The generality of self-control

Chen, Kent	Lee, Rachel	LeRoy, Benjamin	Liang, Jane	Udagawa, Hiroto
kentschen	reychil	benjaminleroy	janewliang	hiroto-udagawa

December 3, 2015

Abstract

Self-control is an interesting field of behavioral research with broad implications for our day-to-day experiences. Being able to appropriately regulate and check our impulses and reactions to various everyday stimuli is necessary for maintaining health and high-functionality in society. Thus, relating a subjects ability to control risky behavior to an area of the brain or focusing on understanding what connects self control to neurological activity has been the focus of many studies. One approach to capturint these neurological facets is to use functional magnetic resonance imaging (fMRI). Our goal is to reproduce and extend the fMRI analysis for a Balloon Analogue Risk Task (BART) study described in *The Development and Generality of Self-Control* [1].

1 Introduction

The Development and Generality of Self-Control [1] and its associated fMRI studies are concerned with the relationship between impaired and normal self control, as well as similarities and differences across the brain relating to self-control. The paper in its entirety explores multiple studies (of multiple study types), but we will just focus on the third study. The full paper compares four different types of self control among healthy adults to see if they are related to each other. Very little relationship was found between these different behavioral tasks, in contrast to the vast majority of existing literature, which argues for a unified notion of self-control. So, we have decided to narrow our focus and data analysis approaches to just the Balloon Analogue Risk Task (BART) study, which purportedly measures control over risky behavior. fMRI scans from the study show blood flow to the brain, which may be relatable to control over risk-taking behavior during participation.

The rest of this report will detail the procedures from the original analysis of the data that we have tried to mimic. We experimented with different procedures to spatially smooth the voxels and the convolve and time-correct the time courses for each subject. After preprocessing the data, we turned to fitting simple and multiple linear regression models to each subject, with the option of including different study conditions. The resulting coefficients from the linear regression models can be used to perform t-tests to examine the significance of activity in different voxels. However, the validity of these tests is highly dependent on the validity of our model assumptions, such as the normality of our errors. Since we are performing a large quantity of tests, multiple comparison corrections are needed to control the false positive rate. We used k-means clustering to further identify regions of the brain across subjects with high activity and compared these results with the results of the original paper.

2 Data

The Balloon Analogue Risk Task (BART) measures risk-taking behavior by presenting participants with a computerized balloon. The participant can earn money incrementally by pumping up the balloon, but after an unknown threshold, the balloon will explode. At any time, the participant elect to cash out his or her earnings, but doing so eliminates the potential to gain additional money through pumps. If the balloon explodes, the participant loses all of the money for the trial. For this study, BART and fMRI data for 24 subjects was collected. The mean age of the subjects was 20.8, and ten of the subjects were female. Four behavioral variables were recorded for each subject: the average number of pumps for each balloon, the average amount of money earned across runs, the number of exploded balloons, and

the number of trials. There were also three model conditions: events for inflating the balloon (excluding the very last inflation of each trial), the last inflation before an explosion, and the event of cashing out (the balloon explosion was not included as an event). Of interest for our work is the blood-oxygen-level dependent (BOLD) imaging data recorded for each subject during the course of task. Each subject's BOLD data was recorded as 64 by 64 image matrices in 34 slices, with a variable number of time points.

Much of our analysis focuses on creating our own procedure for cleaning and preprocessing the raw BOLD data so that the true signal can be readily captured by our analyses. However, a cleaned version of the data was later made available by Ross Poldrack and the OpenfMRI project. The cleaned scans had receieved motion correction, high-pass filtering in time, and registration to the standard MNI anatomical template. We would eventually like to compare the results from using our cleaning procedure versus the provided procedures. One non-trivial issue that must be overcome is to design a reproducible pipeline to perform our analyses on either data format, since the preprocessed scans provided by OpenfMRI are in a space with different-sized dimensions.

It should also be noted that neither set of preprocessed data uses the exact same cleaning procedure as that used in textitThe Development and Generality of Self-Control [1], which describes using various software and black-box methods that may not be available for our use. Additionally, while the paper is mostly concerned with comparing the areas of neural activity across different types of self control, we focused only on a single study on a single type of self control, for feasibility reasons.

3 Methods

3.1 Smoothing

Due to the inherently random nature of human subjects and their movements, a certain level of smoothing must be performed on the spatial dataset. That way, the "noisy" data can be cast off from the data that actually represents significant changes in blood flow in the brain. By doing so, researchers and anyone else investigating the data will be able to distinguish between non-brain scans versus actual brain scans. Each voxel of the brain is represented by a measure of blood flow intensity, and so a series of steps must be taken so that the data is correctly convolved to most closely and accurately depict what was happening at a certain point in the brain at a certain time. After researching quite extensively, we decided to use smoothing involving a Gaussian kernel in order to smooth the three dimensional data. Originally, we were going to try and write a smoothing function from scratch, by implementing a rudimentary average-over-neighbors method. However, discussion with mentors lead us to the scipy module `ndimage.filters` that has a function to performs a Gaussian filter on n-dimensional data. This was exactly what we needed so rather than reinventing the wheel, we will be smoothing the data with this module. An in-depth discussion of the Gaussian filter can be found in the appendix.

3.2 Convolution and Time Correction

3.2.1 Convolution

Our study is structured around event-related neurological stimulus, not block stimulus as was discussed in the in-class example. We predicted the hemoglobin response (HR) related to this type of neurological stimulus.

There were multiple ways in which we approached the problem of trying to better replicate the predicted response function defined by:

$$r(t) = \sum_{i=1}^{n} \psi_i \phi_i (t - t_i) \tag{1}$$

where ψ is the amplitude of the response stimulus (always 1 in our case), and ϕ_i is the hemodynamic response started at the ith stimulation (t_i).

The five approaches we attempted can be divided into 3 subcategories: **(1)** a strict replication of 1; **(2)** a similar function as **(1)** that utilizes matrix multiplication; and **(3)** a function that splits the intervals between each scan (2 seconds) into a certain number of even slices, then puts the stimulus into the closed split, using `np.convolve` on this stimulus and a detailed hrf function, and then reducing the output back into the 2-second time intervals. Detailed exploration of this matter can be found in the appendix.

We compared these methods based on accuracy and speed. [Figure 1] displaces accuracy comparison, and the table in [Figure 2] show the accuracy based off of `ipython`'s `timeit` function.

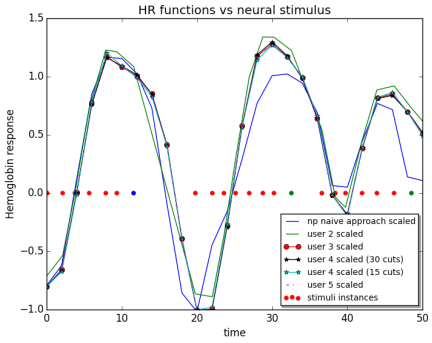

Figure 1: Different convolution functions vs. the Neural stimulus

name in graph	Speed per loop
np naive approach	14.4 μs
user 2	972 ms
user 3	1.15 s
user 4 (15 cuts)	98.3 ms
user 4 (30 cuts)	185 ms
user 5	110 ms

Figure 2: Speed to create HRF predictions for Subject 001, all conditions

The "np naive approach" run has the worst approximation and isn't even theoretically correct. As was stated above, this was the large motivating factor behind the rest of the convolution analysis. The "user 2" and "user 3" runs fall under subcategory **(1)**. The "user 2" was the first approach to matches the theory, but matches the stimulation times and not the scan times. The "user 3" is the most theoretically sound model (and is our standard for accuracy). The "user 5" falls under subcategory **(2)**, "User 5" is our matrix version of the theory, and has the same accuracy as "user 3". The "user 4" model fall under subcategory **(3)**, the methods that use the grid cut usage of `np.convolve` with notations for the number of slices between each scan.

3.2.2 Time Correction

The fMRI machine scans each voxel at a slightly different time. In our case, the lowest horizontal slice was scanned first, with the later scans moving progressively toward the top of the brain. So, there is an observable time drift that needs be accounted for. We did so by shifting the times of stimulus "backwards" for voxels scanned later to directly "correct" for the delay of the scan (assuming that each layer of the scan took 2/34 of a second).

3.2.3 Multiple Conditions

At the beginning of our analysis we tried to see if there was an advantage to separating our 3 conditions and creating separate predicted hemodyamic reposes for each to allow for different amplitudes for each type of condition. We did not observe a large difference in the $\hat{\beta}$ values we got, so we did not continue with this exploration. It is quiet possible that with the new approaches we've applied, differentiating the conditions for the stimulus might see gains in interpretability.

3.3 Linear Regression

A simple and straightforward way to model the voxel time courses is to perform simple and multiple linear regression. As a first attempt, we implemented and performed simple regression on a single subject's 4- dimensional array of voxels against the convolved time course, such that every voxel had an intercept and a coefficient corresponding to the convolved time course. However, examining the effects of the BART experiment conditions on voxel blood flow is also of interest. Thus, we turned to a more sophisticated multiple linear regression model that includes the conditions as dummy variable predictors. We also considered multiple regression including linear drift, Fourier transform and fitted mean terms. In each scenario, we created a design matrix X with the number of rows equal to the number of observed times and the number of rows is equal to the number of predictors.

Our specified model is then $Y = X\beta + \epsilon$, where X is the fixed aforementioned design matrix, Y is the 4-dimensional array of the subject's observed voxel time courses transformed into 2-dimensional space, and ϵ is random noise. ϵ is assumed to be independent and identically distributed with a normal distribution $N(0, \sigma^2)$, and it is also assumed to be independent of X. We will need to check the validity

of these assumptions, as discussed in the *"Normality Assumptions"* section. β is an unknown parameter that must be estimated using matrix algebra: $\hat{\beta} = (X^T X)^{-1} X^T Y$. To do this, we essentially flattened out the first three dimensions (which indicate spatial positions) into a single dimension, while keeping the fourth dimension (time) the same. The resulting $\hat{\beta}$ was transformed back into a 3-dimensional array to maintain the spatial relationships of the voxels.

Another possible way to model the voxels is to use the first two principal components of the voxel by time covariance matrix for each subject as the Y instead of the raw voxel time courses. This reduces dimensionality while still capturing much of the variance of the raw data. However, the choice to use two components is does not have an especially strong justification, as the proportion of variance explained by the components does not have a clear-cut "elbow" or plateau.

To consider the strength of the effects of these predictors, we looked at t-tests of the corresponding estimated coefficients for each voxel, as discussed under *"Hypothesis Testing"*. The validity of these t-tests and their corresponding "p-values" is largely dependent on how good our assumptions of linearity and normally-distributed errors are.

3.4 Normality Assumptions

The validity of our hypothesis tests of the estimated $\hat{\beta}$ values from linear regression is largely dependent on whether we can assume that the errors in our model(s) are independent and identically distributed from some normal distribution mean zero and constant variance. We focus here on checking the normality assumption. It is generally wise to use visualizations, such as residual vs. fitted values plots and quantile-quantile plots to inspect residuals for patterns and abnormalities. However, with the sheer quantity of data we are working with— each of the 24 subjects has $64 \times 64 \times 34$ voxels that can each in turn be fitted to a model— visual inspection is not practical.

For this reason, we turned to using the Shapiro-Wilk test for normality, which tests the null hypothesis that the data in question is normally distributed. A Shapiro-Wilk test was performed for each set of residuals corresponding to a single voxel's time course. That is, each test used around 200 observations, or the number of time points for that particular subject. 200 observations is not an especially large sample size, and for this reason, we express some concern because normality tests have low power for small sample sizes. Shapiro-Wilk may incorrectly fail to reject the null hypothesis due to this bias [2].

3.5 Hypothesis Testing

Our simple linear regression model was created to better understand the relationship between the voxels in a given subject's brain and the convolved time course. In order to measure the strength of the association between these two measurements, we ran a hypothesis test on the coefficients of the simple linear regression model for each subject.

There is an individual linear model associated with each voxel in a subjects image (and a total of $64 \times 64 \times 34$ voxels per subject). Thus we ran a t-test on each voxel's β coefficient that is associated with the HRF response. The null hypothesis for each test was that $\beta = 0$, with the alternative hypothesis that $\beta \neq 0$. Once we had obtained each t-statistic, we compared this value across voxels in two ways. First, we simply compared this t-values with voxels within a subject. In this case, we took into account the sign of the t-statistic in our analysis. Second, we converted this t-statistic to a "p-value", in which case the sign of the t-value will become irrelevant and we compared across voxels without taking into account this sign. Later, we also run a multiple comparison test using a BenjaminiHochberg in order to find the voxels that are most significant.

Having implemented a method to compare the voxels within a single subject, we next examined our results for the same voxels across subjects. Our initial approach was to aggregate the t-statistic data between all subjects for each voxel. This allowed us to decrease the variability of the fit on each voxel and detect a more clear signal.

In order to do this, we ran the hypothesis test as stated above on all 24 subjects of the study. Then for each voxel, we took the average of the t- statistics across the subjects. An issue with our data was the presence of empty space detected by the scanner that is not directly part of the brain. To account for this, we took the masked data of the brain and "cut out" the parts of the images that were not relevant to our analysis. Ultimately, we were left with a single 3-d image with each voxel representing the average t-statistics across all subjects in the study. This image will later be used in our clustering

step in order to pinpoint the regions of the brain that have the strongest relationship with the convolved time course.

3.6 Benjamini-Hochberg Correction

When conducting multiple comparisons, it is important to have an idea of the quantity of Type I errors that may be prevalent in the analysis. In our analysis of voxel data, we decided that limiting/controlling the number of Type I errors is important to the process. The processes of limiting the number of Type I errors are called FDR-controlling procedures. In the grand scheme of things, FDR-controlling procedures give greater statistical power with the cost of more Type I errors that can fall through.

Once we implemented the hypothesis function that would return t-test values and "p-values", we implemented the Benjamini-Hochberg procedure to control the proportion of rejected null hypotheses in the data. The Benjamini- Hochberg procedure works by multiplying each of the "p-values" to a ratio of the number of tests and the chosen false discovery rate – from these adjusted "p-values", only the values that are less than the chosen false discovery rate will be chosen to be returned. This way, we are able to adjust the proportion of null hypotheses that will be rejected and the proportion that will return the desired proportion of significant tests. This will reduce the number of false positives returned in the data and extend greater statistical power in later analysis performed on the voxel dataset.

3.7 Clustering

Now that we have the across-subject average t-statistics for every voxel in the brain, we are left with a 3-d array of t-statistics that contain both negative and positive values. Instead of manually observing patterns in these images, we instead implemented a clustering algorithm to split the entire 3-d images into clusters based on the voxels' relative location to each other as well as the values of their t-statistics.

In order to find a proper clustering algorithm, we decided to treat this problem like a grayscale image segmentation problem and implemented a agglomerative hierarchical cluster using Ward's method. Agglomerative means that the clusters are built bottom up which each observation starting as its own cluster and pairs being moved up the hierarchy. Ward's method creates clusters based on a minimum variance criterion that miniizes the total within-cluster variances. An example of this implimentation for a 2d image is seen here: `http://scikit-learn.org/stable/auto_examples/cluster/plot_lena_ward_segmentation.html`.

In our implimentation, we defined a structure to our data using a connectivity graph in order to ensure that each cluster is spatially constrained. Also, since our scenario uses a 3d image, the connectivity graph will also have to take into account this extra dimension.

Ultimately, the goal of clustering is to have a better understanding of which parts of the brain are related to the signal based on the t-statistics from performing hypothesis tests on the coefficients from linear regression. Once we obtain our clusters, we will both measure the within-cluster mean of t-statistics as well measure the centroids of the clusters. By doing this, we hope to see the parts of the brain that have the strongest relationship with the signal and compare them to the results of the origin research paper.

4 Results

4.1 Linear Regression Results

To develop linear models, we looked at the HR from the neural response as a single feature. Originally we used multiple regression to take into account the 3 different types of stimulus (pump, explode, cash-out) to see if the separation of these stimuli can better describe the response, but it wasn't that good. In figure 5, we can see the different conditions broken up. Using smoothed data, fourier and drift features we got the following fitted values and residuals for a random voxel for subject 001. [Figure 3, 4].

As we also obtained $\hat{\beta}$ values (coefficients) from the linear regression models, we looked at the 3-dimensional reports of the $\hat{\beta}$ values, a less rigorous analysis than hypothesis testing with t-statistics [Figure 6].

The numerous other multiple regression models discussed in *Linear Regression* should be analyzed similarly in the future.

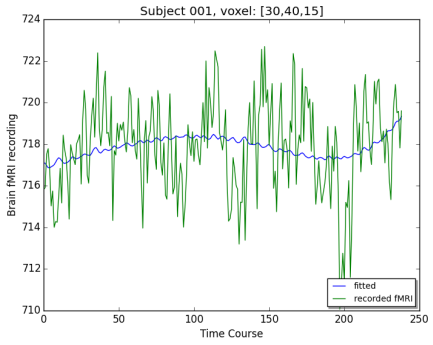

Figure 3: Fitted vs Actual

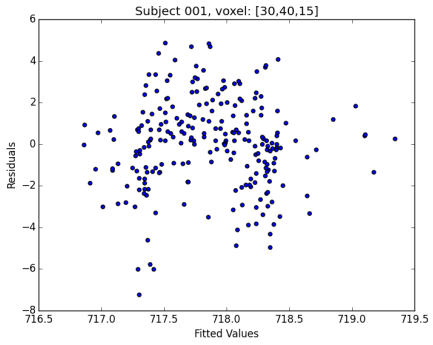

Figure 4: Fitted vs Residual

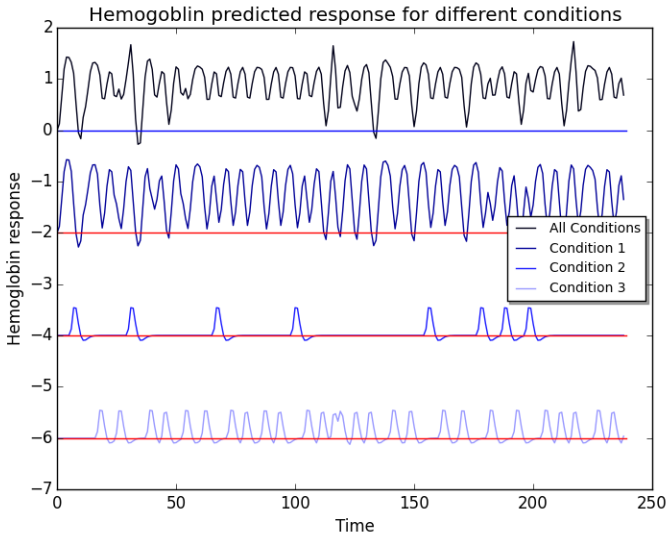

Figure 5: Plotting all predicted HR for conditions.

4.2 Hypothesis Testing Results

The results of our simple linear regression t-statistic comparisons across subjects are shown in [Figure 7]. We can see each slice of the brain from top to bottom in each section of the image. The blue areas shows parts of the brain that had a negative t-statistic while the red parts of the image shows parts of the brain that had a positive t- statistic.

The parts of the image that were cut out by the mask are white so we can more clearly see the contrast in our results. Based on a cursory look at this image, we can see a pattern of dark red (high positive t-statistics) in the lower left parts of the brain, and area of dark blue (high negative t-statistics) in the center and lower middle parts parts of the brain.

4.3 Clustering Results

Not written yet.

5 Discussion

5.1 Discussion of Results

While very much still a work in progress, our analysis thus far includes methods for both data processing and modeling voxel time courses. Prior to doing any serious analysis, we had to smooth the

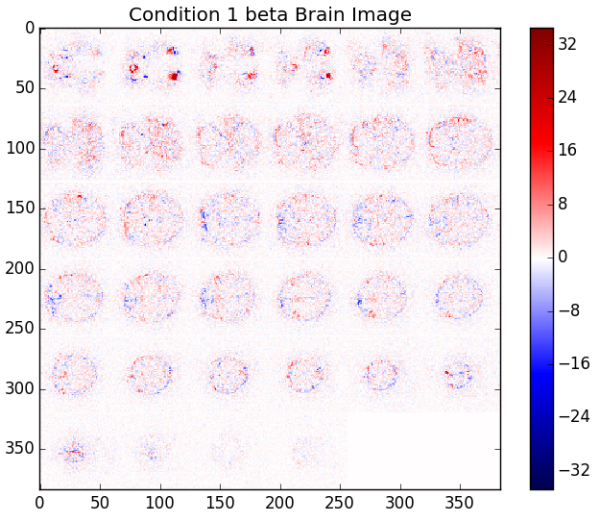

Figure 6: $\hat{\beta}$ values for condition 1, subject 001.

data spatially for each subject. We also generated a reasonable convolved time course with time shift corrections, based on event-related neurological stimuli with non-constant intervals.

A basic but nevertheless important model to consider is linear regression. We implemented both simple and multiple regression models at the individual subject level. In addition to the convolved time course, our multiple linear regression models attempt to account for more of the noise in our data by including terms for event conditions, linear drift, and Fourier transforms. We then checked the assumptions for the fitted models and performed hypothesis testing on the resulting coefficients for each voxel. Though the linear regression models were designed to handle each voxel for each subject's data individually, we aggregated the data across the 24 subjects by taking the means of the t-statistics corresponding to each voxel. However, one major concern for hypothesis testing is the issue of multiple comparisons, which we attempted to address using the Benjamini-Hochberg procedure. Finally, we used k-means clustering to further identify areas of the brain with high neurological activity during the events of the BART study.

5.2 Discussion of Future Work

We have specified several models for linear regression and considered various methods for hypothesis testing across multiple voxels and subjects. Now our main objective is to distill all of our work into interpretable results that can be used to identify brain regions with high neurological activity during the course of the BART events.

Additional future work will concern reproducing our analysis on the clean version of the data provided by OpenfMRI and comparing those results with the results from using our own preprocessing techniques. One issue that arises with the addition of the new cleaned data is that the dimension sizes of the scans are different, which will create challenges for comparing the two data sets.

If we have the time, permutation tests, which have few assumptions and are easy to interpret (but computationally intensive), would be a useful tool for identifying brain regions with significant activity. Additional tests and checks for model assumptions would also be valuable for assessing the appropriateness of our existing hypothesis testing.

References

[1] J. R. Cohen, *The development and generality of self-control*, ProQuest, (2009), p. 164. "`http://gradworks.umi.com/34/01/3401764.html`".

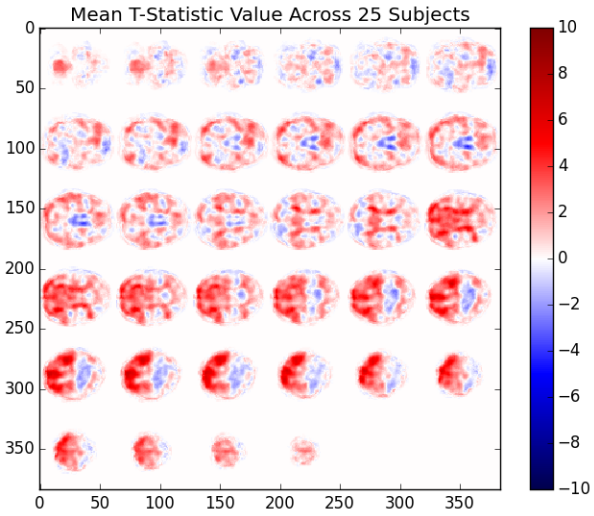

Figure 7: Across-subject mean of t-Statistic per voxel.

[2] A. Ghasemi and S. Zahediasl, *Normality tests for statistical analysis: a guide for non-statisticians*, International journal of endocrinology and metabolism, 10 (2012), p. 486. "`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3693611/`".

6 Appendix

1. Outlier Removal
2. Convolution Analysis
3. Smoothing
4. Clustering (not written yet)
5. Time Series Analysis