Final Project
Stat 215A, Fall 2018

Mengling Liu

December 8, 2018

1 Data Partition

I splitted the data into three parts: a training set, a validation set, and a test set. The training set is used to fit the models, the validation set is used to estimate prediction error for model selection, the test set is used for assessment of the generalization error. A typical split is 50% for training, 25% each for validation and testing set.

2 Model Selection

First, I fitted a LASSO regression model and used 5-fold CV, ESCV, AIC, AICc and BIC for selecting the smoothing parameter in Lasso. Second, I fitted a Ridge regression model and used 5-fold CV, ESCV, AIC, AICc and BIC for selecting the smoothing parameter in Ridge as well.

Now we want to undestand the weakness and strength of each criteria selection method.

1. Cross validation: The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. Also the advantage of this method (over k-fold cross validation) is that the proportion of the training/validation split is not dependent on the number of iterations (folds). The disadvantage of this method is that some observations may never be selected in the validation subsample, whereas others may be selected more than once. In other words, validation subsets may overlap.

2. ESCV: ESCV finds a smaller and locally ES-optimal model smaller than the CV choice so that the it fits the data and also enjoys estimation stability property. We demonstrate that ESCV is an effective alternative to CV at a similar easily parallelizable computational cost. In particular, we compare the two approaches with respect to several performance measures when applied to the Lasso on both simulated and real data sets. For dependent predictors common in practice, our main finding is that, ESCV cuts down false positive rates often by a large margin, while sacrificing little of true positive rates. ESCV usually outperforms CV in terms of parameter estimation while giving similar performance as CV in terms of prediction.

3. AIC, BIC and AICc: AIC and BIC are appropriate for different tasks. In particular, BIC is argued to be appropriate for selecting the ”true model” (i.e. the process that generated the data) from the set of candidate models, whereas AIC is not appropriate. To be specific, if the ”true model” is in the set of candidates, then BIC will select the ”true model” with probability 1, as n goes to infinity. The reason is that, for finite n, BIC can have a substantial risk of selecting a very bad model from the candidate set. This reason can arise even when n is much larger than k2. With AIC, the risk of selecting a very bad model is minimized.

AICc has the advantage of tending to be more accurate than AIC (especially for small samples), but AICc also has the disadvantage of sometimes being much more difficult to compute than AIC. Note that if all the candidate models have the same k and the same formula for AICc, then AICc and AIC will give identical (relative) valuations; hence, there will be no disadvantage in using AIC, instead of AICc.

3 Correlation Calculation

Calculate the correlation between fitted values and observed values based on the predictor for all 20 voxels. Correlation calculation for Lasso regression:

```
    X corr_lasso_cv corr_lasso_escv corr_lasso_aic corr_lasso_bic
1   1     0.4199911     0.440243494     0.28687697      0.4168758
2   2     0.5017827     0.503386289     0.39104262      0.4945761
3   3     0.4500280     0.459503895     0.30970208      0.2407672
4   4     0.4760970     0.480354001     0.31718097      0.4462633
5   5     0.4473752     0.446342633     0.29887635      0.4449211
6   6     0.4360905     0.444573734     0.36404005      0.3863909
7   7     0.4970125              NA     0.42528560      0.4878502
8   8     0.4658905     0.463779906     0.33117879      0.4340256
9   9     0.5492299     0.551363796     0.43062492      0.5410281
10 10            NA     0.180216425     0.07991090      0.1721067
11 11     0.2927225     0.307507379     0.13117281             NA
12 12     0.4159186     0.422816101     0.24633339      0.3879672
13 13     0.1934777     0.215008238     0.18046235      0.1508508
14 14            NA     0.191621477     0.11192226             NA
15 15     0.4569814     0.461336973     0.35098245      0.4185456
16 16            NA     0.008379505     0.05389700             NA
17 17     0.3822960     0.380557870     0.16384265             NA
18 18     0.5034478     0.505995029     0.35138286      0.4985118
19 19     0.3475929              NA     0.20199069      0.2308356
20 20            NA              NA     0.06866961             NA
   corr_lasso_aicc
1       0.29240836
2       0.49940338
3       0.31153275
4       0.32256115
5       0.29887635
6       0.36586427
7       0.42833005
8       0.33117879
9       0.43190560
10      0.08809051
11      0.14385967
12      0.24817754
13      0.18571137
14      0.12467709
15      0.35758891
16      0.06643915
17      0.17004958
18      0.51024170
19      0.20534942
20      0.05788650
```

Correlation calculation for ridge regression:

```
    X corr_ridge_cv corr_ridge_escv corr_ridge_aic corr_ridge_bic
1   1    0.47201250     0.440243494     0.28687697      0.4168758
2   2    0.51864477     0.502251494     0.39104262      0.4945761
3   3    0.46799590     0.459503895     0.30970208      0.2407672
```

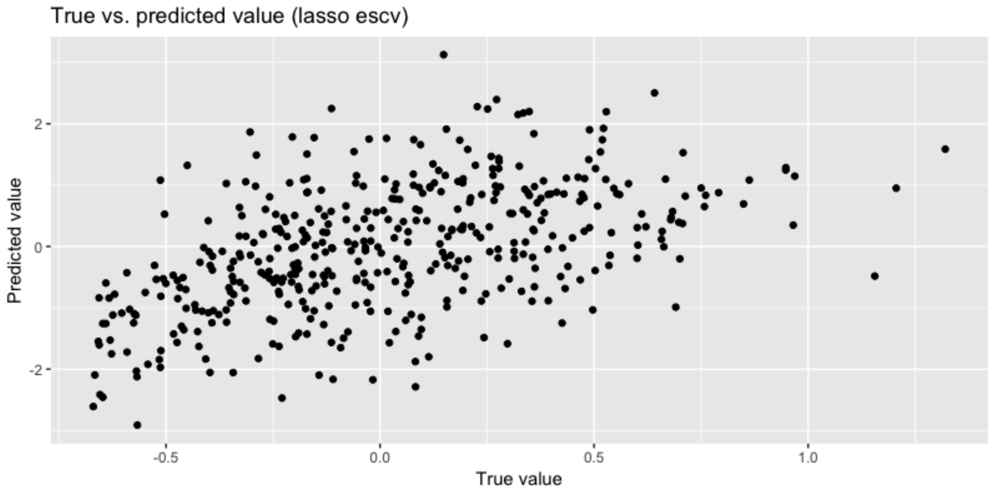
```
4   4    0.44802022     0.480354001     0.31718097      0.4462633
5   5    0.44749209     0.446574242     0.29887635      0.4449211
6   6    0.45131897     0.440049410     0.36404005      0.3863909
7   7    0.48859482     0.507772043     0.42528560      0.4878502
8   8    0.47544394     0.463779906     0.33117879      0.4340256
9   9    0.50513896     0.551363796     0.43062492      0.5410281
10 10    0.15375008     0.175350686     0.07991090      0.1721067
11 11    0.20089666     0.305791676     0.13117281             NA
12 12    0.35184807     0.422816101     0.24633339      0.3879672
13 13    0.16100514     0.215008238     0.18046235      0.1508508
14 14    0.21748065     0.191435053     0.11192226             NA
15 15    0.44976487     0.461336973     0.35098245      0.4185456
16 16    0.03852390     0.008379505     0.05389700             NA
17 17    0.31466673     0.380557870     0.16384265             NA
18 18    0.42292096     0.506115944     0.35138286      0.4985118
19 19    0.31450312     0.368302642     0.20199069      0.2308356
20 20    0.08858058              NA     0.06866961             NA
   corr_ridge_aicc
1       0.29240836
2       0.49940338
3       0.31153275
4       0.32256115
5       0.29887635
6       0.36586427
7       0.42833005
8       0.33117879
9       0.43190560
10      0.08809051
11      0.14385967
12      0.24817754
13      0.18571137
14      0.12467709
15      0.35758891
16      0.06643915
17      0.17004958
18      0.51024170
19      0.20534942
20      0.05788650
```

4 Model Diagnostic

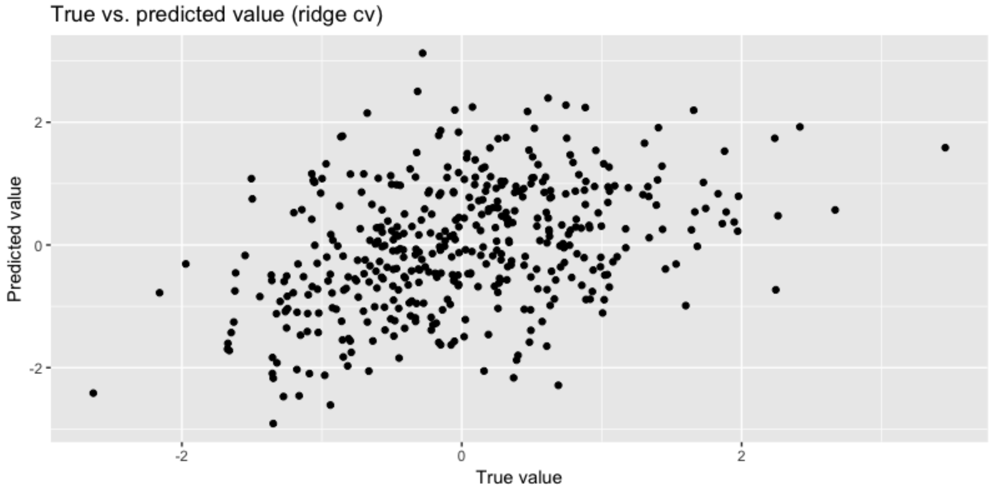
Now I chose a couple of models to further investigate the fit of models. Are there any outliers? Any further discussion on the stability of the prediction results and of the models?

I chose voxel 2 as an example, so for voxel 2, I chose the model with the highest correlation value, which is the Lasso model with escv as a selection criteria. I also conducted similar analysis on ridge regression as well. Ridge model with cross validation as a selection criteria gave the highest correlation value.

Firstly, I wanted to check on the outliers of both models I plotted the true value of y against the predicted value of y from both models. The observation is that most of the dots follow along the ”y=x” line for both cases.

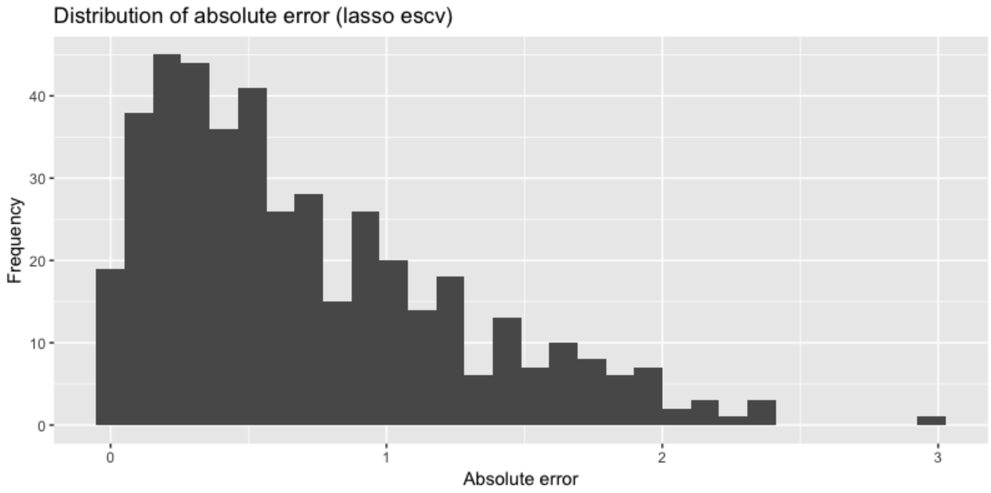


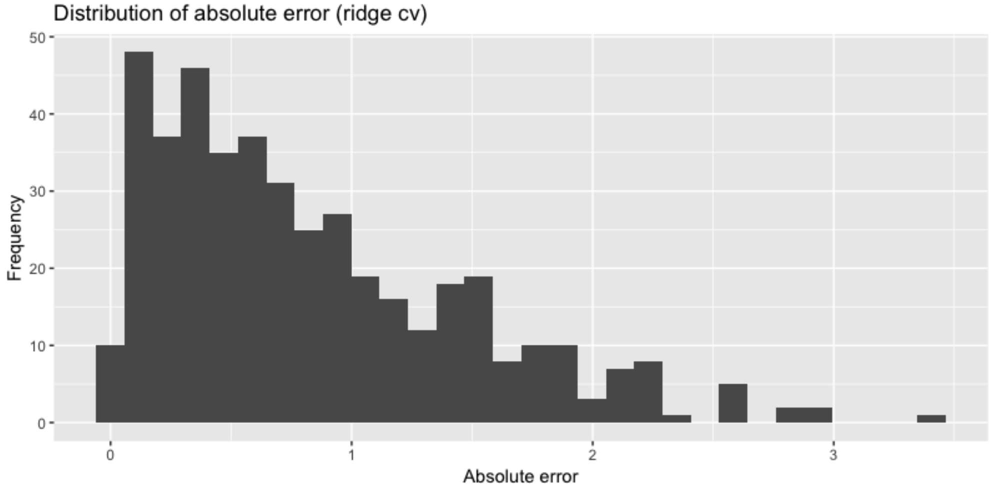
True vs. predicted value (lasso escv)



True vs. predicted value (ridge cv)

Secondly, I also plotted the distribution of absolute error for both ridge and lasso models. From both sets of graphs, we can conclude that both models generated relatively stable prediction results. The absolute errors are controlled within the range from 0 to 3.

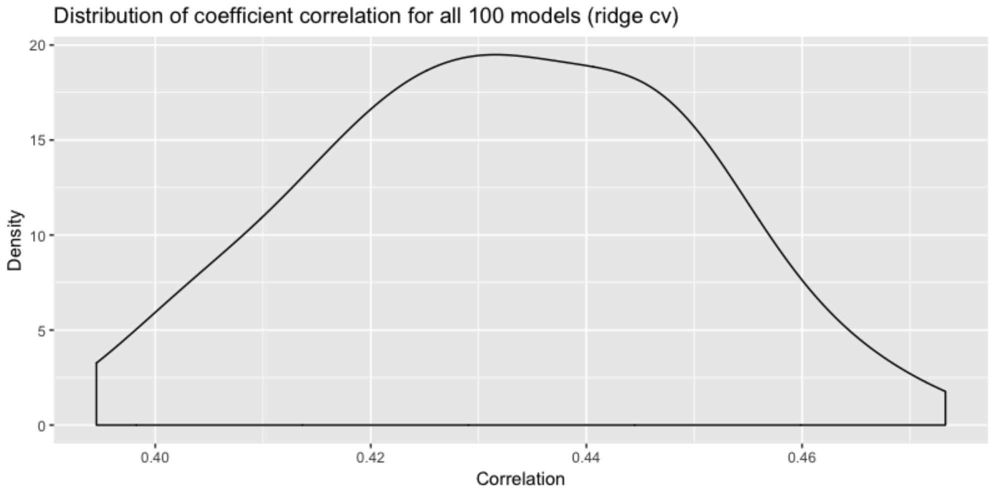

Distribution of absolute error (lasso escv)

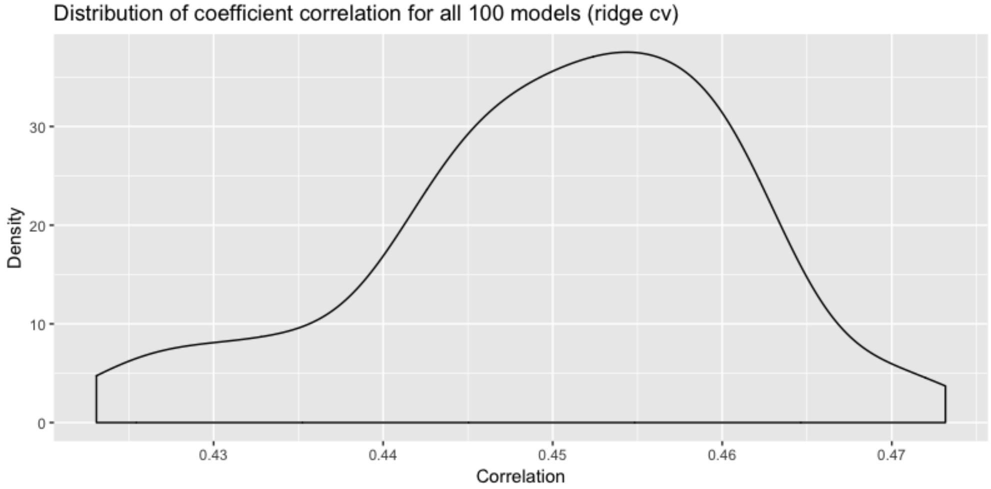

Distribution of absolute error (ridge cv)

Now I wanted to deep dive into the features. Do the models share any features? Which predictors are important for which voxels? What features are stable across different bootstrap samples? What can be learned about how voxels respond to images?

In order to test the stability of the prediction results of the models, I want to adopt the bootstrap methodology. I first sample from the training data sets with replacement for 100 times, and then fit the lasso and ridge models. I then check the prediction results for each of the 100 models. Here for lasso model, I chose the one with the highest correlation value - escv criteria. For the ridge model, I also chose the model corresponding with the highest correlation value - cv criteria.

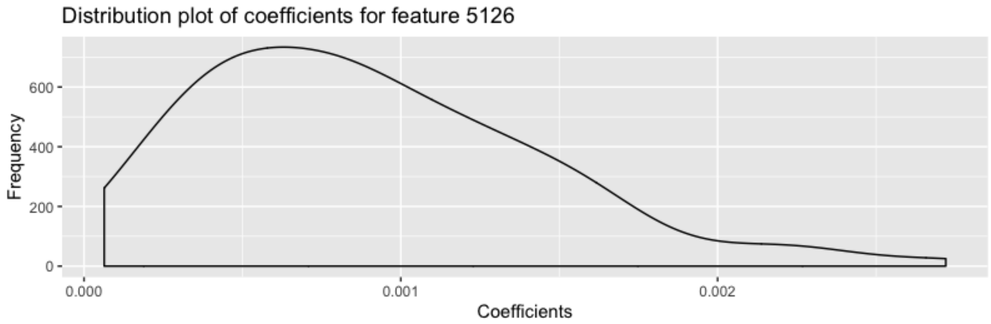
In order to undestand whether the lasso and ridge models are stable, I examined the distribution of correlation among all coefficients for all 100 models. I noticed that for both models, the correlations fall largely in the range around 0.45, indicating that the model is quite stable.

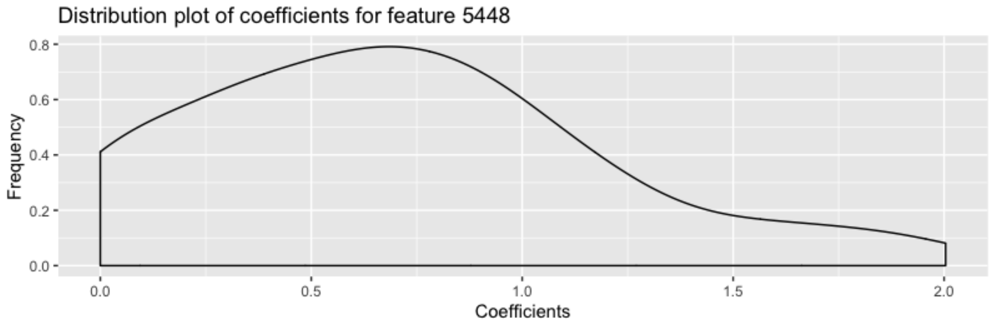


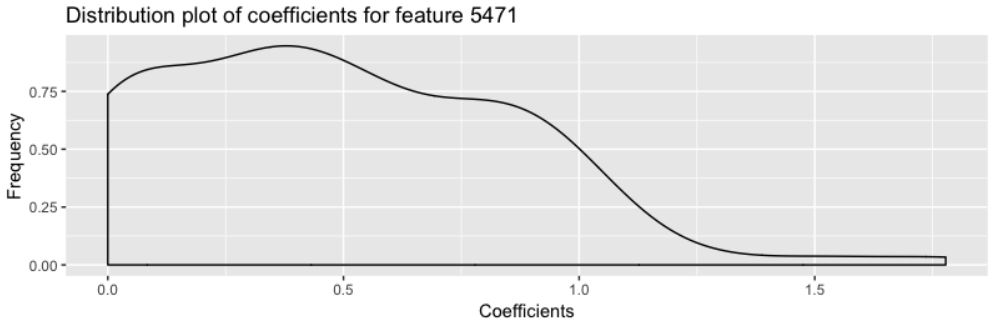


In order to understand whether the models share any features and which predictors are important for which voxels. Lasso regression can perform variable selection. I extracted all the 10,921 coefficients for all the 100 models. I defined important predictors as the ones with high occurance of non-zero coefficients and here I used the threshold of 60%: I counted the total number of coefficients with more than 60 models showing non-zero values. The final output shows the coefficients that fall above this threshold, and the corresponding number of times non-zero values occur for those coefficients. Under lasso regression model with escv as the selection criteria, we observe 3 important predictors: feature 5126 with 64 non-zero values, feature 5448 with 93 non-zero values and feature 5471 with 86 non-zero values.

I also wanted to understand what features are stable across different bootstrap samples under lasso regression model, so I plotted the distribution of coefficients for important features selected using 60% threshold. From the graph, we observed that the features are stable across different bootsrap samples as the variance of the distribution is relatively small and the plot is long-tail right-skewed distribution with majority of the data concentrated within a narrow range.







What can be learned about how voxels respond to images? We can conclude from above analysis that different features contribute to how voxels respond to images with different weights.

5 Model Fitting

I used my best lasso model to predict the response of the first voxel on the 120 images. I chose lasso model with escv as a seleciton criteria as it outputs the highest correlation among all the methods for lasso regression.