Final Project - fMRI Stat 215A, Fall 2018

Daniel Soriano

11/19/2018

1 Introduction

In this report, we analyze fMRI data from the Gallant Neuroscience Lab at University of California, Berkeley measuring the brain's responses to visual images. In the experiment, subjects are shown images of everyday objects and the fMRI responses are recorded. The fMRI response is recorded at the voxel level, where each voxel is a cube-like unit of the brain that corresponds to a tiny volume of the visual cortex. We model responses in 20 voxels, located in the region of the brain responsible for visual functions, to images, selecting an optimal model for each voxel. We then evaluate our prediction models and interpret them in relation to how voxels respond to images.

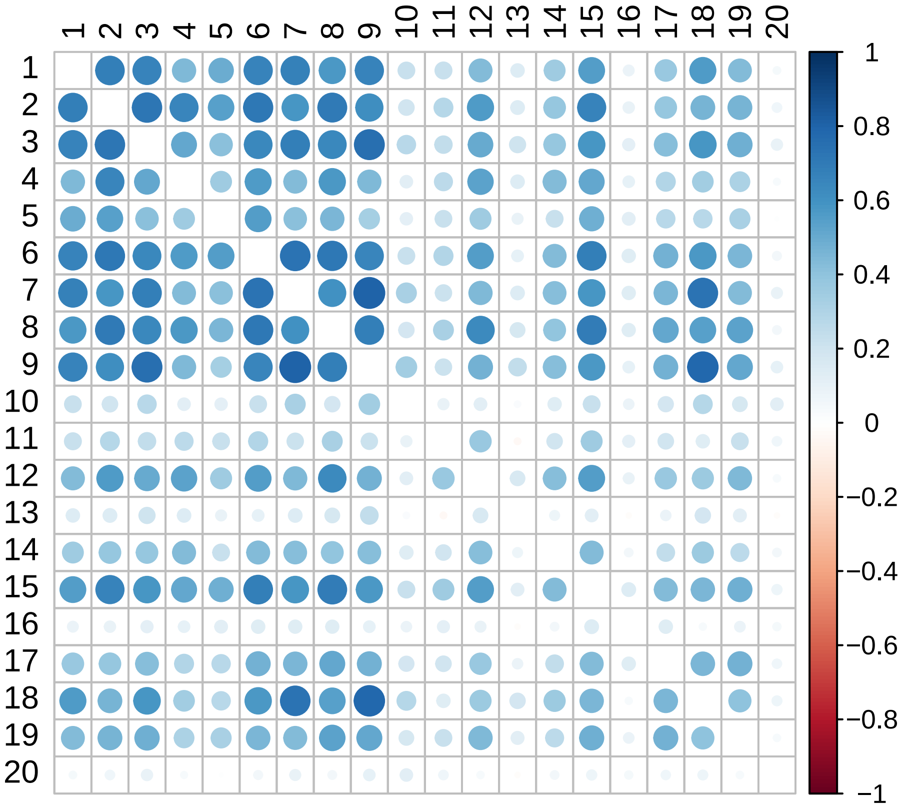
2 Exploratory Data Analysis

2.1 Data

The primary data that we analyze to build our prediction models are responses from the 20 voxels to 1,750 images. Each image is a 128 pixel by 128 pixel gray scale image, which can be represented by a vector of length 16,384. Each image vector is reduced to a length 10,921 feature vector through a Gabor transformation, while the response of each voxel to each image is reduced to a single number. Ultimately, we use a 1,750 (images) by 10,921 (Gabor features) Gabor feature matrix and a 1,750 (responses) by 20 (voxels) voxel response matrix to build and evaluate our prediction models. Additionally, we use our models to make predictions on a separate validation set of 120 transformed images without corresponding responses.

In Figure 1, we plot the correlation of the responses for the 20 voxels. We observe that there is high correlation for the responses of the first 9 voxels, especially voxels 1 through 3 and 6 through 9. Additionally, the responses for voxels 12, 15, and 17 through 19 show high correlation with the initial group of 9 voxels. Furthermore, the responses for voxels 13, 16, and especially 20 are largely uncorrelated with the responses for all other voxels.

Figure 1: correlation of responses for the 20 voxels



3 Methods

To start, we partition our data into a training set, a validation set, and a test set. We randomly assign 50% of our data to the training set and 25% each to the validation and test sets. We use the training set to fit our models, the validation set for model selection, and the test set to assess how well our final chosen model performs on new images.

After partitioning our data, we use regression methods to build a model to predict the response of all the voxels to new images. We predict each voxel's responses separately. We use ridge regression, lasso, and elastic net, selecting the tuning parameter for each using multiple model selection criteria.

3.1 Ridge Regression, Lasso, and Elastic Net

Ridge regression, lasso, and elastic net are techniques for regularizing the regression coefficient estimates by shrinking them towards zero. All three techniques help prevent overfitting the training data and can significantly reduce the varaince of the coefficient estimates. As with least squares, ridge regression, lasso, and elastic net seek coefficients that make the residual sum of squares small; however, they also impose a shrinkage penalty that shrinks the coefficient estimates towards zero and use a tuning parameter, λ, to control the extent of the shrinkage. When $\lambda = 0$, there is no penalty and the coefficient estimates are the least squares estimates. As λ increases, the coefficient estimates shrink towards zero.

How the three methods differ is the shrinkage penalty. While ridge regression imposes an l_2 penalty, lasso imposes an l_1 penalty. Unlike ridge regression, which shrinks coefficients towards zero but never exactly to zero, lasso will force some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large. Thus, lasso performs variable selection, leading to increased interpretability compared to ridge regression.

Elastic net imposes a combination of l_1 and l_2 penalties based on a mixing parameter, which can be beneficial when there is high correlation between features.

3.2 Model Selection Criteria

When performing ridge regression, lasso, and elastic net, we must select a value for the tuning paramater λ. We use cross-validation (CV), estimation stability with cross validation (ESCV), Akaike's information criterion (AIC), AICc, and Bayesian information criterion (BIC) to select the tuning parameter.

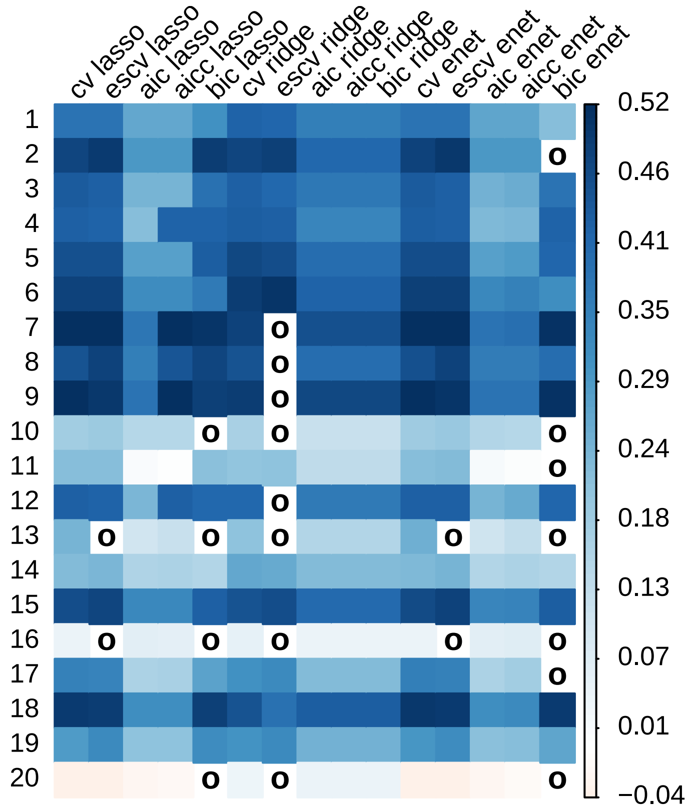
AIC and BIC indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting. The penalty on the training error increases as the number of predictors in the model increases, with BIC generally placing a heavier penalty on additional predictors and thus choosing smaller models. AIC often selects complex models that overfit the data in small sample sizes, so AICc adds a correction for small sample sizes to AIC. The main advantage of the information criteria is that they are easy to compute. On the other hand, their validity rely on model assumptions. Additionally, they are derived from asymptotic results, so even when model assumptions are satisfied, they may not work well in the finite sample case.

Conversely, CV directly estimates the test error by relying on data resampling. Commonly used to select the regularization tuning parameter, CV has many advantages. The method can be used with any loss function, provides a direct estimate of the test error, and makes fewer assumptions about the true underlying model. As for disadvantages, CV can be unstable, particularly in high-dimensions, with potentially large variability on the model chosen. ESCV, based on an estimation stability (ES) metric and CV, finds a locally ES-optimal model smaller than the CV choice so that the it fits the data and also enjoys estimation stability property. For dependent predictors, often the case in practice, ESCV performs better than CV for parameter estimation and significantly better for model selection. Their prediction performances are comparable, unless the predictors are independent.

3.3 Selection of Models:

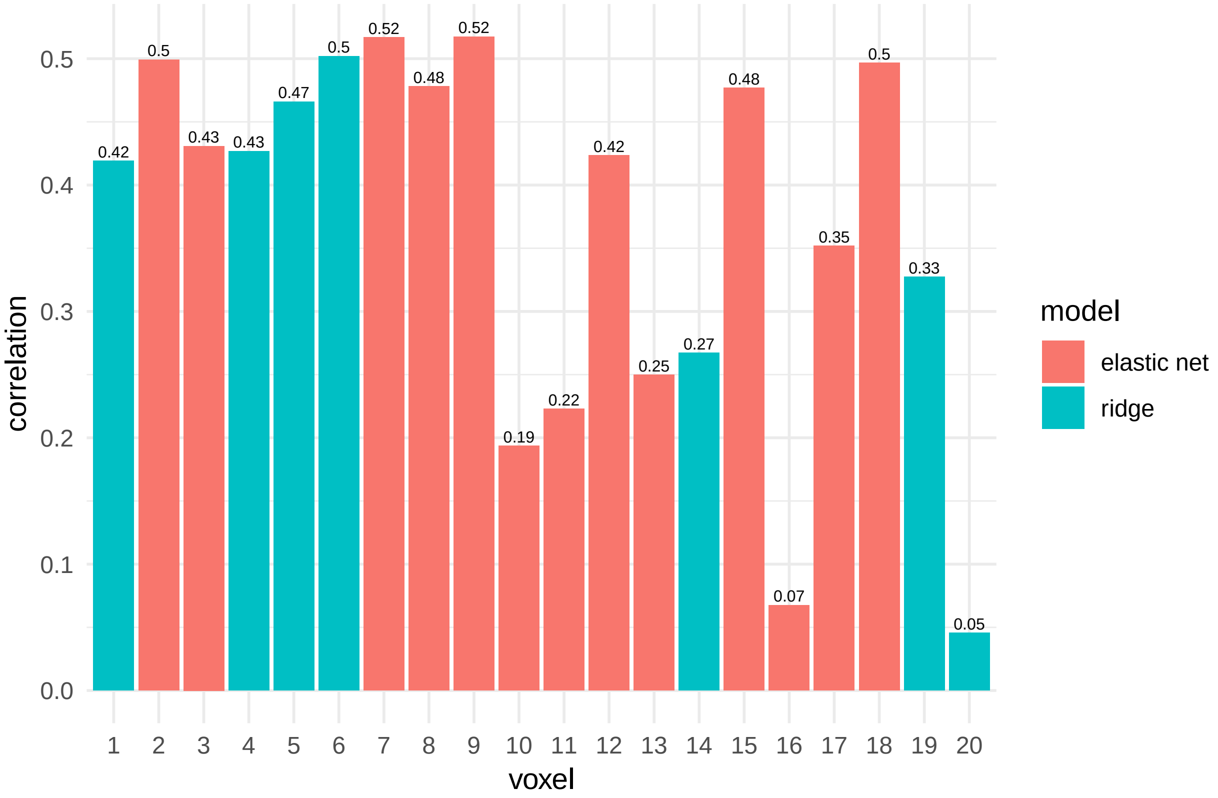
For each of the 20 voxels, we select optimal models for ridge regression, lasso, and elastic net (with the mixing parameter set to 0.5) for each model selection criteria (CV, ESCV, AIC, AICc, and BIC) after fitting on our training set. Then, with 15 potential models for each voxel, we select an optimal model for each voxel based on the correlation between the fitted values and observed values on our validation set, the same performance indicator used in the Gallant lab. Ultimately, we select elastic net models for 13 voxels and ridge regression models for 7 voxels, while we do not select the lasso for a single voxel. As Figure 3 displays, the correlations range from a minimum of around 0.05 for the voxel 20 ridge regression model to a maximum of around 0.52 for the voxel 7 and voxel 9 elastic net models.

Figure 2: correlation by model, criteria, and voxel



note: circles indicate models which predict the same response for each image, so the correlation cannot be calculated.

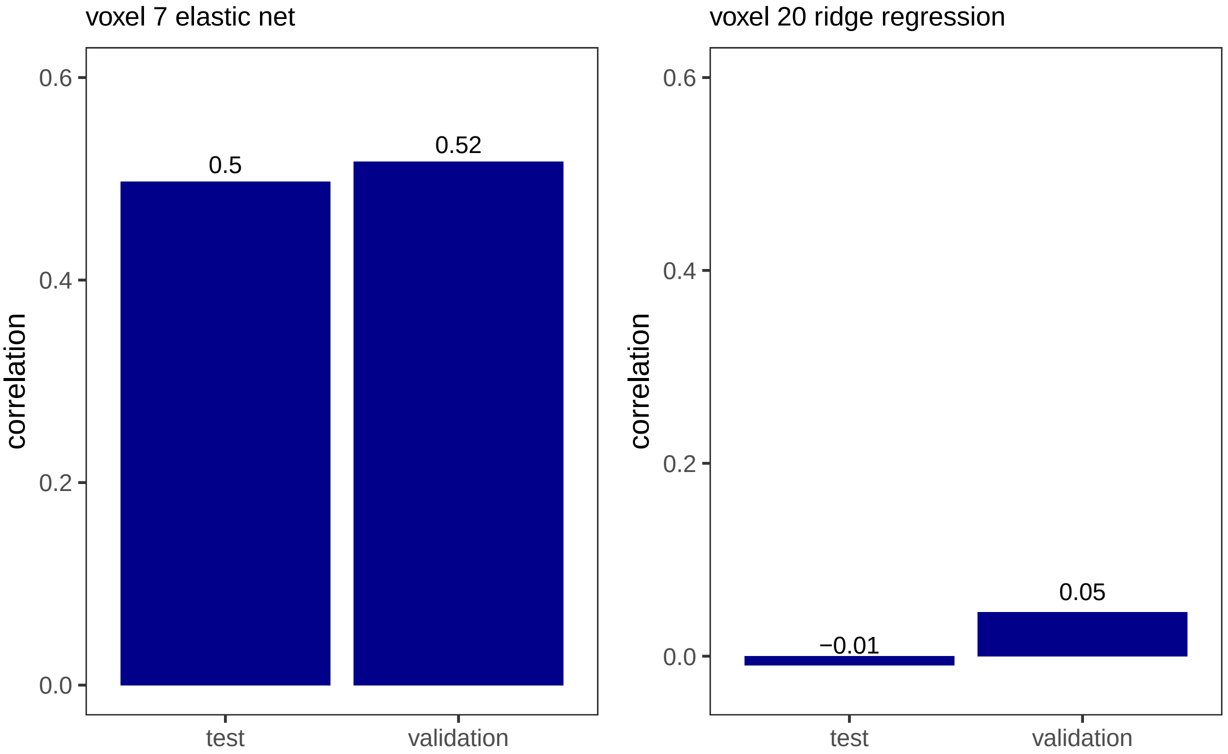
Figure 3: correlation between fitted and observed values by voxel



4 Diagnostics and Interpretation

For further investigation, we examine the models corresponding to the voxels for which we do the best and worst in predicting responses. The elastic net model with $\lambda = 0.24$ for voxel 7 and the ridge regression model with $\lambda = 1.34$ for voxel 20 are among our best and worst according to the correlation performance metric, respectively, so we choose those to examine. To assess the fit of our models, we begin by making predictions on our test set and similarly inspecting the correlation between the fitted values and observed values. As shown in Figure 3, both models perform slightly worse on the test set than they do on the validation set. Across all voxels, the models perform comparably on the test and validation sets. Using our selected model for voxel 1, we make predictions on a separate validation set of 120 transformed images without corresponding responses. We expect similar performance to the results we obtain when making predictions on our test set.

Figure 4: correlation performance metric for predictions on test set vs. validation set



To search for potential outliers, we inspect images for which our models perform particularly well and particularly poorly. For our voxel 7 model, we observe that 4 of the 6 images for which our model performs worst contain people, while none of the images for the 6 best predictions contain people. Figure 4 displays image 1,294, one of the images containing people for which our voxel 7 model inaccurately predicts the response. On the other hand, 4 of the 10 images for which our voxel 20 model performs best contain a monkey (for example, see Figure 5), while only 1 of the 10 images corresponding to our least accurate predictions contain a monkey. We might suspect that our model for voxel 7's responses does not give accurate predictions to images of people and that our model for voxel 20's responses is most accurate for images of monkeys; however, making these conclusions would require far more investigation and domain knowledge.

Figure 5: image 1294

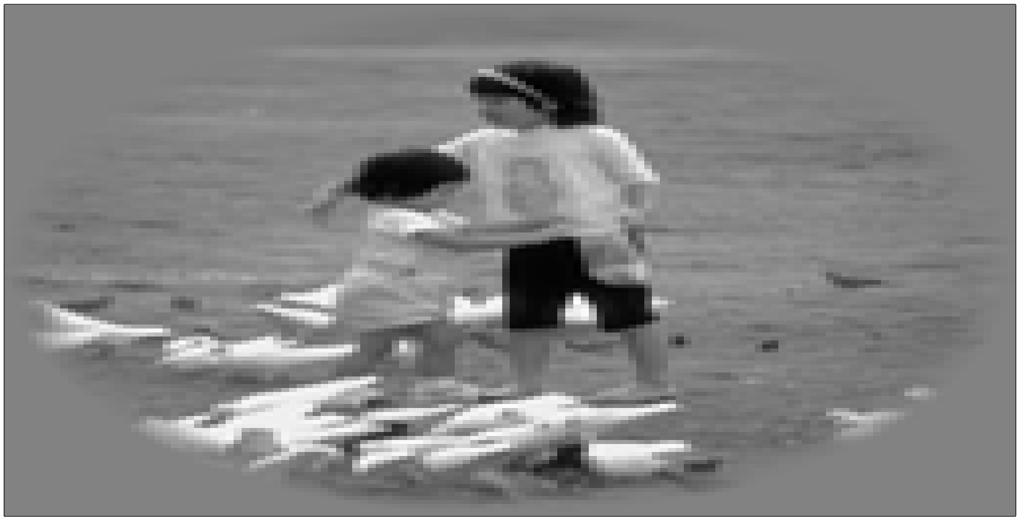
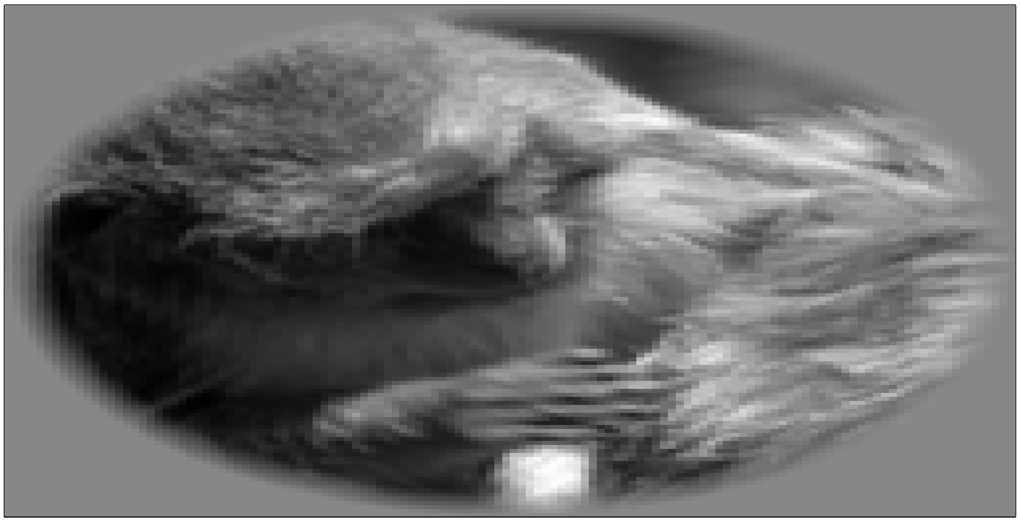


Figure 6: image 748



Next, we examine the stability of our prediction results and models. To do this, we fit our voxel 7 and voxel 20 models on 100 bootstrap samples from the training set and test on the validation set. For increased interpretabilty, we consider our lasso model for voxel 7 instead of the "optimal" elastic net model since they perform similarly. For both the voxel 7 and voxel 20 models, the prediction results are fairly stable. The correlation performance metric for the voxel 7 model among bootstrap samples has standard deviation equal to 0.02 and mean 0.50, compared to the 0.51 correlation we observe when training on the standard training set. The voxel 20 model's prediction results are a little bit less stable, with the correlation among bootstrap samples having standard deviation equal to 0.04 and mean 0.03, compared to the 0.05 correlation we observe when training on the standard training set. Figure 7 displays a box plot of the correlation values across the bootstrap samples for both models.

To assess the stability of the models, we examine the stability of the values of the selected tuning parameters and the features. The tuning parameter for the voxel 7 model among bootstrap samples has standard deviation equal to 0.06 and mean 0.13, compared to the 0.10 λ value from training on the standard training set. This seems fairly stable. The voxel 20 model's tuning parameter value seems less stable, with the correlation among bootstrap samples having standard deviation equal to 0.18 and mean 1.87, compared to the 1.34 λ value we observe when training on the standard training set. As Figure 6 shows, the range of lambda values from the bootstrap samples for the voxel 20 model does not contain the lambda value we observe. On average, the voxel 7 lasso model, which performs variable selection, includes approximately 43 of the 10,921 Gabor features, compared to the 50 features selected when training on the standard training set. Out of all features, the 5,721st feature is selected most frequently, being included in 88 of the 100 bootstrap models. Only 8 features appear in more than 50% of the models, while 986 of the features appear in at least 1 model. Since ridge regression does not perform variable selection, we examine the 100 features with the highest coefficient values for each of the 100 models from the bootstrap samples for the voxel 20 ridge regression model. Three features appear in the top 100 for all 100 models, while 165 features appear in the top 100 for at least 1 model. Overall, the features are not very stable, most clearly illustrated by the large variation in the features selected by the voxel 7 lasso model across the models trained on the bootstrap samples.

Figure 7: prediction correlation and lambda values across 100 bootstrap samples

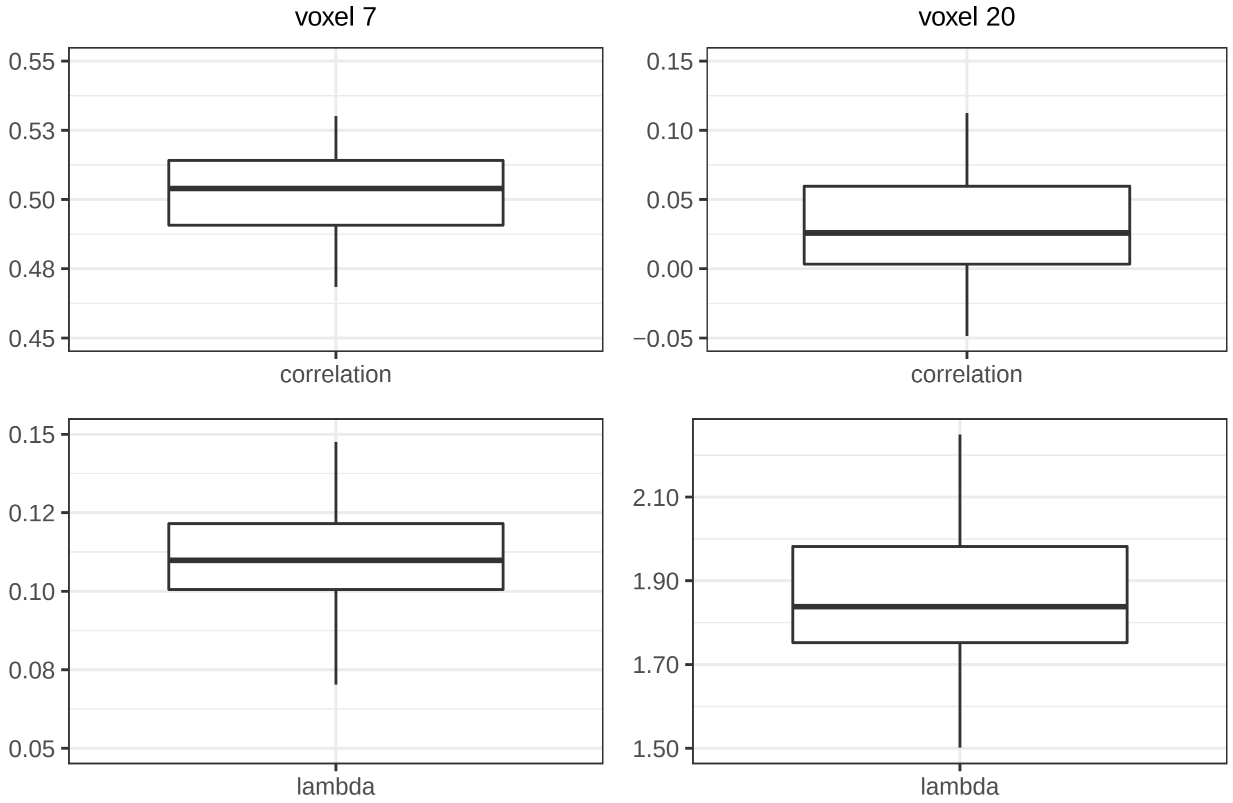
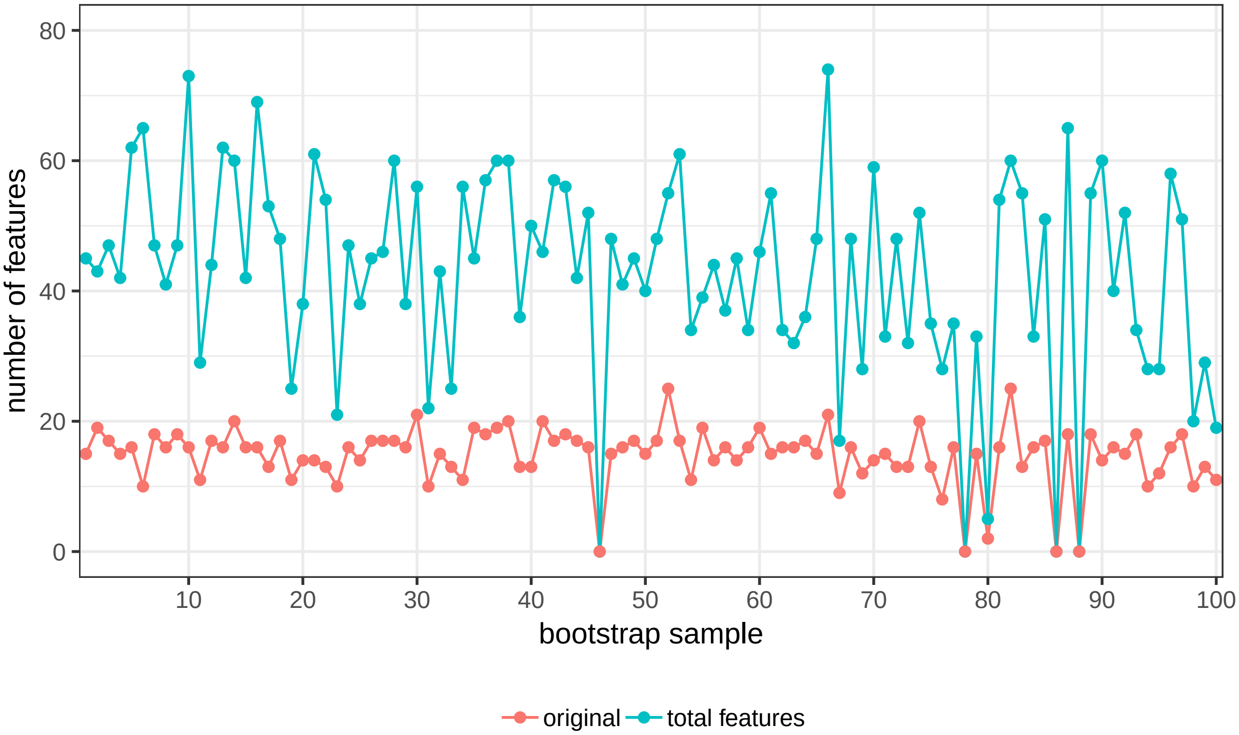


Figure 8: total features and original features by bootstrap sample for voxel 7 model



Before interpreting our models, we wish to highlight how difficult it can be to intepret results in high dimensions. In the high dimensional setting, we face extreme multicolinearity since any feature can be written as a linear combination of all other features. Thus, it is impossible to know exactly which variables are truly predictive of the outcome, and we cannot identify the best coefficients for regression. Since there are likely many sets of features that would predict the voxel responses comparably to our selected models, it would be incorrect to conclude that our models' features predict the responses better than the features not included our models. If we repeated the same procedures on an independent data set, it is likely that our models would contain a different, and maybe even drastically different, set of features.

Having emphasized the caution that must be taken when making interpretations, we attempt to interpret our models. Comparing the 50 features selected by lasso for our voxel 7 model and the 100 features with the highest coefficient values for our voxel 20 model, they share no features in common. In fact, across the 100 models across the bootstrap samples for both voxels, they share only feature number 10,876. Since there is high variability across bootstrap samples, we believe that the most important features are those that are stable across different bootstrap samples. For the voxel 7 lasso model, features 5,721, 5,699, and 1,544 appear in 88, 71, and 69 of the 100 bootstrap samples, respectively. For the voxel 20 ridge regression model, features 10,707, 4,011, and 9,387 are among the 100 features with the highest coefficient values for each of the bootstrap samples. Thus, there is evidence that these features are most important to voxels 7 and 20, respectively. We could do hypothesis testing on the estimated parameters in a similar way to the Gallant lab through a bootstrap procedure to estimate statistical significance of predictive power.

5 Conclusion

Overall, the fMRI data measuring the brain's responses to visual images highlight some of the complexities faced when conducting statistical analysis in high dimensions. While the predictions that our models make are reasonably stable, the features included are highly variable. Thus, when translating from the statistical model back to the real-world problem, it is important to do so with caution. For example, it would not be prudent to conclude that the features with high coefficients in the model predicting voxel 7's responses to images are the most important. In order to begin to make similar claims, you must examine the results on independent data and tie in domain knowledge.

6 References

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

[2] Lim, C. and Yu, B. (2013). Estimation stability with cross-validation (ES-CV). Available at arXiv.org/abs/1303.3128.

[3] Hastie, Trevor J., et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

[4] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: With Applications in R. New York: Springer-Verlag; 2013.