Final Project - Predicting Brain Responses to Visual Images, Stat 215A, Fall 2017

Amy Ko (24978168)

December 10, 2017

1 Introduction

Reconstructing the mental images from brain activities is an important problem in neuroscience. In this report, we aim to model and predict the brain's responses to visual images. We use the fMRI data provided by the Gallant Lab [1], which measured the responses to visual images at various cubic units, or voxels, in the brain. Through Gabor transformation of the images, we model the connection between the images and the brain responses as a linear relationship. The first part of the report will explore several linear regression models to best predict the brain response to images. The second part aims to interpret the models in order to better understand how the brain responds to images.

2 Data

The data contains a single subject's brain responses to 1750 images in 20 voxels located in the region responsible for visual functions. Although the fMRI response is a function of time, the response was reduced to a single number, resulting in a vector of size 1750 for each voxel. Each image, originally a 128 pixel by 128 pixel grey scale image, was Gabor-transformed to produce a vector of size 10921. The training data, therefore, contains a response matrix of size 1750 x 20 (one column per voxel) and the feature matrix of size 1750 x 10921. In addition to the features and the responses, we are also given the 3-dimensional spatial location of each voxel.

For model selection, I split the data into three sets: training, validation, and test sets [2]. I randomly chose 50 percent of the data as the training data, 25 percent as the validation data, and the remaining 25 percent as the test data. As I will discuss in detail in later sections, the training data was used to select λ in the regularized regression models; the validation set to select the best model; and the test set to measure the performance of the selected model.

3 Model Selection

3.1 Exploratory Data Analysis

Before we dive into model selection, we will first explore what the data looks like. Figure 1 shows the distribution of the response at each voxel. As we can see, the distribution of each voxel is approximately normal, which suggests that the normality assumption in the various models we will consider may be satisfied.

Next we explore how voxels are correlated with each other. We see in Figure 2 that there are clusters of voxels that show fairly high correlation, such as voxels 6 through 9. Interestingly, all voxels show positive correlation with each other. However, as shown in Figure 3, the voxels that are physically close to each other do not necessarily show high correlation. For example, pairs of voxels with distance of 1 have a wide range of correlation values, from less than .1 to over .8. But if the correlation is high, then the distance tends to be smaller.

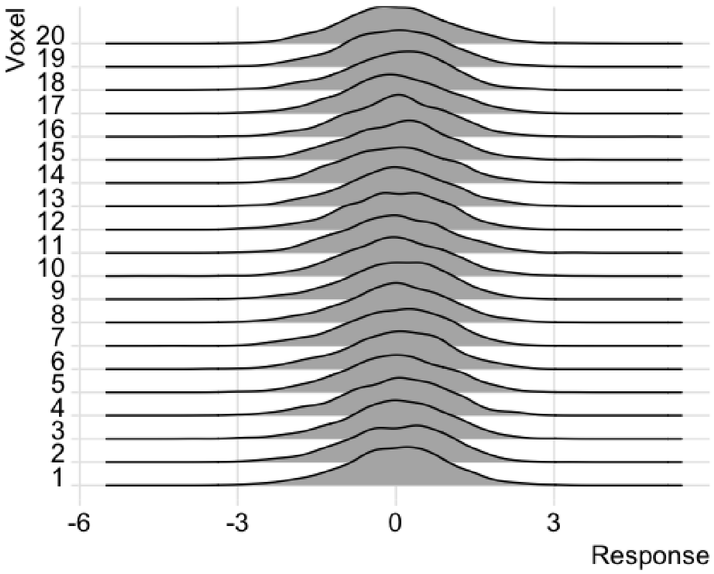


Figure 1: Distribution of the response at each voxel.

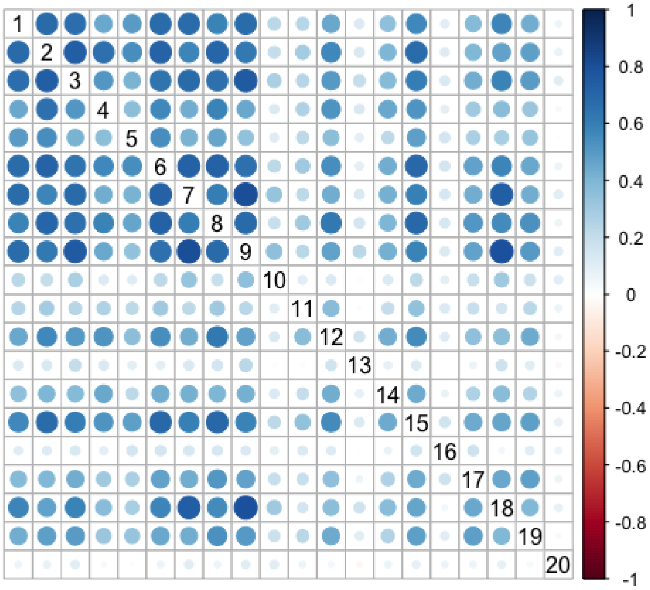


Figure 2: Correlation between the voxel responses. The diagonal indicates the voxel number and the color indicates the direction and the strength of correlation.

3.2 Model Selection Criteria

To select the best model for our data, I explored two regression models: LASSO and ridge. Here I fitted each voxel separately. To select the tuning parameter λ, I considered five selection criteria: 10-fold CV, ESCV, AIC, AICc, and BIC. I will briefly describe ESCV, AIC, AICc, and BIC below.

ESCV is a model selection criterion based on an estimation stability metric and CV. The estimation stability measure is meant to help the CV selection, which may be unstable in high dimensions. To compute the estimation stability, or ES, we first divide the training data into M blocks. For a given λ, we fit the data for each block i and estimate its coefficients, β_i. Then the ES is defined as

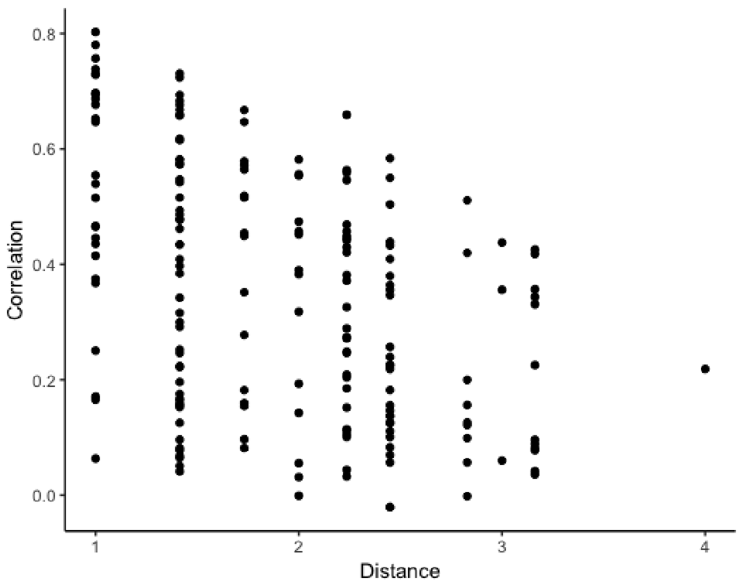


Figure 3: Euclidean distance between a pair of voxels vs. correlation between their responses.

$$ES = \frac{\frac{1}{M}\sum_m \|X\hat{\beta_m} - X\hat{\hat{\beta}}\|^2}{\|X\hat{\hat{\beta}}\|^2},$$

where X is the design matrix and $\hat{\hat{\beta}} = \frac{1}{M}\sum_m \beta_m$. We compute this stability measure for each choice of λ and choose the one for which ES is the smallest and is no smaller than the CV choice. For our problem I chose to divide the training data into 10 blocks ($M = 10$).

AIC measures the goodness of a statistical model by a modified maximum likelihood. For regression models with Gaussian errors, we can write AIC as

$$AIC = n \log RSS + 2k,$$

where RSS is the residual sum of squares, n is the number of observations, and k is the dimension of the model. For our problem, we define the dimension of the model as the number of predictors for which the coefficient is non-zero. We then seek to minimize AIC.

AICc is similar to AIC but corrects for the finite sample size and can be written as

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}.$$

Note that the correction factor is negative when $n < k + 1$, which would make AICc smaller. This is indeed possible for our problem since the number of features (p = 10921) is much greater than the number of observations (n = 880) in the training data. So AICc may not be an appropriate metric in our case since this correction term may overpower the penalty of adding more features, favoring more complex models.

BIC comes from a Bayesian point of view is approximately equivalent to choosing the model with the highest posterior probability if the prior assigns equal probability to every model. BIC imposes a stronger penalty on the dimension of the model than AIC for large n and can be written as

$$BIC = n \log RSS + k \log n.$$

These model selection criteria have their own strengths and weaknesses. Unlike AIC, AICc, and BIC, both CV and ESCV are model-free and can be applied to a wide range of models. On the other hand, AIC, AICc, and BIC assume a model and may require good estimates of the model parameters to work well.

Furthermore, CV and ESCV use the left-out data to directly estimate the test error, whereas the other criteria use the entire training data to select the model, which may lead to overfitting. CV and ESCV, however, do not directly take into account the complexity, or dimension, of the model. Furthermore, CV and ESCV are generally more computationally intensive, which could be problematic for large data sets.

3.3 Selecting λ

Figure 4 shows the values of various model selection metrics against λ. Here, only the results for voxel 1 are shown for visualization (the other voxels show similar patterns). We see that AIC does not go to a minimum as we hoped; instead, the plot suggests that the minimum might be somewhere near zero. This may be that the penalty term in AIC is not strong enough to select a less complex model. This is further evidenced by Figure 5, which shows the AIC values computed on the validation data (i.e. data that was not used in fitting the model). Here, the AIC curve goes to a minimum as we expect. This suggests that AIC is somehow overfitting the training data. AICc also shows a strange behavior. More specifically, there is a discontinuity at a small value of λ, which corresponds to the point where the correction term discussed before flips sign. This confirms our previous expectation that AICc is not a suitable selection criterion for our problem. On the other hand, BIC, CV, and ESCV go to a minimum. We can see that BIC imposes a stronger penalty than others, resulting in a larger value of best λ around .2. CV and ESCV, on the other hand, select a smaller value of λ around .1.

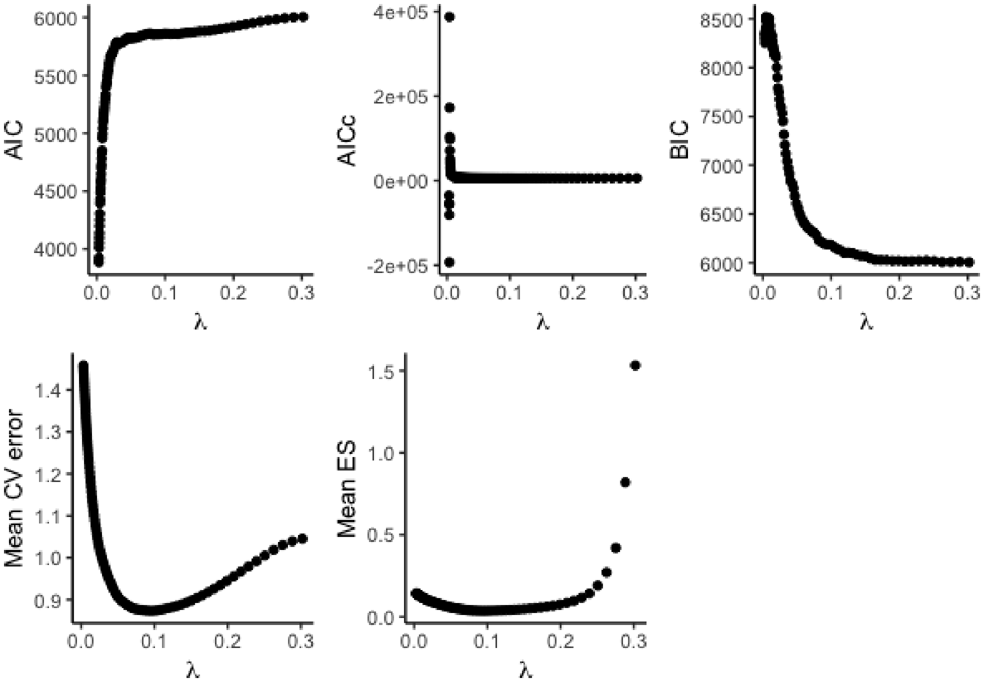


Figure 4: LASSO: model selection criteria at various values of λ. Only voxel 1 results are plotted for visualization.

For ridge regression, I used only two model selection criteria: CV and ESCV. Since ridge regression selected the same number of features at various values of λ, the other criteria would have been constant. We see in Figure 6 that ESCV selects a much greater value of λ than CV. Although CV seems to give a clear minimum value of λ around 70, the ES value at that point is quite high.

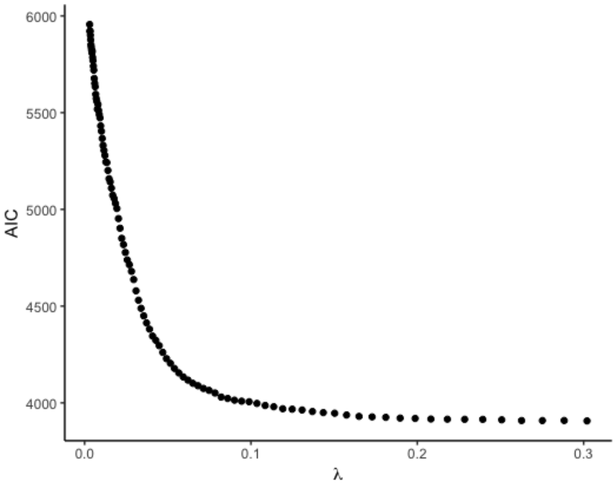


Figure 5: AIC curve using the validation data. Here, LASSO was fitted on the training data but AIC was computed using the validation data.

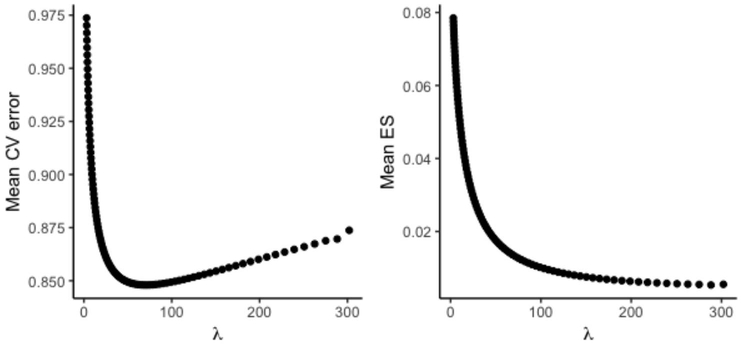
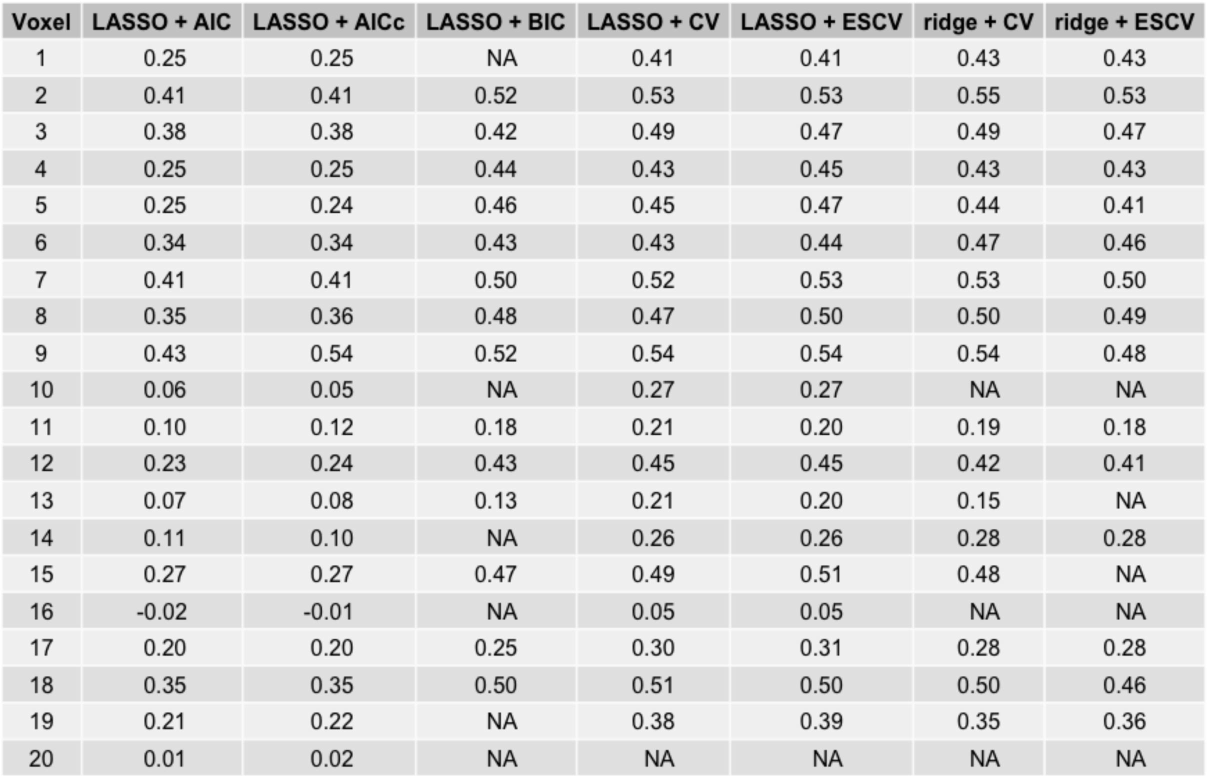


Figure 6: Ridge: model selection criteria at various values of λ. Only voxel 1 results are plotted for visualization.

3.4 Correlation between Fitted and Observed Values

To select the best model, I used the correlation between fitted and observed values on the validation set as the performance measure. More specifically, the response of each voxel was estimated for each of the competing models: LASSO + AIC, LASSO + AICc, LASSO + BIC, LASSO + CV, LASSO + ESCV, ridge + CV, and ridge + ESCV, where LASSO + AIC means LASSO with λ selected by AIC, and so on. Table 7 shows the correlation for each voxel in each of the models. The ”NA” entries correspond to models that predict a constant value for the response, making the correlation undefined. Voxels 10, 16, and 20 often resulted in this kind of model across many model selection methods. Interestingly, the response of each of these voxels had low correlation with those of other voxels (Figure 2).

I chose to select two models to further investigate, one from LASSO and one from ridge. The best model among each category was chosen based on the correlation value averaged over the 20 voxels. To make a fair comparison, I removed voxels for which the correlation was ”NA” in any of the models being compared. This led me to choose LASSO + ESCV and ridge + CV, which will be called LASSO and ridge for convenience in subsequent sections.

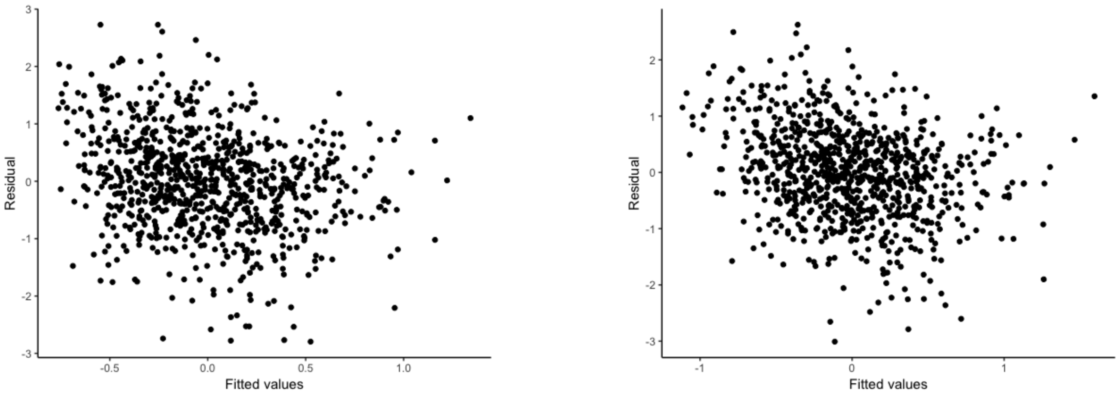


Voxel	LASSO + AIC	LASSO + AICc	LASSO + BIC	LASSO + CV	LASSO + ESCV	ridge + CV	ridge + ESCV
1	0.25	0.25	NA	0.41	0.41	0.43	0.43
2	0.41	0.41	0.52	0.53	0.53	0.55	0.53
3	0.38	0.38	0.42	0.49	0.47	0.49	0.47
4	0.25	0.25	0.44	0.43	0.45	0.43	0.43
5	0.25	0.24	0.46	0.45	0.47	0.44	0.41
6	0.34	0.34	0.43	0.43	0.44	0.47	0.46
7	0.41	0.41	0.50	0.52	0.53	0.53	0.50
8	0.35	0.36	0.48	0.47	0.50	0.50	0.49
9	0.43	0.54	0.52	0.54	0.54	0.54	0.48
10	0.06	0.05	NA	0.27	0.27	NA	NA
11	0.10	0.12	0.18	0.21	0.20	0.19	0.18
12	0.23	0.24	0.43	0.45	0.45	0.42	0.41
13	0.07	0.08	0.13	0.21	0.20	0.15	NA
14	0.11	0.10	NA	0.26	0.26	0.28	0.28
15	0.27	0.27	0.47	0.49	0.51	0.48	NA
16	-0.02	-0.01	NA	0.05	0.05	NA	NA
17	0.20	0.20	0.25	0.30	0.31	0.28	0.28
18	0.35	0.35	0.50	0.51	0.50	0.50	0.46
19	0.21	0.22	NA	0.38	0.39	0.35	0.36
20	0.01	0.02	NA	NA	NA	NA	NA

Figure 7: Correlation between fitted and observed values for each voxel. ”NA” means that the model predicted a constant value for the response, making the correlation undefined.

4 Diagnostics and Stability

Here we check the fit of the two models, LASSO and ridge. Figure 8 shows the residual plots for LASSO and ridge for voxel 1. As we can see, there are no obvious outliers or distinct patterns in the residuals, which seem to suggest that the fit was reasonable for both LASSO and ridge.

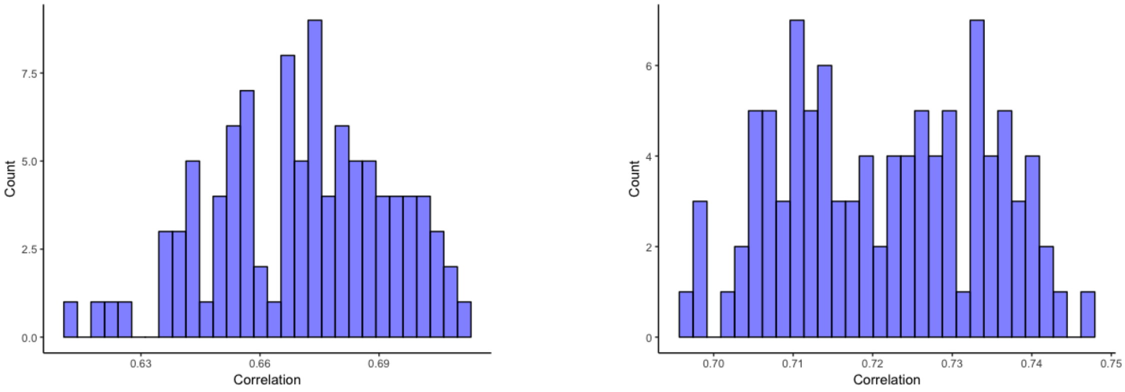


(a) LASSO

(b) Ridge

Figure 8: Residual plots for voxel 1.

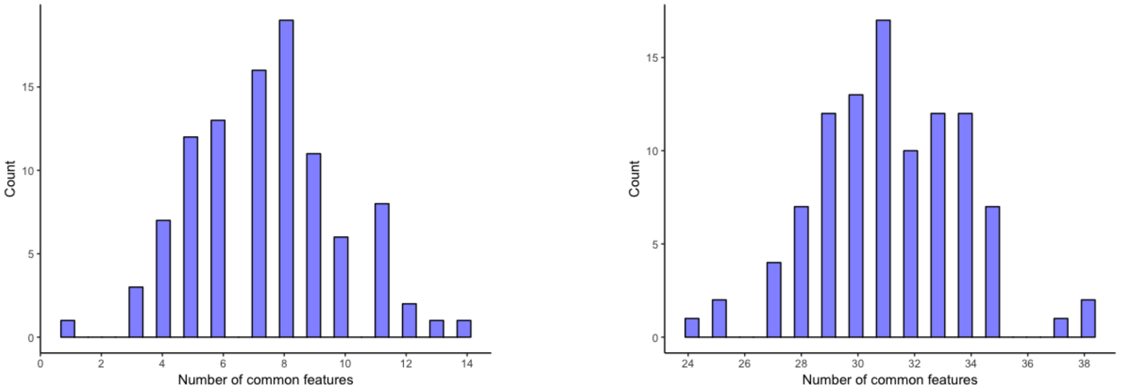
Figure 9 shows the distribution of the correlation between the fitted and observed values for voxel 1 over 100 bootstrap samples. We can see that for both LASSO and ridge, the correlation values fall within a narrow range. In other words, the predicted response is fairly stable over different bootstrap samples. To check whether the top features (i.e. features with the largest coefficient magnitudes) are stable across the bootstrap samples, I extracted top 50 features from each bootstrap model and counted the number of common features in each pair of bootstrap models. Figure 10 shows the distribution of this count. For LASSO, a pair of bootstrap models shares around 8 out of 50 top features. For ridge, on the other hand, they share about 30. This suggests that some of the features may be highly correlated. The effect is especially pronounced in LASSO, in which there may be many different sets of features that can be selected to produce good prediction accuracy. The correlated variables also affect the ridge estimates. Even though ridge selects the same number of features over different bootstrap samples, the coefficient estimates may vary significantly.



(a) LASSO

(b) Ridge

Figure 9: Distribution of correlation between fitted values and observed values for voxel 1 over 100 bootstrap samples.



(a) LASSO

(b) Ridge

Figure 10: Distribution of the number of common top features between each pair of bootstrap models. This is based on the top 50 features for voxel 1 in 100 bootstrap models.

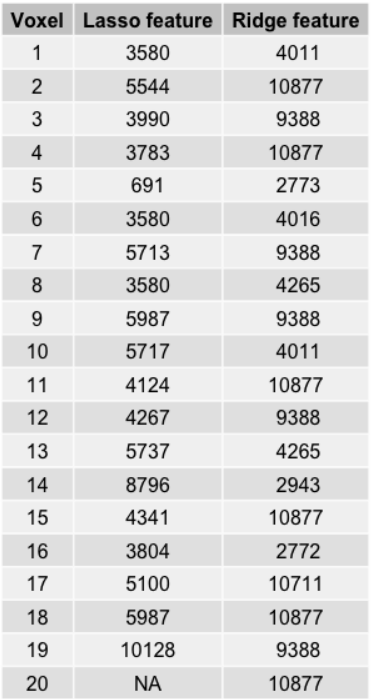
5 Model Interpretation

Here we turn to interpreting the model in order to better understand how the brain responds to visual images. To investigate whether the two models share any features with each other, I extracted top 50 features from each model and found the intersection between them. Surprisingly, the two models did not share any features. Again, this is probably due to the presence of highly correlated features.

Next, we turn to the most important feature for each voxel. Again, the importance of a feature was measured by the magnitude of the estimated coefficient (Figure 11). For LASSO, feature 3850 was the most important variable in three of the voxels (1, 6, 8). For ridge, feature 10877 appeared as the most important feature in six of the voxels (2, 4, 11, 15, 18). Interestingly, the voxels that share the most important feature do not necessarily show high correlation in their responses (e.g. voxels 2 and 11 in Figure 2). This suggests that there are features that are important across voxels that are not necessarily correlated.

To see which features are stable across the bootstrap samples, I again restricted our analysis to voxel 1. For each bootstrap sample, I fitted the data and extracted the top ten features. Then I tabulated how many times a particular feature appeared in the top ten list across 100 bootstrap samples. For voxel 1, feature 3580 was the most stable where it appeared in the top ten list 36 percent of the time; the next most stable feature was 5360 at 16 percent. For ridge, feature 2772 appeared 81 percent of the time and feature 2773 appeared 72 percent of the time. In general, ridge regression showed more stability in which feature is estimated as the most important than LASSO. This could mean that the brain response is better modeled by ridge regression, where many features have small effects on the response, unlike LASSO where only a small number of features have effects. But it is important to note that the presence of correlated features can challenge this interpretation.

Hypothesis testing on the estimated coefficients is tricky. First, both LASSO and ridge shrink the coefficients toward zero, making the estimates biased. Hypothesis testing in LASSO is particularly problematic in the presence of correlated features since a different set of features could be selected to give similar prediction accuracy. To get a sense of whether the coefficient estimate is non-zero in LASSO, one idea is to use the selected model in the OLS setting and carry out hypothesis testing in a normal way. For ridge, we could estimate the standard error of each coefficient via bootstrapping and test the significance. However, I would only use these methods to get a rough sense of the importance of features and would not over-interpret the resulting p-values.



Voxel	Lasso feature	Ridge feature
1	3580	4011
2	5544	10877
3	3990	9388
4	3783	10877
5	691	2773
6	3580	4016
7	5713	9388
8	3580	4265
9	5987	9388
10	5717	4011
11	4124	10877
12	4267	9388
13	5737	4265
14	8796	2943
15	4341	10877
16	3804	2772
17	5100	10711
18	5987	10877
19	10128	9388
20	NA	10877

Figure 11: Top feature for each voxel in LASSO and ridge. ”NA” for voxel 20 of LASSO indicates that no feature was selected for this model.

6 Performance of the Selected Model

To select the best model between LASSO and ridge, I computed the correlation between the fitted and observed values for each voxel and averaged over them. Voxels 10, 16, 20 were excluded since the correlation was undefined for these in at least one of the models. The average correlation value .42 for LASSO and .41 for ridge. Since LASSO performed slightly better, I chose to use LASSO (with ESCV) as my final model. Finally, the correlation on the test data is shown in Figure 12. The correlation varies across the voxels, with the values ranging between .08 (voxel 16) to .52 (voxel 8). Again, we note that the voxels with poor performance (e.g. 10, 13, 16) also showed low correlation in the response with the other voxels (Figure 2).

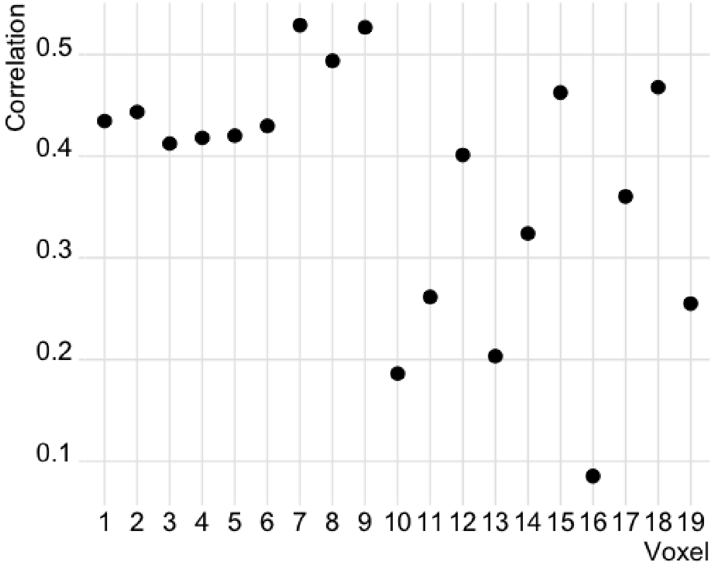


Figure 12: Correlation between fitted and observed values for each voxel on the test set by the selected model (LASSO + ESCV). Voxel 20 is omitted because the model predicted a constant value for the response.

7 Conclusion

In this report, we explored the relationship between visual images and brain responses using regularized linear regression. The model selection process showed that some selection criteria are more suitable than others. For example, AICc tended to favor more complex models due to the correction term, making it inappropriate for our data.

The prediction results suggested that there are many correlated features in the data, which is evidenced by the inconsistent set of important features selected by the models despite the relatively consistent estimates of the responses. We also saw that there are features that are important for several voxels, indicating that there may be a set of features that affect many or all voxels, whether the voxels are the correlated or not. Furthermore, we observed that the prediction accuracy is not uniformly good across all voxels. Voxels with poor accuracy were generally uncorrelated with the other voxels. A further study could involve modeling all 20 voxels simultaneously to take into account the correlation between the voxels. In conclusion, this study illustrates how we can predict the brain response to visual images and begin to understand the underlying relationship between them.

References

[1] Kay, Kendrick N. et al. Identifying Natural Images from Human Brain Activity. Nature 452.7185 (2008): 352355. PMC. Web. 10 Dec. 2017.

[2] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. New York: Springer series in statistics, 2001.

[3] Lim, Chinghway, and Bin Yu. ”Estimation stability with cross-validation (ESCV).” Journal of Computational and Graphical Statistics 25.2 (2016): 464-492.