Lab Final - Stat 215A, Fall 2017

Yizhou Zhao

December 10, 2017

1 Introduction

Understanding how the human brain functions remains one of greatest challenges in medical and psychological research. Medical tools such as functional Magnetic Resonance Imaging (fMRI), which is a class of imaging methods developed to demostrate the metabolism in the brain, gives tangible ways for scientist to analysis the mechanism of visual cognition of human-beings. From fMRI response from 1750 images collect by Gallant lab, I am trying to study what parts of our brain are active in visual processing and how they interact.

2 Exploratory Data Analysis

2.1 Data Background

Gallant lab provided me with several data sets:

- **Response of brain voxels.** During the fMRI, scientists recored the response of 20 voxels from 1750 images. The response of each voxel is a continuous variable, ranging from -4.86 to 4.78.
- **Image data.** Each one of the 1750 images is a 128×128 images with gray scale.
- **Wavelet Feature data.** Gabor wavelet pyramid transform the information of the image into frequency domain. The real part of 10921 Gabor wavelet is given and another non-linear transformation extracts the 10921 features from each image.
- **Voxel locations.** The 3D coordinates of the 20 voxels in the brain. Figure 3 shows the distributions of the responses for 20 voxels. All curves look similar from normal distribution and the differences between voxels are small.

Figure 1: An example for fMRI experiment. The image has the round shape with 128 by 128 pixels with only gray scales. The images include figures, animals, natural sights and etc. All images were transformed into 16384 a vector. Figure 1 is image one rotated by 90 degree.

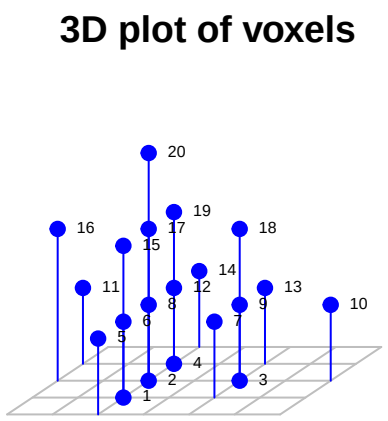

Figure 2: 3D plot of the locations of the 20 voxels. The distribution of the voxels is not exactly symmetric. Voxel 5, 6, 7, 8 are in the center, the voxels at the boarder are 10,11,16,20 and etc.

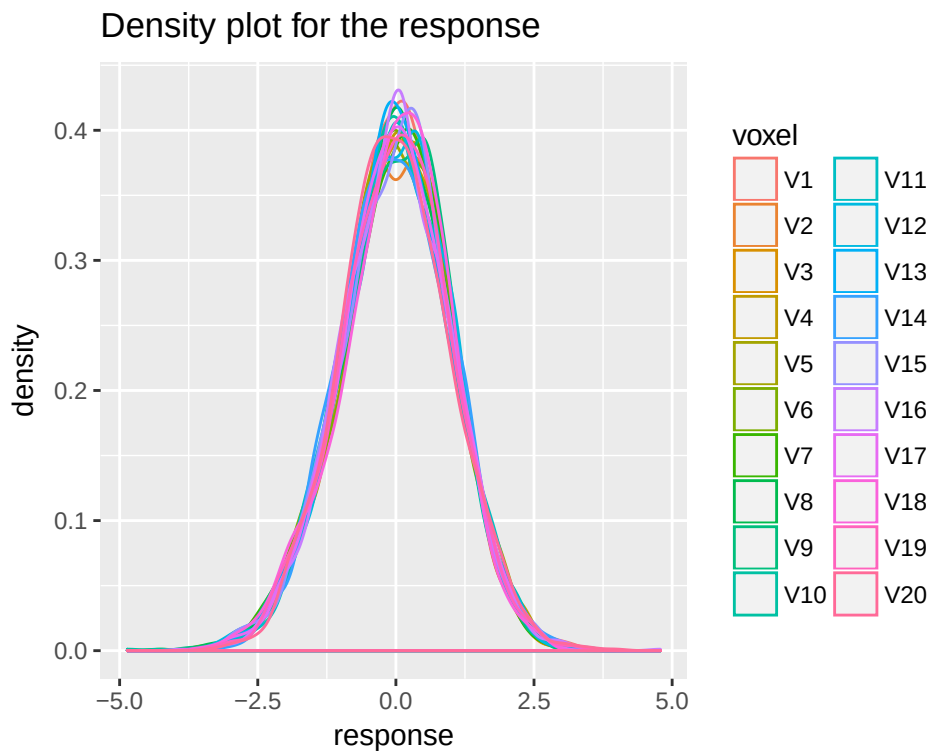

Figure 3: Histogram of the responses of 20 voxels. All the density curves look similar to each other with a little difference at the center. Thus, it is not possible to distinguish voxels only from the distributions of the response.

3 Training and analysis

3.1 Methodology

The experiment in Kay's article upon fMRI based on mainly two stages. In the first stage, model estimation, fMRI data were recorded from visual areas while each subject viewed 1,750 natural images.Gabor wavelet pyramid described tuning along the dimensions of space, orientation and spatial frequency. In the second stage of Kay experiment, scientist tried to identify images from brain's response. I will only focus on the previous stage, to discover the relationship between the features of images and voxel responses.

I decided before training these methods on a few approaches:

- **Individual lasso without feature screening**: In this approach, I will treat the response of every voxel individually with all the 10921 features included.
- **Individual lasso with feature screening**: As the same as the above method, I will treat the response of every voxel individually. However, I will apply some methods for dimensional reduction to get a pre-screening of the features. Specifically, inspired by Fan's book [3], Sure Independence Sampling will select M features with the highest M absolute correlation with responses. In my study, I select 1000 features from fit_feat.
- **Multiple lasso regression**: This model treats all the response together as a matrix and preforms a multiple regression task.

  Also several regularization methods will be considered in our model including **AIC**,**AICc**,**BIC**,**ESCV**[2] and **Mean square error**.

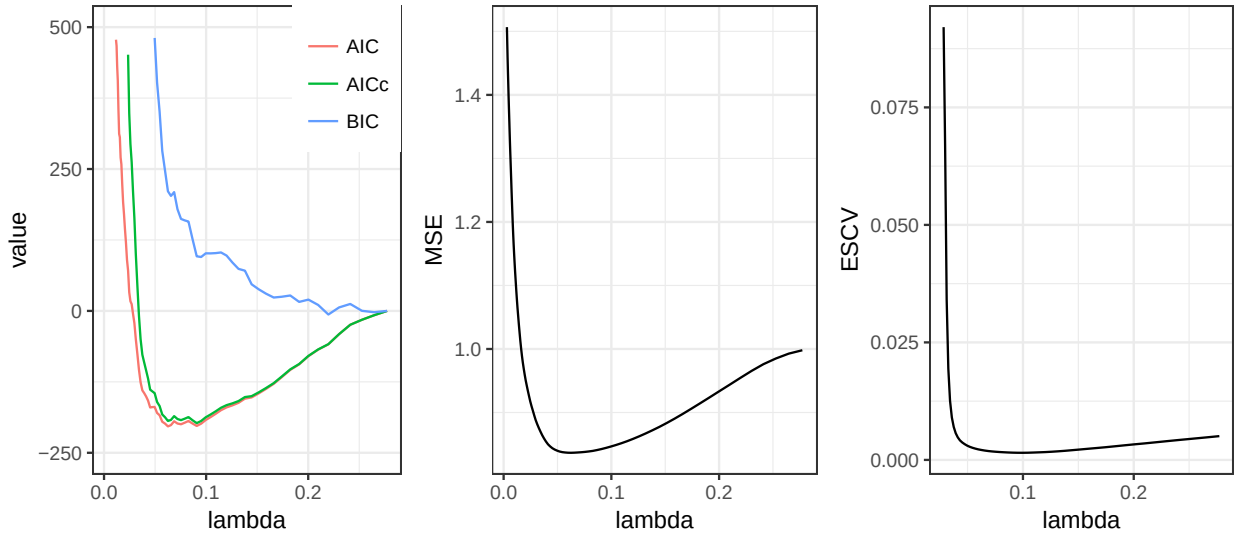

Figure 4: Validation plots for lasso regression for voxel one. The left panel shows the trend of AIC,AICc and BIC. The middle one shows the mean square error and the right one is for ES-CV. Each different lambda is examined by 10-folds validation.

3.2 Lasso

Lasso is the linear model trained with L1 penalty as regularization. The optimization objective for Lasso is:

$$\frac{1}{2n}||Y - X\beta||_2^2 + \lambda||\beta||_1$$

Lasso is able to achieve both of these goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients.

3.3 Feature Screening

Fan and Lv (2008) introduced Sure Independence Screening for variable screening via independent correlation learning that tackles ultrahigh dimensional linear models. Sure Independence Screening is a two-stage procedure.

- First filtering out the features that have weak marginal correlation with the response, effectively reducing the dimension p to a moderate scale below the sample size n. In my study, I select N features with highest N absolute correlation with responses.
- Then performing variable selection and parameter estimation simultaneously through a lower dimensional penalized least squares method such as LASSO.

3.4 Model selection

In my model, I divide my data into 1400 samples for training and validation, and the rest 350 for testing. The cross-validation process for LASSO has 100 different values for hyper parameter λ, and each validation has ten folds.

The above figure 4 gives an example of applying different model selection criteria. BIC shows a special pattern and it keeps going down when λ increases. I cannot choose BIC criterion in this case because it does not provide me with any meaningful insight for λ. The trends of AIC, AICc, MSE and ESCV are similar to each other, with the λ minimizing those criteria close to 0.06. It is worth mentioning that ESCV has a slight different pattern: dropping really fast to zero and then slowly going up as λ increases.

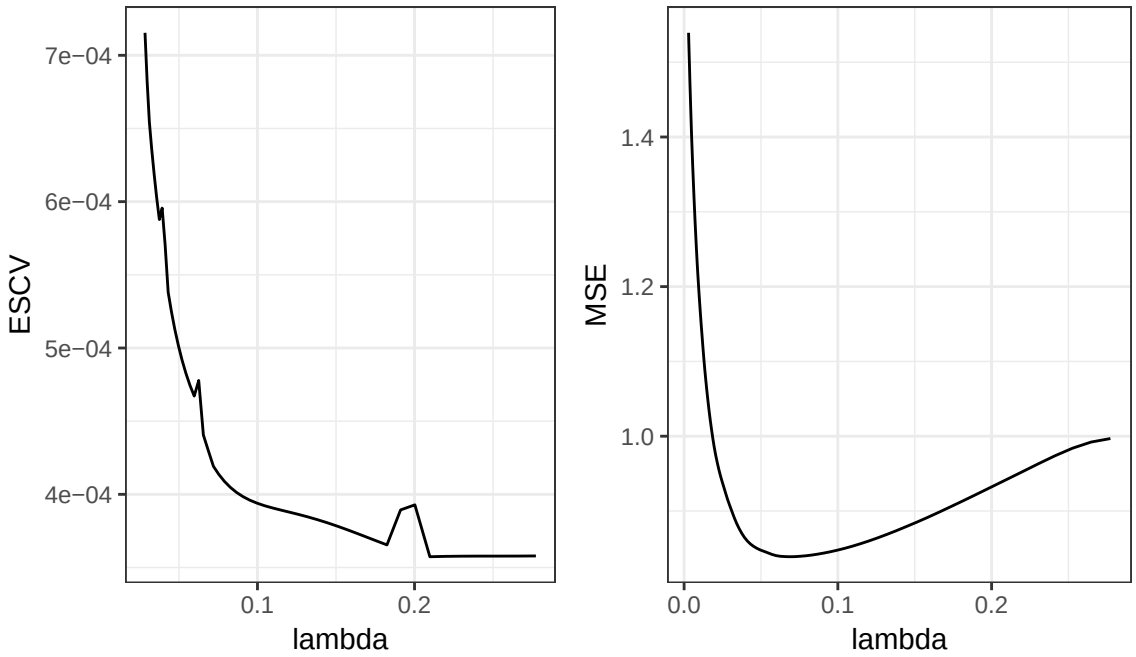

Figure 5: Ridge regression for voxel one with criterion ES-CV and MSE.

I mainly use MSE as reference; however, I am not choosing the λ that minimizes mse. We often use the "one-standard-error" rule when selecting the best model; this acknowledges the fact that the risk curves are estimated with error, so errs on the side of parsimony. We often use the "one-standard-error" rule when selecting the best model; this acknowledges the fact that the risk curves are estimated with error, so errs on the side of parsimony[4]. By following this suggestion, I choose $lambda.1se$ as hyper-parameter for the rest of my analysis.

3.5 Model Diagnosis

For comparison, I use another regularized linear model: ridge regression in my analysis. I first run ridge regression for the response of the first voxel as an example. Since the parameters of Ridge regression do not necessarily shrink to 0, to AIC, AICc or BIC criteria are meaningless. Figure 5 shows the ESCV scores and means squares for different hyper-parameter λ.As λ goes up, ESCV goes down with several irregular results(for example when λ is near 0.2). The mean square reaches the minimum at 0.063, which is close to λ minimizing the mse in Lasso.

Figure 6 shows the correlations between predicted and real responses for each one of the 20 voxels. Generally, multiple LASSO performs slightly better than individual LASSO. We can also read from the figure that the range of correlations is from near 0 to the highest 0.6. The correlations for voxels 10,13,16,20 are the worst. Three different LASSO methods all show poor results for those voxels. This is probably because of locations of those voxels. 3D plot of the locations tells that voxel 10,13,16,20 all locate at the border, which may be the indication to try some non-linear models for those voxels.

Figure 7 plots the lambda chosen for Lasso and Ridge shows that the regression results for different voxels vary from image to image. For each voxel, the Lasso/Ridge regression was run ten times with the 80% training data random shuffled. The lambda.1se is the most stable index with the least average standard deviation(0.0545) among all the voxels. In Ridge regression, the lambda.1se for voxels 10, 11, 13, 16, 20 are significantly lower than the rest, indicating again the regularity of those voxels.

In fact, the features select from LASSO if rather stable. The left panel in figure 8 plots the frequencies of times of being selected for all the features. I should remind that most of the features were never selected in the 10 training processes, some of the features were selected one or two times. However, it is worthing mentioning that there are a dozen of features being selected every time regardless of the random shuffling.

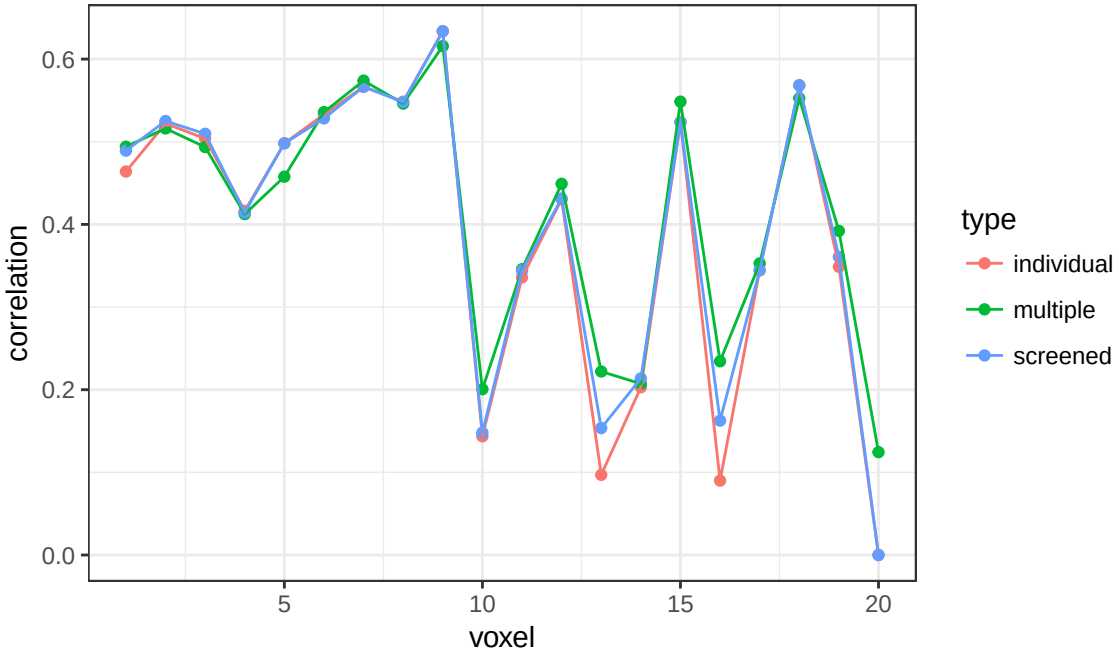

Figure 6: Correlations for test set for all the voxels from individual Lasso regression without feature screening, individual Lasso with feature screening and multiple Lasso regression with feature screening. The lambda of each is chosen as lambda.1se.

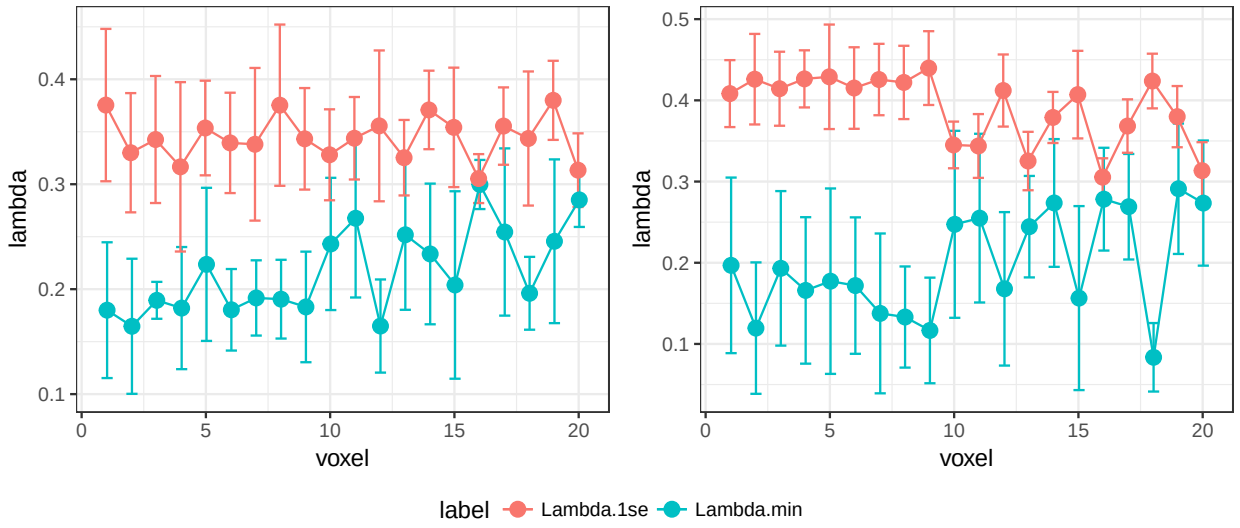

Figure 7: Histogram of the correlation similarity measure for different values of k

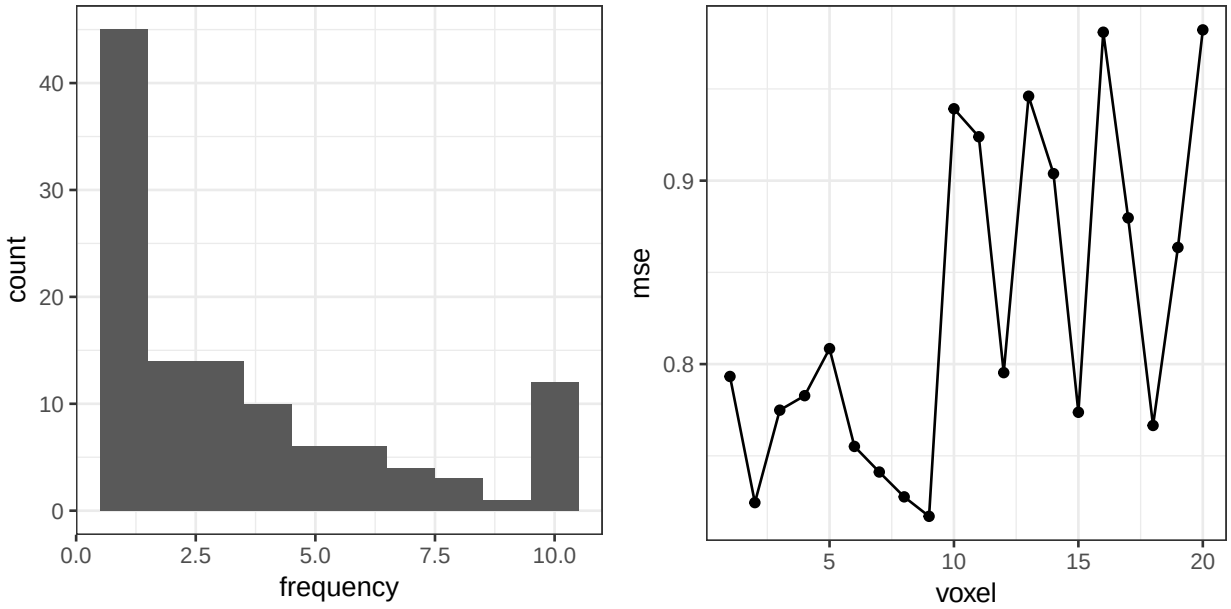

Figure 8: Tesing the stability of multiple Lasso regression by random shuffling 10 times the training data set and take 80 percent of the sample for training. The left panel shows the freqencies of the selected features in the training processes. The frequency zero is omitted. The right panel shows the mean of MSE for each voxel.

Those features(1111,1238,1397, 1408,...,5448) are considered to be the most significant and most stable ones. The right panel shows the mean of mean square error for each voxel, which is in line with the correlation plot(figure 6): the voxels with higher MSE have lower correlations.

3.6 Interpretation

This part is for the interpretation for the best LASSO method–the multiple regression. In details, after trying λ for cross-validation and choosing the one with least mean square loss, I get the model with 120 features(including one constant term).

My regression model indicates that voxels respond to contrast in certain directions and in particular areas of an image. Most important features seem to consider only upper half of image, with few on the border.

Figure 9 shows the collection of the real parts of the Gabor wavelets which are selected as the features from LASSO. I got several inspirations from the figure. First, the Gabor waves gather at the top left and top right parts of the image, which may be due to the habit of observation: we start looking the top left part and then come to top right part. Interestingly, the exact center of the image does not draw much attention; however, the area around the center is usually the most significant part of the image. Second, the scale of wavelets are really small, meaning that we see objects from local characteristics and those local characteristics may arouse the strong reaction from our brains.

Figure 10 gives a more concrete example. the circle shape of the image makes us focus on the center. And upper left part draws most of our attention, then right part and the center.

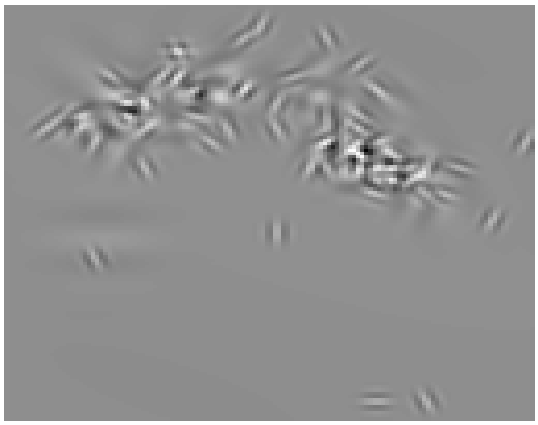

Figure 9: Plot of the real parts of Gabor wavelets weighted by the coefficients selected by multiple Lasso regression. Gabor wavelets have different scales, angles and locations. The feature waves selected by Lasso tend to be have small scales, all kinds of angles and concentrated locations.

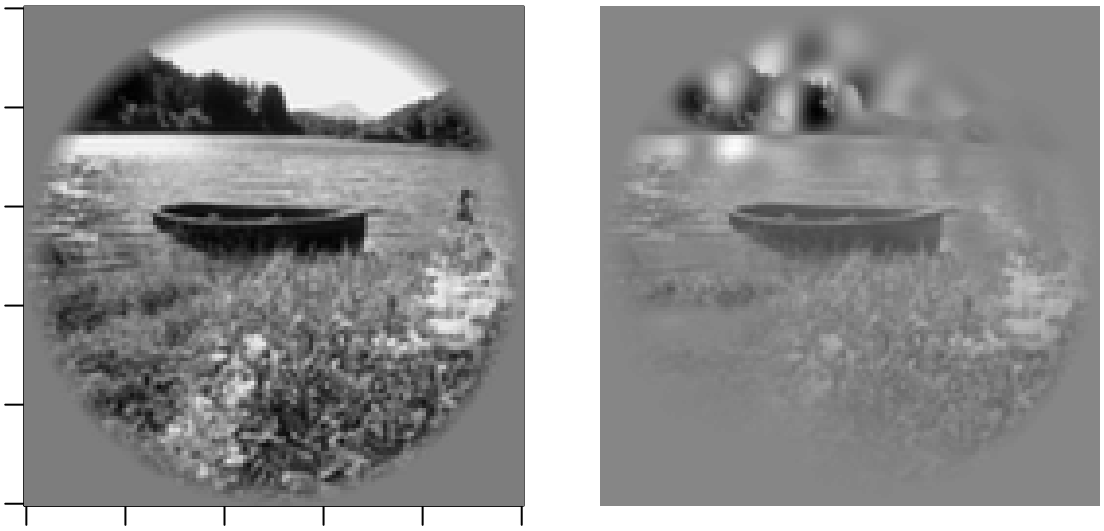

Figure 10: Plot of Image-10(left panel) and the image convoluted by the real part all the selected wavelets(right panel).

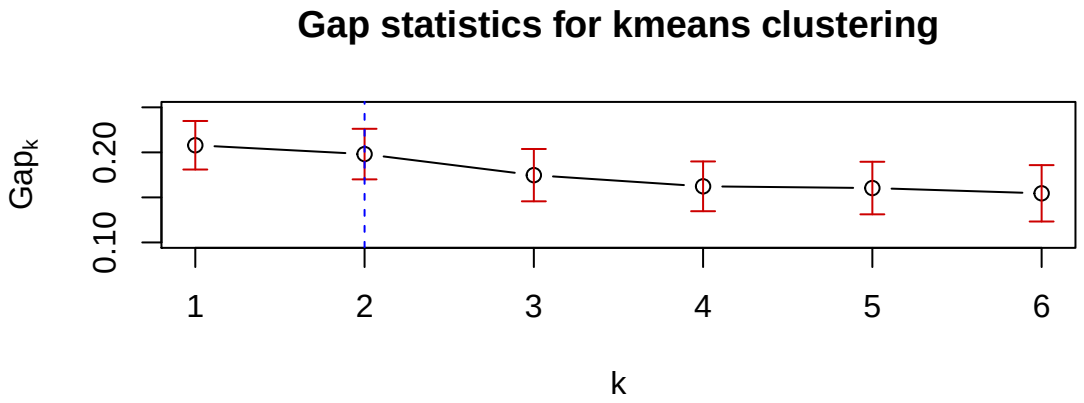

Figure 11: Plot of gap statistics for clustering.

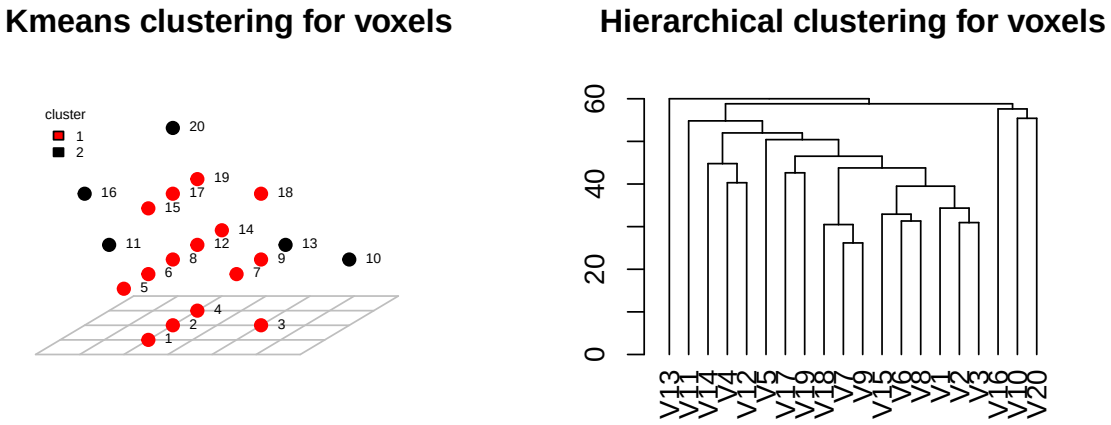

Figure 12: Cluster results from kmeans and hierarchical clustering. Kmeans shows a more intuitive grouping: inner voxels and outer voxels

4 Prediction

My previous study indicates that the voxels on the border(10,13,16,20) are different from the voxels in the center according to the correlations in the testing part. By separating the voxels into several groups and apply different models to the group could probably get a better result in prediction. I will apply the kmeans algorithm for clustering first and the gap statistics to decide the number of clusters.

Even though figure 11 tells that the gap statistics reaches maximum in the case of only one cluster, the gap statistics of two clusters is just slightly lower. Inspired by the results of correlations from Lasso regression and the lambda.1se chosen in Ridge regression, I decided to divided the voxels into two groups.

Hierarchical Clustering and Kmeans shows similar results. Hierarchical clustering distinguish voxel 13 from all other voxels, then the group of voxel 10,16 and 20.

Table 1 shows the "mixed model" I used for prediction. For the inner voxels, Lasso and Ridge are good enought for prediction. However, Random forests are an ensemble learning method that combine the prediction of multiple component decision trees in order to make a probabilistic determination of the class of

Multiple Lasso Regression	Individual Random Forest
V1,V2,V3,...,V9,V12,V14,V15,V17,V18,V19	V10,V11,V13,V16,V20

Table 1: Two prediction models

Voxel	Correlation of Random Forest	Correlation of Lasso Regression
10	0.19	0.19
11	0.38	0.34
13	0.27	0.21
16	0.21	0.24
20	0.15	0.11

Table 2: Comparison between two prediction models.

an input. When training random forests, I consider:

- **Number of trees**: Usually, more trees lead to a more stable prediction. Since each tree uses different baseline data and different features, overfitting is usually not a big concern. However, the time espense for high dimensional data is so high.
- **A moderate number of maxnodes**: A deep tree may result from overfitting; and few maxnodes suffer from underfitting.

  Tuning those two variable took me a long time. Finally I selected the number of trees to be 1000 and maxnodes of one tree to be 100. The training time last ten hours on the SCF clusters. I did not try parallel computing and it seemed that it was the biggest mistake I made. Table 2 shows the testing correlations of those two methods. Random forest regression is slightly better.

5 Conclusion

In this report, I have explored and modeled various regression methods to perform analysis voxel responses from fMRI, based on Gabor wavelet features of images. My data are all of high dimensions, each picture has 16384 pixels and 10921 features for wavelets. Regression models were trained on 80% of the data.

Features were extracted mainly from Lasso regression. Ridge regression was also conducted, but only for comparison. Validity of several Lasso models was assessed via correlations and cross -validation technique with some model selection criterion such as AIC and BIC. ESCV turns out be another strong model selection criteria, which has similar results as AIC and AICc.

For prediction clustering methods were performed to separate the voxels into inner ones and out ones. I just used the Lasso regression model for the prediction of the inner ones, and applied random forest to predict the responses of outer voxels.

Careful analysis leads me to identify that geographical locations of voxels do have a strong influence on the response. Feature waves that I extracted mainly focus on the upper center of pictures, are likely due to the habit when we see pictures.

6 Reference

[1] Identifying natural images from human brain activity, Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, Jack L. Gallant, Nature, 06713.

[2] Estimation Stability with Cross Validation, Chinghway Lim, Bin Yu,arXiv:1303.3128v1, 2013

[3] Sure independence screening for ultrahigh dimensional feature space, Jianqing Fan, Jinchi Lv, 10.1111/j.1467-9868.2008.00674.x, 2008

[4] Regularization Paths for Generalized Linear Models via Coordinate Descent,Jerome Friedman, Trevor Hastie, Rob Tibshirani, Journal of Statistical Software, Volume 33, Issue 1, 2011.