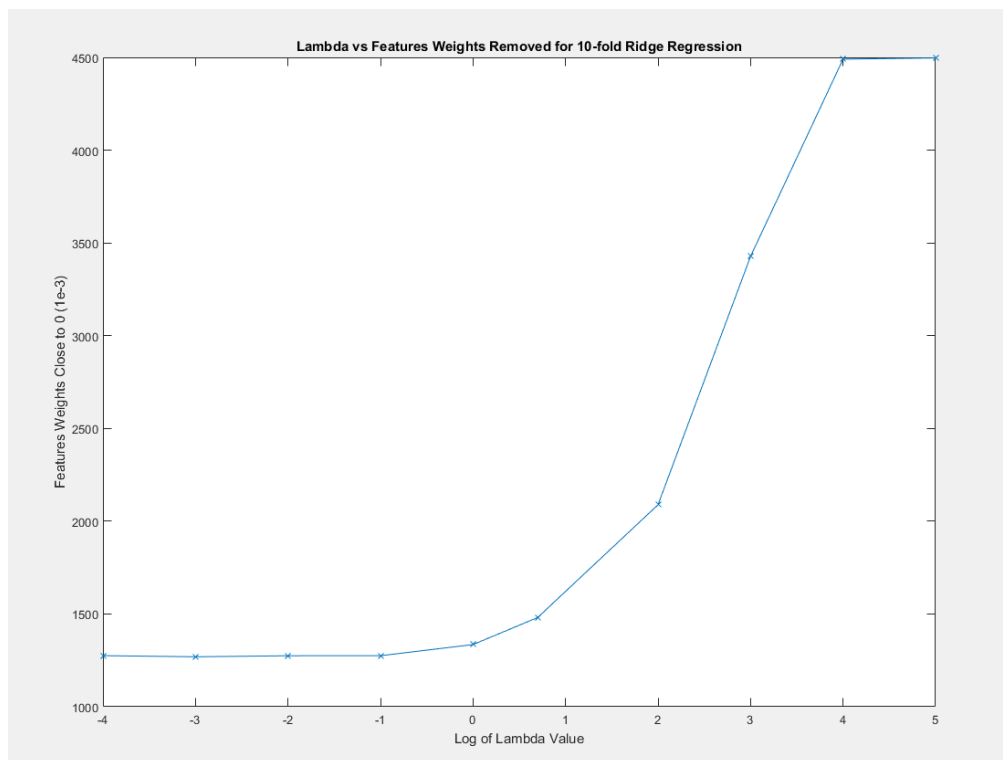
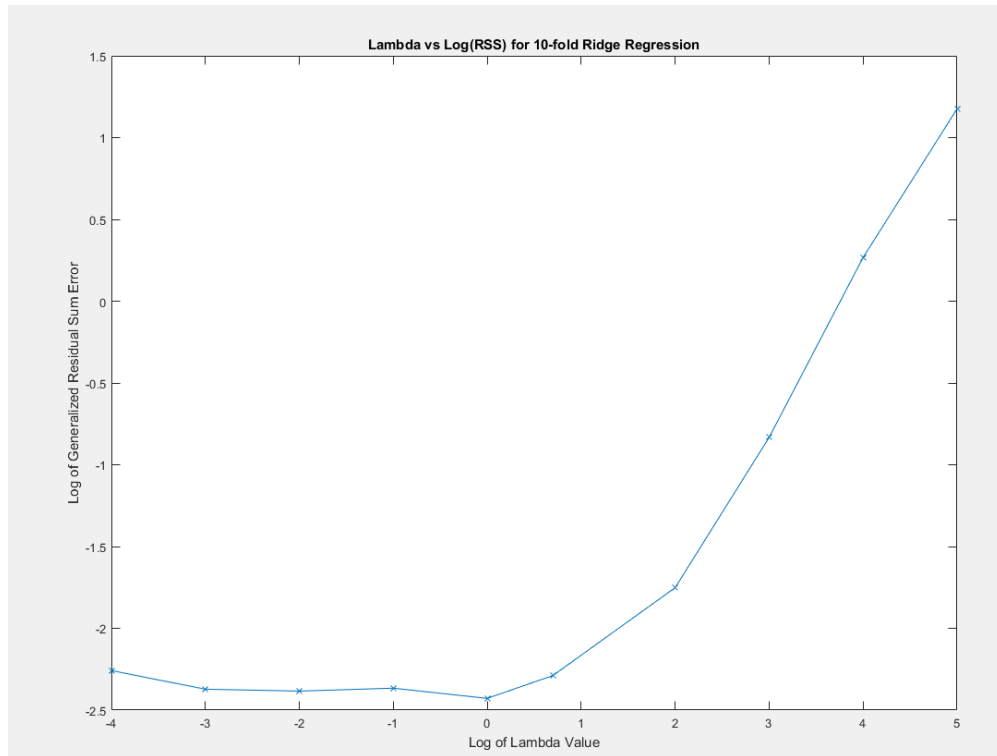
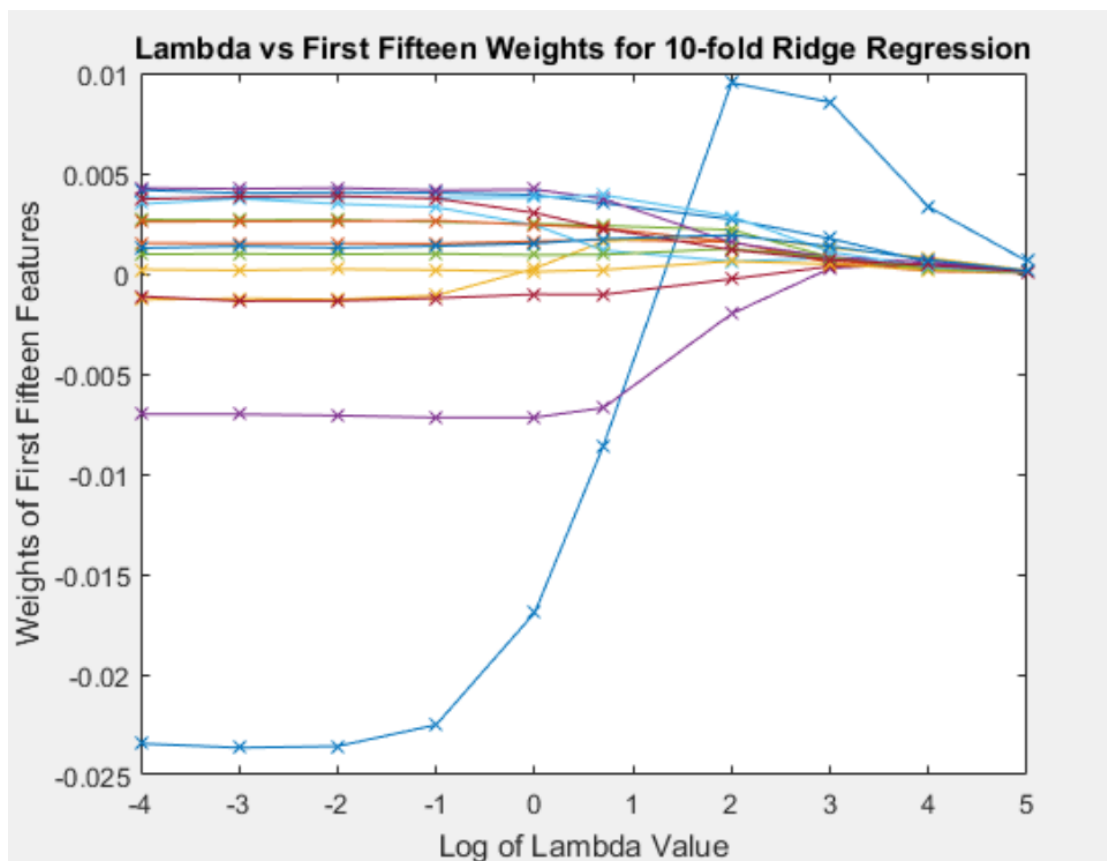


● Q1

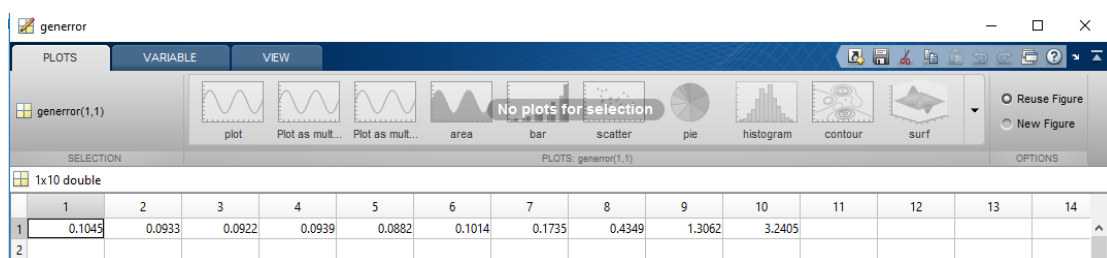




From the first plots, we will choose $\lambda = 1$ as our optimal constrained parameter value which minimize the residual sum error.

From the second plots, we see that there are around 1350 parameters closed to 0. One shuffled data report 1371 parameters to be closed to 0, this indicated that under λ equal to 1, there are $4497 - 1371 = 3126$ non-zero coefficients.

The 10-fold cross validation generalization error is reported by following graph.



● Q2

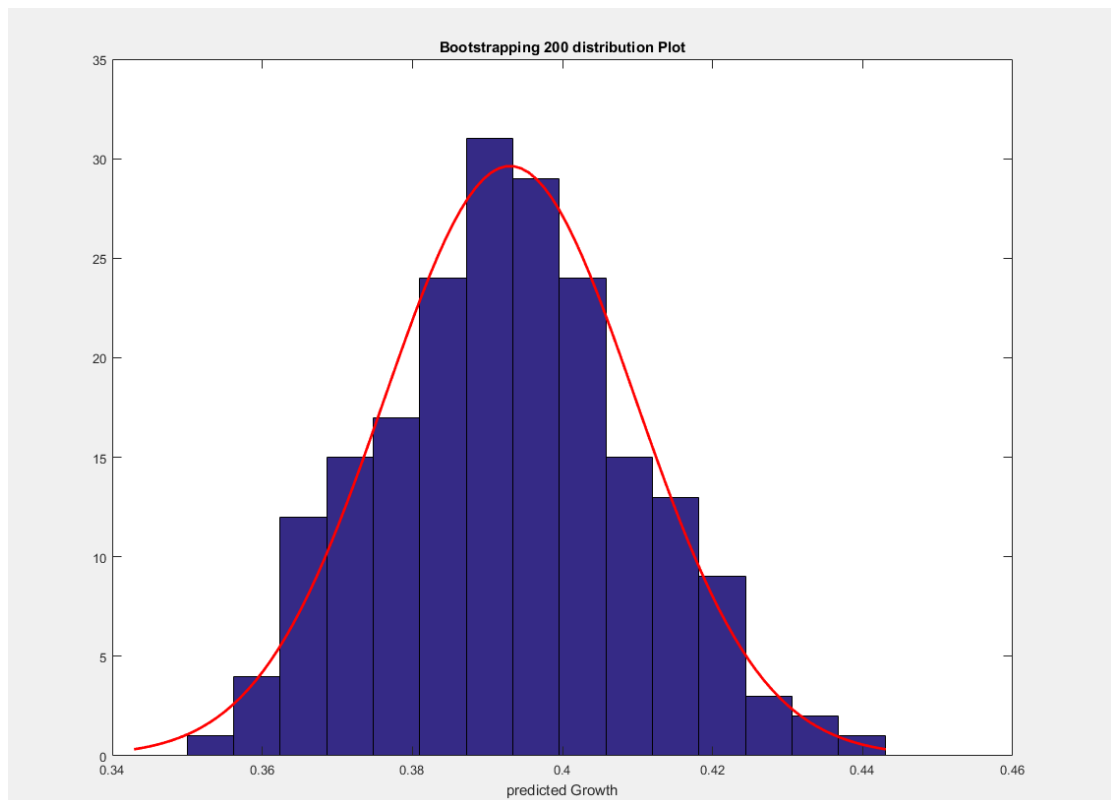
So, I bootstrap the dataset for 200 times (large dataset, unable to do more iteration) with replacement. Then I will use these W_{boot} along with new observations to predict 200 Growthrate, then check if the distribution of these Growthrate like any distribution. Then choose the method to apply Confidence Interval to the predict.


By using the non-parametric bootstrap, we don't need to assume normal distribution for the dataset

using CLT($n > 30$). Sometimes, this assumption might cause problems. We need very little assumption that the bootstrap dataset with n time iteration can represent the population. This is true when n is larger. Then, we can estimate the standard error of prediction thus give us confidence interval.

- Q3

By using the 90% confidence level, The distribution of bootstrap predict seems normally distributed.(Shown below) So, I just applied percentile method to get the confidence interval.

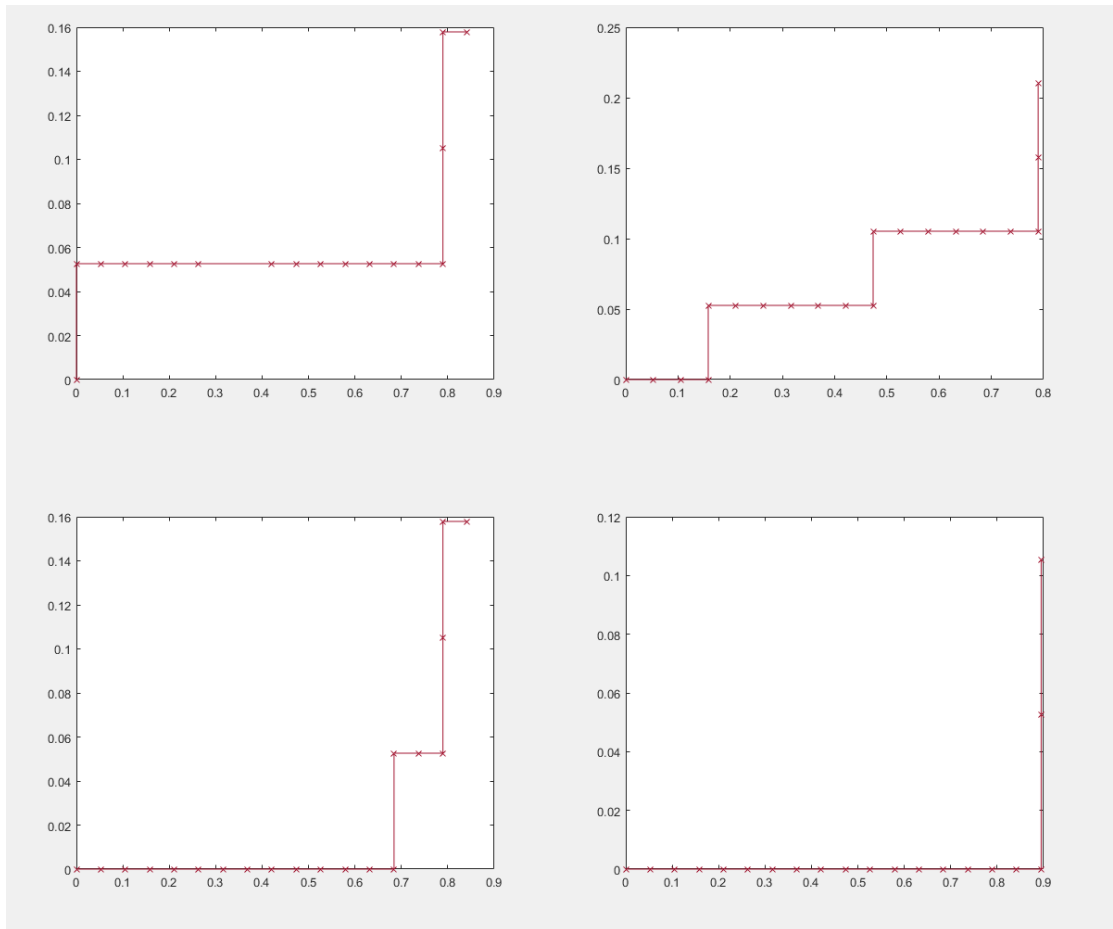


Estimated prediction:  predict 0.3929

Lower Limit:  cilow 0.3688

Upper Limit:  ciup 0.4157

- Q4



The plot above shows ROC curves for Strain type, medium type, environmental and gene perturbation. (subplot order)

- Q5
The generalized accuracy is around 7% for the composite SVM. This might due to the different combination of each composite situation is more than 30, which means we only have about 5 observations per case. So, the result might not be surprising. We better off building separate classifiers to predict features.
- Q6
From the following plot, PCs have great performance compared to using other method. So we can say it retain most of the classification performance while reducing the dimensionality.

