Final Project - Natural Image and Voxel Responses

Linqing(Waverly) Wei

12/07/2018

1 Introduction

Human brain is comprised of complex strucutres and delicate subunits. People's curiosity towards this charming object builds up the exciting research field, neuroscience. Neuroscience is aiming to decode brain activity, and the invention of fMRI in the 1990s brings neuroscience to an even more exciting stage. fMRI discretizes 3D volumes of the brain into cube-like voxels. Therefore, researchers are able to analyze brain's response to stimuli at a granular level. The previous researches mostly used simple stimuli, but this study is focusing on brain's response to complex visual stimuli: images.

2 Data

fMRI data is collected by showing each subject 128 x 128 grey scale images, and each image can be represented as a vector of length 128^2. Through transformation, the length is reduced to 10921, representing the features of each image. resp_dat (1750 x 20) contains responses from 20 voxels to 1750 images. fit_dat(1750 x 10921) contains the features of 1750 transformed images. val_feat contains a validation set of 120 transformaed images. loc_dat contains 3D spatial location of the voxels.

2.1 Exploratory Data Analysis

The given data is very clean, no need for extra data cleaning. Both feature matrix and response vectors contain continuous variables with some columns containing constant numerical values.

3 Model Selection

In order to investigate how voxels respond to images, we first experiment with various prediction models and select the best performed ones to proceed to the future analysis. Before model construction, the original data is split into training, validation and test sets with 60:20:20 ratio. LASSO and Ridge are used as the two regression models for comparison. Penalty parameter for each model is selected with different criteria: CV, ESCV, AIC, AICc and BIC. The performance indicator is the same as the one used in the Gallant lab. Performance of each model is evaluted by finding the correlation between the fitted values and observed values on a validation set for each voxel. The higher the correlation measure, the better the performance of the model.

Each selection criteria has its pros and cons. AIC tends to be a more conservative selection criteria since it prefers a model with more parameters. In contrast, BIC penalizes additional parameters more seriously by introducing the penality term $m * ln(n)$.m represents the number of parameters and n represents the number of observations. BIC is better applied on nested models compared to AIC. BIC is well known for selecting the true model, meaning that if the true model is within the candidates, BIC will select the true model with probability 1 as n→ ∞. However, when n is finite, AIC tends to select a better model than BIC since AIC is good for finidng the best approximating model. AICc is known as the small sample corrected AIC. Therefore, when n is large, the performance of AIC and AICc will converge. When n is finite, AICc tends to perform better than AIC.

CV is intuitively more straightforward and is actually asymptotically equivalent to AIC criteira. It is sensitive to both the functional form and the number of parameters. However, since CV averages over multiple partitions, the partition proportion also plays a role in CV performance. Partition may potentially cause "loss of data", leading to instability in CV. When CV is applied onto a sparse model, like LASSO,it tends to become unstable in high dimensions. ESCV would be a better substitute to handle this situation. ESCV is built on CV, but it cuts down the false positive rate and finds a smaller model locally with high stability. However, cutting down the false positive rate at the same time will sacrifice true positive rate.

To investigate how each model performs in practice, in Figure 1, we plot the correlation of LASSO and Ridge models in combination with the five model selection criteria. Apprently, CV and ESCV combined with both LASSO and Ridge outperform the models with selection criteria AIC,AICc and BIC.The reason might be that CV and ESCV are non-parametric parameter selection methods, which would be more flexible to react to large and noisy feature sets. BIC performs the worst in this case because when the "true" model is not within the candidates, its non-conservative selection algorithm would put constraint on its performance.

However, the performance of ESCV and CV is hard to distinguish from the plots. Therefore, we zoom in ESCV-LASSO, CV-LASSO, ESCV-Ridge, and CV-Ridge to select the best performed model.

In Figure 2,ESCV-LASSO, which is coded in red, has an overall better performance than the other three models. However, the performance is not drastically better than CV-LASSO. ESCV-Ridge has the third best performance and followed by CV-Ridge. By comparing the four models, ESCV seems to be a better selection criteria than CV, and LASSO has an overall better performance than Ridge. Therefore, ESCV-LASSO is selected as the best performed model for future analysis. This result verifies the previous statement that ESCV handles sparse model better than CV.

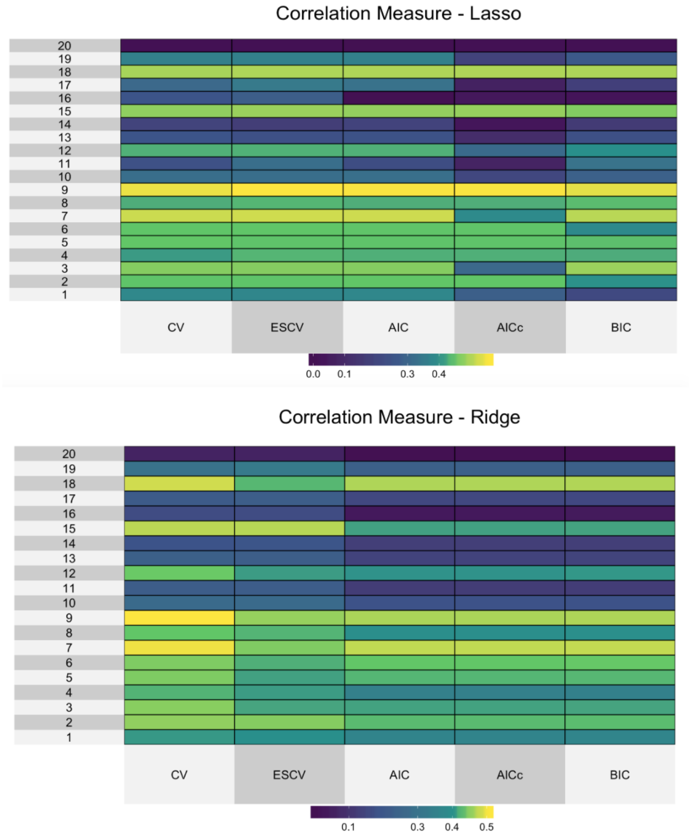

Figure 1: Correlation measures of five models for lasso and five models for ridge. Each model corresponds to a different selection criteria (CV,ESCV,AIC, AICC, BIC).

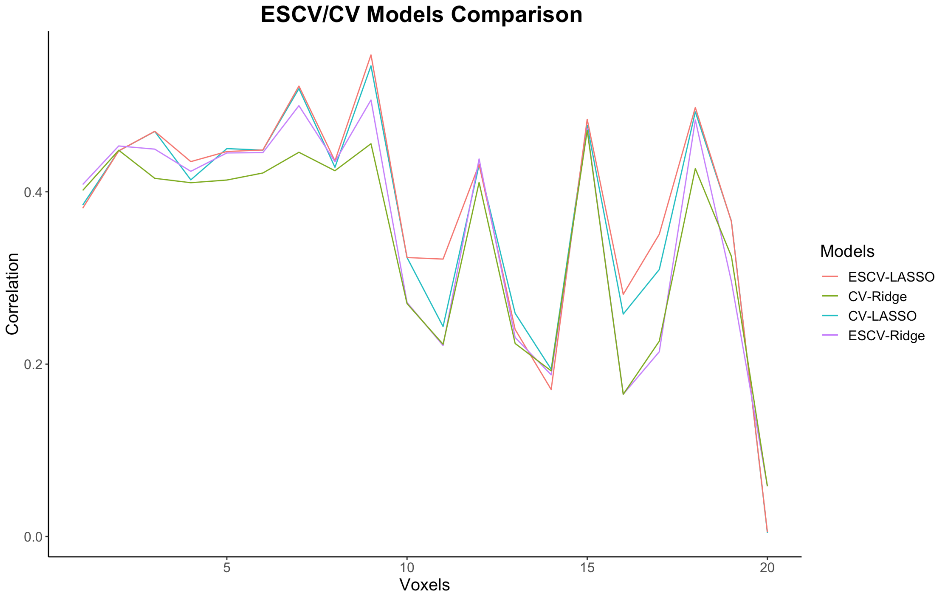

Figure 2: Performance comparison of ESCV-LASSO,CV-LASSO,ESCV-Ridge,and CV-Ridge.

4 Diagnostics:

ESCV-LASSO and CV-LASSO have been shown to have the best performance during the model selection stage. However, the stability and prediction accuray need to be further investigated. First, we run the two models on voxel 1 and voxel 2 to compare the correlation measures with the previous results. This will give us a sense of prediction stability. Also, we would like to verify whether ESCV always outperform CV to assess the model stability. To do this, we run the selected models on the test set.

In Table 1, we summarized the correlation measures of ESCV-LASSO and CV-LASSO for voxel 1 and 2 on validation set, together with the results run on test set. Both models have even higher prediction accuracy running on the test set. Therefore, the fit of the models seem to be constantly good on new data set. However, for voxel 2, CV has a slightly higher correlation measure than ESCV when running on the test set. This indicates that the relative performance of CV-LASSO and ESCV-LASSO may vary in response to new data set. Overall, ESCV-LASSO achieves higher correlation but there could be cases that CV-LASSO actually outperforms ESCV-LASSO.

4.1 Outliers

Running on 20 voxels, we do detect some outliers. For example, voxel 16 and voxel 20 have extremely low correlation across all models. Voxel 20 even have correlation measures close to 0 even running with ESCV-LASSO. The reason might be that voxel 20 does not have a fixed or regular response pattern to visual stimuli. The collected features are not good predictors to capture the response of voxel 20 to images. Therefore, with the restriction of collected data, even a stable and well-performed model, like ESCV-LASSO, fails to make an accurate prediction.

4.2 Further Improvement of the Model

After running with ESCV-LASSO and CV-LASSO, we get both prediction results and also selected features by LASSO. One thought is, if we use only the selected features from ESCV-LASSO to run other models, can

Table 1: Model Diagnostics: Voxel 1 and 2

	CV	test.CV	ESCV	test.ESCV
Voxel1	0.3809017	0.4371393	0.3846245	0.4371832
Voxel2	0.4474537	0.4647489	0.4475132	0.4638718

we achieve better performance? To test the idea, we use the selected features from ESCV-LASSO and run random forest only with selected features.

In Figure 2, we compare the performance of using ESCV-LASSO and using selected features from ESCV-LASSO to run random forest on voxel 1 and voxel 2. It turns out that the performance is consistently better on both voxels when running the model with random forest with only the selected features from ESCV-LASSO. The reason could be that we use ESCV-LASSO to largely reduce the noise within the dataset. Random forest itself has great prediction performance and avoids overfitting. Combining these two models thus improves the correlation measure.

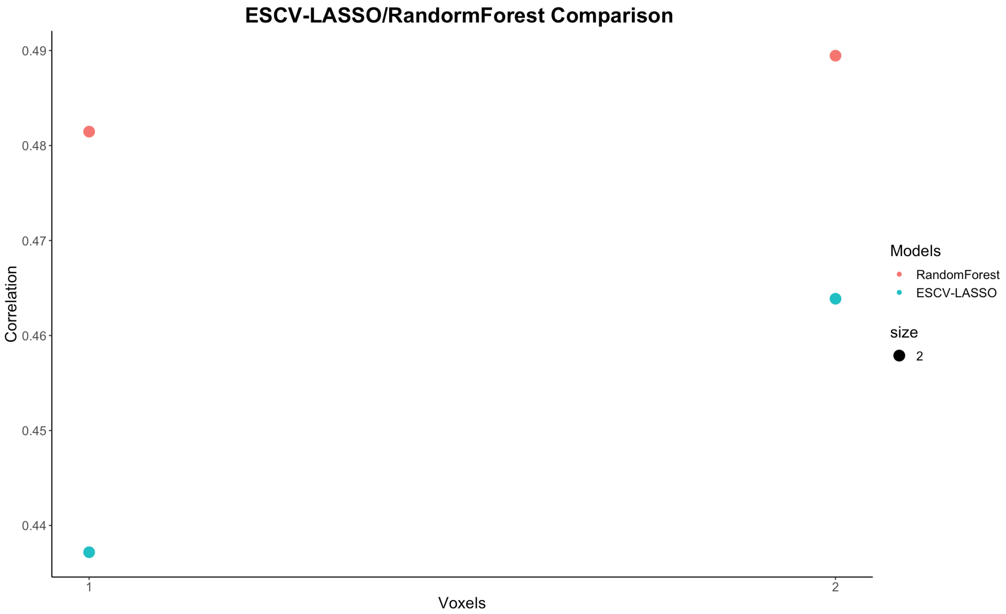

Figure 3: Compare correlation with only ESCV-LASSO vs. further processing with RandomForest

5 Feature Analysis

After thoroughly comparing and choosing the best model for predicting the voxel response to images, we now shift our focus to interpret the results. For the following sections, we will map our findings back to the images to see what the numerical values mean in reality. ESCV-LASSO and CV-LASSO are the models we choose to generate the following results.

5.1 Features Shared Across Voxels

First, we are curious to know if there are any features shared by 20 voxels.In fact, 32 features are shared across all 20 voxels using ESCV-LASSO model. For example, feature 915,994,1012,1152,1260,1405,4099,4119, etc. There is an interesting pattern shared by some features such as feature 915 and feature 994. They always locate at the dark empty part of the image but never onto the identifiable part(e.g.human face,eyes). In

Figure 4, we map feature 915 onto two randomly selected images: 1 and 903, one with human and another one with animal. The assumption is that all voxels share feature 915, 994 beacause those features are used to identify the background information, which is a necessary functionality for each voxel, and they may very likely respond to the background information in the same manner.

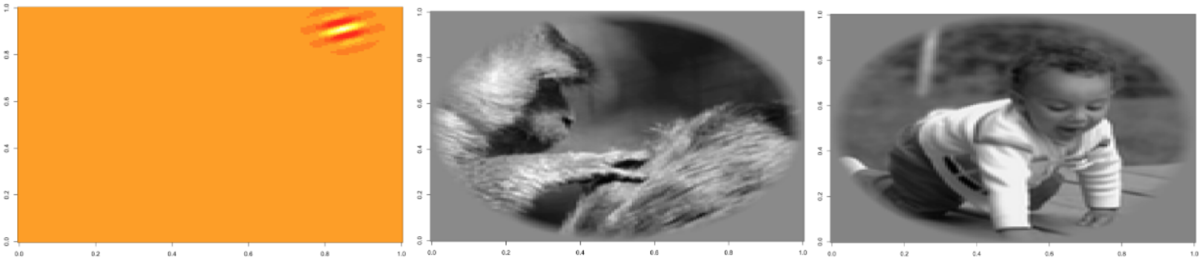

Figure 4: Map feature 915,which is shared by all voxels, onto image 1 and image 903. It identifies the background of each image.

5.2 Features Shared Across Models

Now, we would like to see if there are any features shared across models. We understand that instability exists within our models, so if the features could combat model instability, then they must encode important information. We zoom in voxel 1 to explore interesting patterns. Using ESCV-LASSO and CV-LASSO to select predictors for voxel1, only feature 1145 is shared across models. In Fig 5, we map this feature onto images. Most of the time, this feature locates at the head part. More specifically, it identifies hair. The identification is not only restricted to human head. When mappin onto image with animals, such as image 903, it also identifies the monkey's hair. Feature 1145 is a realtively stable singal across models. Therefore, it does unfold part of the mystery of voxel 1. One assumption is: voxel 1 is sensitive to the visual stimuli "hair."

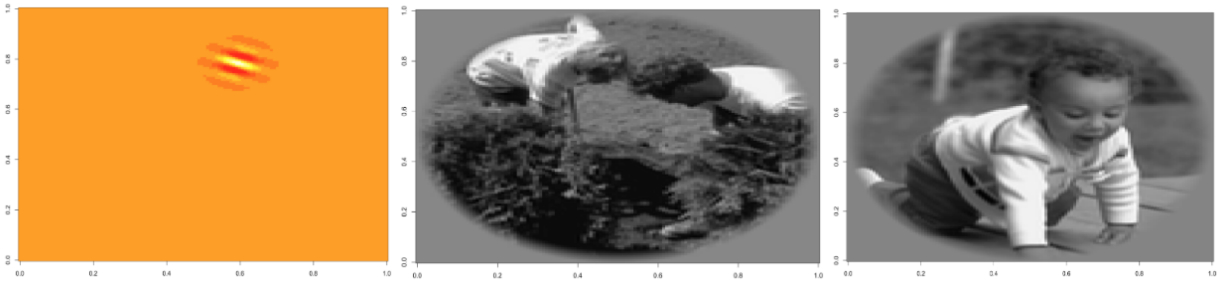

Figure 5: Map feature 1145,which is a predictor of voxel 1,shared across models,onto image 3 and image 3. It identifies human head, especially the hair part.

5.3 Features Shared Across Bootstrap Samples

To further investigate model's internal stability of selecting features and further investigate how voxel 1 responds to images, we apply bootstrap for 20 iterations and use ESCV-LASSO as the prediction model on voxel 1. There are no features shared by all 20 bootstrap samples for the following reasons: 1) the model is not stable enough in terms of feature selection 2) the total amount of features in the original data is huge, leading to unavoidable noise 3) the total number of bootstrap samples is small due to computation time limitation.

Table 2: Features shared across bootstrap samples with high frequency

feature	count
1145	9
5163	5
5460	5
5733	6
5875	5

Therefore, instead of searching for the stable features across all bootstrap samples, we decide to count the occurence of each selected feature to detect the relative stability of features across bootstrap samples.

In Table 2, we summarized the features with occurence larger than 5 times out of 20 bootstrap samples. Feature 1145 is selected 9 out of 20 times, which once again verifies the significance of feature 1145 as a predictor foe voxel 1. Other features that have high occurence are 5163,5460,5733,5875.

In Fig 6, we map these five features onto one single image to see as a group of features, how they could possibly unfold a hidden story. These five features all surround the human head in this image. They work together to identify the contour of human head. It is reasonable to conclude that the features remaining stable across bootstrap samples are the ones important for voxel 1 to capture the core imformation from an image, for example, human head. However, this part of analysis is quite limited. Extending to large number of bootstrap samples and to other 19 voxels would definitely reveal more exciting patterns.

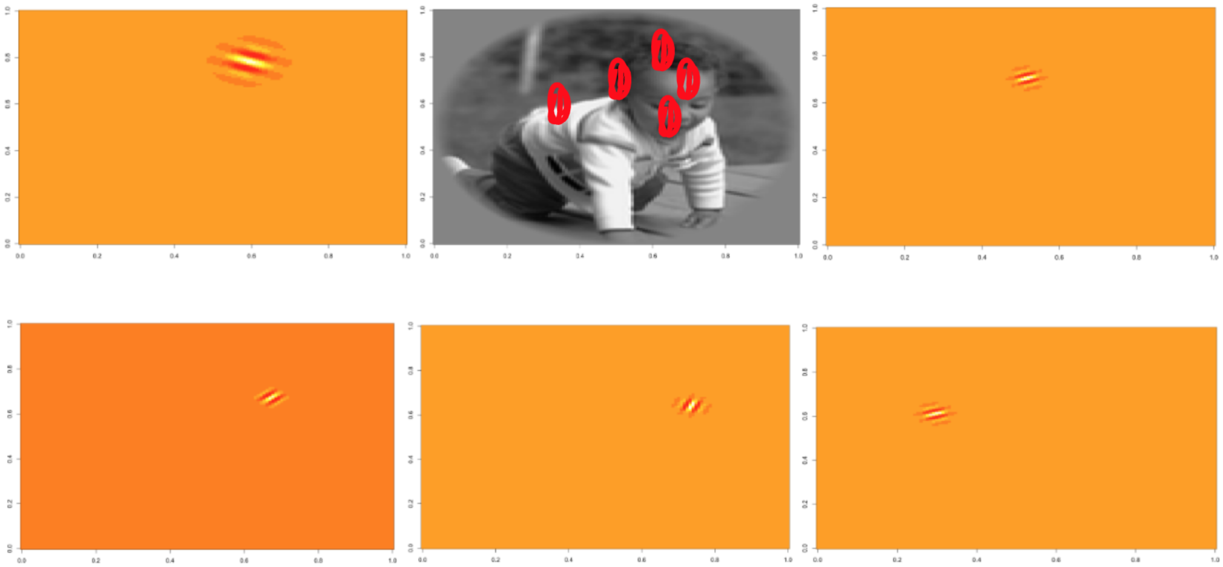

Figure 6: Map relatively features after running bootstrap onto image 1. From left to right, top to bottom: feature 1145,5163,5460,5733,5875.These features circle around human head

5.4 Hypothesis Testing

In the previous analysis, we were able to extract features and the corresponding beta values from the model we chose. A natural follow-up question would be: are those features/betas statistically significant? To answer this question, we could conduct hypothesis testing.

Hypothesis testing can be conducted at the following steps:

1. Pick predictors that are stable across bootstrap samples and extract their corresponding β values. Calculate the statistics of each individual beta such as mean and standard error.
2. The null hypothesis is that the observed beta is not statistically significant. We calculate t-statistics to evaluate the significance of each individual beta using the following formula. The subscript B represents bootstrap samples which betas are extrtacted from.

$$t = \frac{\hat{\beta}_B - \beta_B}{SD(\hat{\beta}_B)}$$

3. If $|t| < 1.95$, then we say the true beta belongs to a 95% confidence interval. Otherwise, we reject the null hypothesis.
4. Now we are only looking at 20 bootstrap samples and only one beta is relatively stable across all the samples. However, if we extend to large number of bootstrap samples, n = 100000, there might be multiple betas that remain stable and need to conduct hypothesis testing on. In that case, we would face a multiple hypothesis testing problem and we need to adjust the significance threshhold. There are multiple ways to approach this. For example, we could do Bonferroni Correction: α/n to account for multiple betas we would like to test on. Alternatively, we could use False Discovery Rate, which is the proportion of false positives among all significant results.

6 Conclusion

Decoding fMRI data to unfold humna brainactivity related to image identification is an exciting direction to dive into. The decoding process requires the construction of well-performed prediction models. ESCV-LASSO is a great combination of handling feature selection, prediction, and model stability. A well structured statistical model provides trsutworthy and meaningful clues to help us decode voxel response patterns to images. It is interesting to find out that features that are shared across models and withstand the stability test are associated with important identification information of each image. However, these findings only reveal a small portion of the voxel mystery.

The future direction would be to extend the stability test and also apply the analysis onto all voxels. At the same time, we would like to explore prediction models with even higher accuracy to capture the intricate pattern we might have missed in this study.

7 Bibliography

[1] Kay, K., Naselaris, T., Prenger, R. and Gallant, J. (2008). “Identifying natural images from human brain activity.” Nature 452(20):352-355.

[2] Lim, C. and Yu, B. (2013). “Estimation Stability with Cross Validation.” Jounral of Computational and Graphical Statistics 25(2):1-31.