.17	0.18	0.22
Pik3cd	Pla2g5	Sphk2	Map2k1	Pik3r3	Ptk2	Nras	Nos3
0.18	0.17	0.20	0.11	0.16	0.07	0.10	0.23
Pik3r1	Pik3ca	Ppp3cb	Map2k2	Nfatc4	Mapk13	Rac1	Nfat5
-0.14	0.07	0.09	0.12	0.09	0.35	0.17	0.14

Table 3: Best Subset Selection Model

	Coefficient	Std. Error	t value	p-value
Intercept	0.5258	0.2363	2.225	0.03430*
Pik3r3	0.3142	0.1024	3.069	0.00473**
Ppp3cb	0.5770	0.2085	2.767	0.00991**
Rac1	0.3084	0.1025	3.009	0.00549**

Table 4: Ridge Regression Model (Coefficients Only)

Intercept	-0.071910679
Cdc42	0.042179421
Pla2g6	0.014992862
Akt2	-0.043831003
Plcg2	0.038318820
Rac2	-0.036493596
Rik	0.042208164
Mapkapk2	-0.016795923
Pik3cd	-0.040691433
Pla2g5	-0.019706383
Sphk2	0.014137448
Map2k1	0.052870017
Pik3r3	0.050869743
Ptk2	0.039006080
Nras	0.017951470
Nos3	0.002665847
Pik3r1	-0.023654937
Pik3ca	0.014952254
Ppp3cb	0.052714022
Map2k2	0.014861960
Nfatc4	0.015130791
Mapk13	-0.004848481
Rac1	0.044434043
Nfat5	0.037665012



sanity check.jpeg

Figure 9: ”Sanity check” for lambda value (Ridge)

sanity check.jpg

Figure 10: ”Sanity check” for M components (PCR)

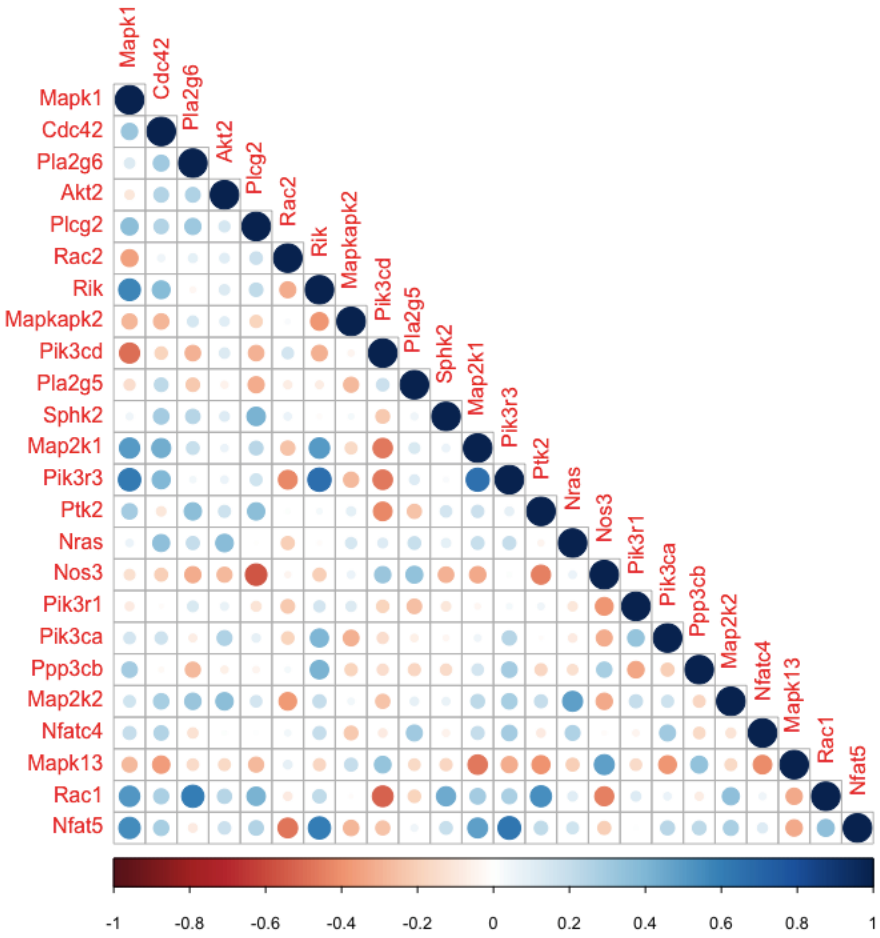


Figure 11: ”corelation plot”