Project Epsilon progress report
The Neural Basis of Loss Aversion in Decision-Making Under Risk

Jo, Min Gu `mingujo`
Kaam, Soazig `soazig`
Li, Zhuangdi `lizhua`
Yu, Timothy `timothy1191xa`
Zhi, Ye `ye-zhi`

December 1, 2015

1 Introduction

The study *Neural Basis of Loss Aversion in Decision-Making Under Risk* [1] focuses on decision-making process, especially on the correlation between the neural activity and the reluctance to lose. 16 people were presented 255 gambling situations with a 50% of success. Each situation was associated with a potential gain and loss that were randomly selected. The gains were ranging from $10 to $40 while the losses from $5 to $20. The participants were asked to assess their level of willingness to accept or reject the gamble using a 4-point likert scale [1: strongly accept, 2: weakly accept, 3: weakly reject, 4:strongly reject]. The response time was also recorded for each case. The imaging data were collected using the fMRI method. They were processed and analyzed in order to identify the regions of the brain activated by the decision making process. This study also investigated the relationship between the brain activity and the behavior of the subjects towards the gambling situations using a whole-brain robust regression analysis.

2 The data

The data we are using can be found on the OpenfMRI website at the following address: `https://www.openfmri.org/dataset/ds000005`, the dsnum is ds005. For our project, we are specifically using the behavior data and the BOLD data that are organized.

For each of runs per subject (3), the behavior data contains the timestamp of each survey question (onset), the gain/loss combinations (gain and loss), the response for the particular trial (respnum) from the 4-point likert scale. The researcher created a response category (respcat) to be used in their binary choice model that combines the “reject” answers together on one hand and the “accept” answers together on the other hand. BOLD data contains compressed 4-dimensional brain images for each subject’s run. The folder also comport Quality Assurance (QA) files and a report.

3 Our Work

3.1 Initial work

First, we downloaded the data and used the checksums.txt to validate. However, there was one issue that downloading checksum.txt can also be distorted to validate the data. After exploring the structure of data folder, we did some data fetching and preprocessing procedures. For behavior data, we wrote a function to merge three runs for each subject since we would like to look at the total observations within one subject. Also we noticed that in the data set of each run, there existed observations with ”-1” in ”respcat”, which was meaningless and might be an error during the experiment so we took them out. For bold data, we also wrote a function to unzip all nii.gz files for further analysis.

3.2 Behavior Data

We did some explanatory data analysis and regression analysis using behavior data. For explanatory data analysis, we generated some summary statistics, including correlation among variables and simple plots to better understand the behavior. And then we used regression analysis to mainly answer two scientific questions. The scientific questions that we have are:

- If gain/loss would be significant for individuals who choose to participate and how much time it would take for them to respond.
- If gain/loss would be significant for whether individuals would like to participate in the gamble.

3.2.1 Linear regression

To answer the first question, we built three linear regression models.

1. Response Time ∼ gain + loss
2. Response Time ∼ diff(gain-loss)
3. Response Time ∼ ratio(gain/loss)

Result
We use the result for one subject to explain our findings. In the first model, p value for the predictor loss is 0.002644 while p value for the predictor gain is 0.547262. Basically, we can conclude that people would actually care more about loss than gain. In the second model, p value for the predictor difference is 0.413985. It turns out that we can conclude that the difference between gain and loss wouldn't be a factor for how much time it would take for a individual to respond. In the third model, p value for the predictor ratio is extremely small, showing that ratio has a huge impact on individuals' response time.

3.2.2 Logistic regression

To answer the second question, we fit logistic regression between Accept/Reject Gamble and gain/loss. According to our analysis, the decision to whether take the gamble of most of subjects, in general, is more affected by loss amount rather than by gain amount. For example, below is the analysis on the subject 3. The regression line shows that it well follows the border between the two decisions: 1 (gamble) and 0 (not gamble). Right side of the line illustrates the decision to not gamble and it takes up more area relative to the opposite decision.

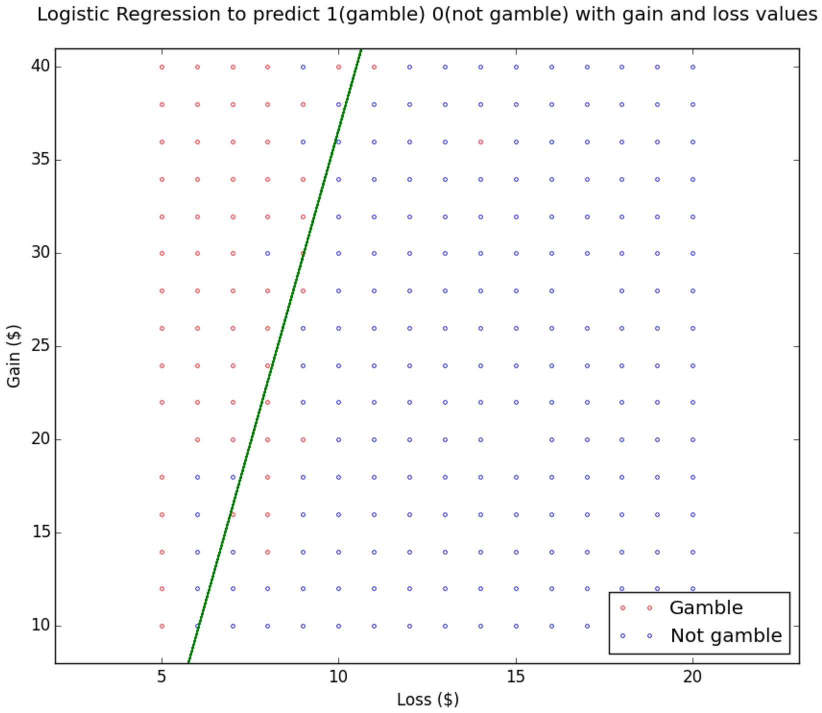

Figure 1: Logistic Regression for Subject 3

3.3 BOLD data

For image data, we did several explanatory data analysis to better under the BOLD data.

- Reproduce Quality Assurance(QA) plots
  We reproduced some of the QA Plots provided in the BOLD folder, including mean signals, Framewise Displacement and DVARS(root mean squared signal derivative over brain mask). The shapes were almost the same, except the y axis. We assumed there was some transformation or preprocessing in the original QA plots.

- Correlation
  We calculated the correlation between task-on/task-off vectors and voxel time courses to identify the active region of the brain. Below is an example of correlation figure of the middle slice in subject 1.

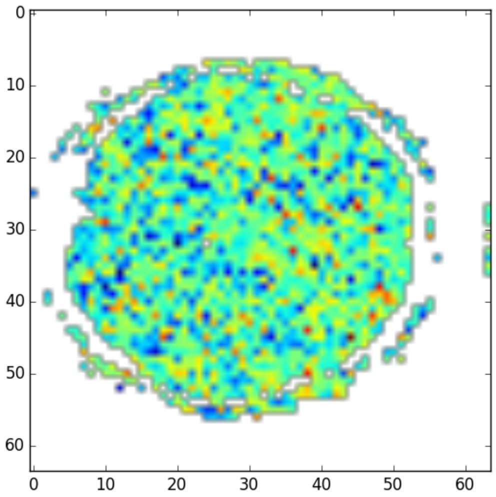

Figure 2: Correlation between On/Off and Voxel Time Course

- Associated the brain image data with behavior data
  We started to associate brain image data(mean/standard deviation across time courses) with gain/loss/ratio, choices(accept/reject) and respcat(strongly/weekly accept/reject).

4 Our Research Plan

4.1 Behavior data

- Use other classification methods than logistic regression to predict gamble/not gamble
  For now, we fit logistic regression to use gain/loss/ratio to predict whether a person would take the gamble. Also, we might try other machine learning/classification methods, such as K-nearest neighbor and Tree based model. Other than using gain/loss as predictors, we can use other independent variables such as subjects' response time and their confidence level in their decision to see how they affect their decision. We can select our train set to classify the whole data for each subject and test the classification.

- Explore correlation between neural activity (image data) and behavior data (survey data)
  For example, we can try to fit linear regression model between each voxel and gain/loss/ratio and other predictors.

4.2 Model validation

- Check assumptions of the regression models: normality, independence, equality of variance

- Use cross validations to test our model accuracy: divide our data sets into testing and validation sets

4.3 Modeling voxels for each participant

- Use convolved hemodynamic response and linear regression for each participant
  We can train the model using two randomly selected runs and validated the model using the third run.

- Investigate for more suited shape parameters for the Gamma function primarily with literature review

4.4 BOLD image data analysis

Prior to analysis, we would undergo preprocessing of our image data by slice timing correction. We will try to shift each voxel's time course so that we can assume as if they were measured simultaneously. Then, using convolved data on our image data of subject from HRF as a beta of our linear regression model, we might try to predict each voxel of an image and later combine to have a 3D-image of a person who decided to gamble or not.

5 Our Process

5.1 Challenges

One of the biggest challenges was to understand the paper and the fMRI data. Especially in our case given that we want to make a literature review in order to find the proper parameters (especially) the shape parameter) of the Gamma function used to model the hemodynamic response. As mentioned before, we also had difficulties with the workflow on git and version control management, especially with the branch management and Travis checks. Consequently, our code review process was not working well. As a team, because of our conflicting schedules, we had difficulties to meet at 5. Besides the class time, we were eventually able to set up a weekly meeting. The rest of the communication was done via emails essentially or on the GitHub platform.

5.2 Improving reproducibility

In order to be successful, we will need to improve the reproducibility of our project. For that, we need to work on adopting a systematic approach for code organization following the combo: functions - scripts - tests, and also our code syntax following python guidelines (PEP 0008). We also need to work on our code review process. We also need to generate the supporting documentation to improve the usability of our project. We also would like to exchange with other groups working on the same study in order to complement our understanding of the data and analysis.

6 Problems and Issues

When approaching the project, we have faced many problems. They can be divided into three main categories.

6.1 understand the data

The first main problem that we have encountered is about understanding the data. Although we have a large dataset, there is no specific file that introduces the data structure or explain the meaning of the variables. Due to the lack of documentation of these relative information, we have spent much time reading through the paper and searching the website. Even so, there are still some problems unsolved: for example, we had no ideas about the variable called "PTval" in the behavior data and did not use it for our analysis. Meanwhile, since the data we are analyzing is fMRI data and we are all unfamiliar with its format and processing procedure, therefore, in the beginning, we found that the technical field studies are difficult to understand. The insufficient description of the analysis methods also made us hard to reproduce the work. For example, there are no descriptions about the QA section. When we are trying to reproduce the plots or calculate certain statistics, we can not get the exact results: the plots have the same shapes and patterns but the values on the y-axis are more than two times smaller than those in the QA report. As a result, we doubted that they might perform some transformations on the data set but did not mention them in the paper. We also attempted to do a generalized logistic regression for the behavioral data by including gender and age of each subjects, but there was no specific guidance on how to perform this.

6.2 Coding

The second main problem that we have faced is about the coding. In order to assure the reproducibility of our analysis, we need to do everything from the terminal, instead of simple point and click. We write a lot of functions, scripts and tests to download, unzip, load the files and travel through different folders. We also use Git to do version control. However, it might cause some conflicts or misunderstandings when we just merge the pull requests. In order to avoid these situations, we keep track of others' code, try to add more descriptions to our code and review others' pull request more carefully.

6.3 Data Analysis

The hardest problems come from the data analysis. Our image data and behavior data for each run and each individual are in different folders, so sometimes it's hard to work on data across subjects and runs. Meanwhile, the length of behavior data, which is 239, does not match the length of image data, which is only 86. Therefore, when we tried to combine them together and check if there are some linear relationships between them, we didn't how how to deal with the difference. If we fill in NA, we will lose much information; if we fill in 0, it is meaningless to run a linear regression that have many zeros in it. We also have many questions regards to the convolution. In the lecture or our homework, the convolution matrix is given to us, but now, we need to create a convolution matrix by ourselves. We are also unable to determine the parameters of the gamma distribution for the HRF. Most of online papers are using gamma quadratic (2 gamma functions + intercept term) and the common practice is to use 6 and 12 and an intercept of -0.35; however, the results are not reasonable for our project. We will continue working on it, try different parameters and find more online sources.

7 Feedback on the class

Here are a few feedback we have for the class:

- We appreciate the model with 3 (or 4) supervisors having their own expertise.
- We would like more exposure to machine learning.
- The lectures on git workflow and collaboration were very useful but fast-paced.
- The lecture on linear algebra was a good refresher but a little bit too theoretical for this project-based class.
- We would like more linear regressions course focusing on the implementation.

We also have a ideas of improvement for the class:

- We propose to support the lecture with slides or handouts with the fundamentals (e.g. git) command for collaborative work) we can refer to after class.
- We propose to provide a support document with the most used statistics definitions for a good analysis design and interpretation of our data.

References

[1] S. M. Tom, C. R. Fox, C. Trepel, and R. A. Poldrack, *The neural basis of loss aversion in decision-making under risk*, Science, 315 (2007), pp. 515–518.