Adjusting for multiple testing problems in fMRI data

Claire Guo, Kim Ting Li

STAT30850 SPRING 2018

Final Project

Introduction

fMRI data can be used to answer many biologically interesting questions, but it is first crucial to address the challenging multiple testing problems that the data structure poses. When comparing the activity of thousands of voxels collected during brain imaging under several different settings, a large number of p-values are generated and thus a major multiple testing problem is created. Moreover, due to the complex functionality of the human brain, voxels are not independent but are correlated across the entire brain with various strength and patterns. This lack of independence violates the underlying assumptions of the multiple testing adjustment procedures we studied this quarter. Thus we set out to examine the performance of the methods we discussed in class when correlations are present, test modifications to those methods under inspirations from the real data, and use the best method to answer experimental questions about the fMRI dataset. In order to test these methods, we needed to avoid creating an additional layer of multiple testing problems by repeatedly using the real fMRI data to find the best adjustment method. We therefore needed to come up with simulations that we believe would closely follow the structure of the real data and use the simulated data to assess the performance of adjustment methods. In summary, our project addresses two main problems: which adjustment method or modification works the best and how to create simulations that would closely resemble the pattern of the actual fMRI data to verify the effectiveness of the methods.

For this project, we are using the fMRI dataset originally collected by Marcel Just and his colleagues in Carnegie Mellon University's CCBI. It contains the brain activities of one subject (subject 04847) across 40 trials. During each trial, the subject's brain reactions were divided into 54 different time intervals which were either a period of stimuli-absent resting state or a period

of sentence and picture presentation. For each of these time intervals, 4698 voxel activities were collected.

Group-adaptive Benjamini-Hochberg procedure based on ROIs

The original dataset has assigned each of the 4698 voxels into 24 different regions of interest (ROIs), with the exception of 7 voxels which were left unassigned. Inspired by the group-adaptive Benjamini-Hochberg (BH) procedure we encountered in Problem Set 2, we considered it reasonable to expect that each ROI would have a different proportion of nulls as regions play different roles and do not all activate uniformly. We therefore thought the group adaptive BH procedure could be appropriate and consequently chose to group the voxels by ROI. The ROIs varied in size, with the smallest ROI having 6 voxels and the largest having 440 voxels. The details of the group-adaptive procedures are shown below:

Group adaptive Benjamini-Hochberg procedure:

For G groups of sizes $n_1, n_2, ..., n_G$ and $n = n_1 + n_2 + ... + n_G$:

(1) For each group, estimate $\hat{\pi}_0^g = \frac{\#\text{ of p-values in group g which are } >\gamma}{n_g(1-\gamma)}$

(2) Define $\tilde{P}_i = \hat{\pi}_0^g P_i$ for each p-value in group g

(3) Find the largest k such that there are k many p-values $\tilde{P}_i \leq \frac{\alpha k}{n}$, but never rejecting any $P_i > \gamma$

From STAT30850 Problem Set 2 by Rina Barber

To create simulated data to test the procedure on, we assumed that all the voxels within the same ROI are uniformly correlated (with voxel variances to be 1), while voxels in different ROIs are independent from each other. Using these assumptions, we simulated signal p-values by converting randomly generated Z values from a multivariate normal distribution with a mean of 2 (a vector of 2s) and our created correlation matrix as $\mathbf{\Sigma}$, along with null p-values generated by random samples from the Uniform(0,1) for each ROI. To illustrate the performance of the group-adaptive BH under correlated scenarios, we varied the proportion of nulls and the

covariance between voxels in the same ROI and compared the resulting false discovery rates (FDR) and powers. We tested the method under the conditions that 10%, 50% and 90% of the data in each ROI were signals while the others were nulls. Under each condition, we then compared the effects of setting the correlations between voxel pairs within an ROI to be 0.1, 0.5 and 0.9. We used a γ of 0.5 across all tests. For each simulated condition, the setting was repeated 250 times and we calculated the FDR and mean power from these 250 iterations (Table 1). The resulting FDRs were very close to 0.05, which was our α significance level. However, the power decreased as the proportion of nulls increased which indicated that the method is unreliable when examining subtle differences where the majority of the results are in fact nulls.

Table 1: Group adaptive Benjamini-Hochberg procedure on simple simulation setting

Proportion of Signals/Nulls	***Correlation within ROI***	***FDR***	***Power***
0.5	0.5	0.051	0.535
0.5	0.9	0.052	0.535
0.5	0.1	0.051	0.535
0.9	0.5	0.052	0.985
0.1	0.5	0.055	0.118

Of course this simulation is an overly-simplified version of the real data we have because, in reality, brain voxels are highly unlikely to be correlated uniformly. Furthermore, different regions of the brain will still be correlated as the brain's activities are highly interconnected with constant communications between regions. However, the results under this simple setting led us to think

that the group-adaptive BH could work well when there is a high proportion of signals, even though the power was not ideal.

Therefore, we then tested the group-adaptive BH on a simulation that more closely resembles the real structure of the fMRI data. We computed the covariance matrix from the real dataset without examining the labels so as to not create another multiple testing problem, and simulated signals using this covariance matrix. We then determine a proportion of nulls we want to use and randomly substitute the correct number of signals with null p-values we generated from Uniform(0,1). This setting allows us to more accurately capture the structure of the data, without actually looking at our data and creating a higher level of multiple testing issue. We again tested the group-adaptive BH procedure with different proportion of nulls, with the results shown in Table 2.

Table 2: Group adaptive Benjamini-Hochberg procedure on simulated data using MRI covariance matrix

Proportion of Nulls	***FDR***	***Power***
0.05	0.05	0.98
0.2	0.21	0.67
0.5	0.5	0.29

At an initial glance, the FDR and power of the group adaptive BH seemed promising when the proportion of nulls was low. However, we noticed that the FDR closely followed the proportion of nulls that we set. This indicates that the group adaptive BH procedure is not performing better than random, motivating us to test other adjustment methods.

Regular Benjamini-Hochberg procedure with a modifying factor c

After we found that the group adaptive BH procedure was not appropriate for our data structure, we decided to try the original Benjamini-Hochberg procedure, with an additional modifying factor that we could determine based on the performance of the method on the real data structure:

Data determined modified Benjamini-Hochberg procedure:

(1) Find the largest k such that there are k many p-values $P_i \leq c\frac{\alpha k}{n}$, but never rejecting any $P_i > \gamma$

We now compare the p-values to $c\frac{\alpha k}{n}$ instead of $\frac{\alpha k}{n}$ and use the performance of the procedure on the simulated data when c = 1 to determine how we should modify the value of c. When running the regular BH method with c = 1 on our simulated data based on the real data covariance matrix, we got an FDR of 0.02 and a power of 0.49. As we expected, regular BH applied on the correlated data turned out to be overly conservative. The low level of FDR and power indicated to us that we could afford to raise our cutoff value for judging if a p-value is a real signal or not; in other words, we could explore values of c larger than 1. With that in mind, we experimented with a range of different values of c from 1 to 3 (preliminary examination showed us that when c = 3, FDR exceeded 0.05). The results are shown in blue on Figures 1, Figure 2(a), and 2(b) and in Table 3. We can see from the graph that the best combined result of low FDR and high power occurs when c takes the value of 1.6 (Table 3).

Modified Benjamini-Hochberg procedure on spatially correlated simulated data

While using the covariance matrix computed from the real data to create simulated data is a plausible method, it also has its shortcomings. For instance, because we have not examined the conditions under which the covariance matrix was generated, there could be a lot of noise in the correlation patterns. Therefore, we also attempted to capture the structure of the

data by using spatial correlation. Using the 3-dimensional coordinates provided in the dataset, we calculated the distance between all voxel pairs and created a correlation matrix such that voxels closer to one another are more highly correlated:

Spatial correlation:

$\rho_{ij} = (constant)^{d_{ij}^2}$ where d_{ij} is the distance between voxel i and j

For all experiments, the value of the constant used to calculate the spatial correlation was 0.5. This spatially correlated structure makes sense given that voxels next to each other are more likely to have similar activation patterns than those further away. While there is still the possibility that voxels far away from each other could be highly correlated since brain regions are highly interconnected, the spatial correlation still captures a major element of the fMRI data and the performance of our modified BH method on this simulation can give us more information on how it will perform on the real dataset. The results can also be found on Figures 1, 2(a), and 2(b) in green and the details on FDR and power for each factor value are listed in Table 3. We can see from the graph that the FDR from spatially correlated simulations follow the real data covariance simulation pretty closely. The power of the method on the spatially correlated data was relatively low when c was less than 2. The best combination of FDR and power occurs at around c = 1.8 (Table 3).

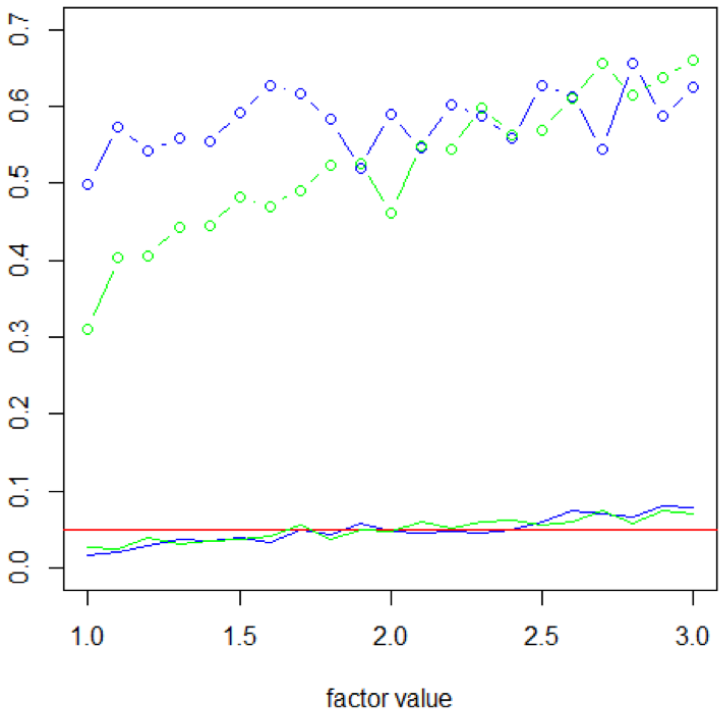

Figure 1: Power (open circle points and dashed line) and FDR (solid line) from MRI data covariance (blue) and spatial correlation (green) simulated data verse value of c from data determined modified B-H procedure. Red horizontal line indicates the value of **α** we set for our experiments.

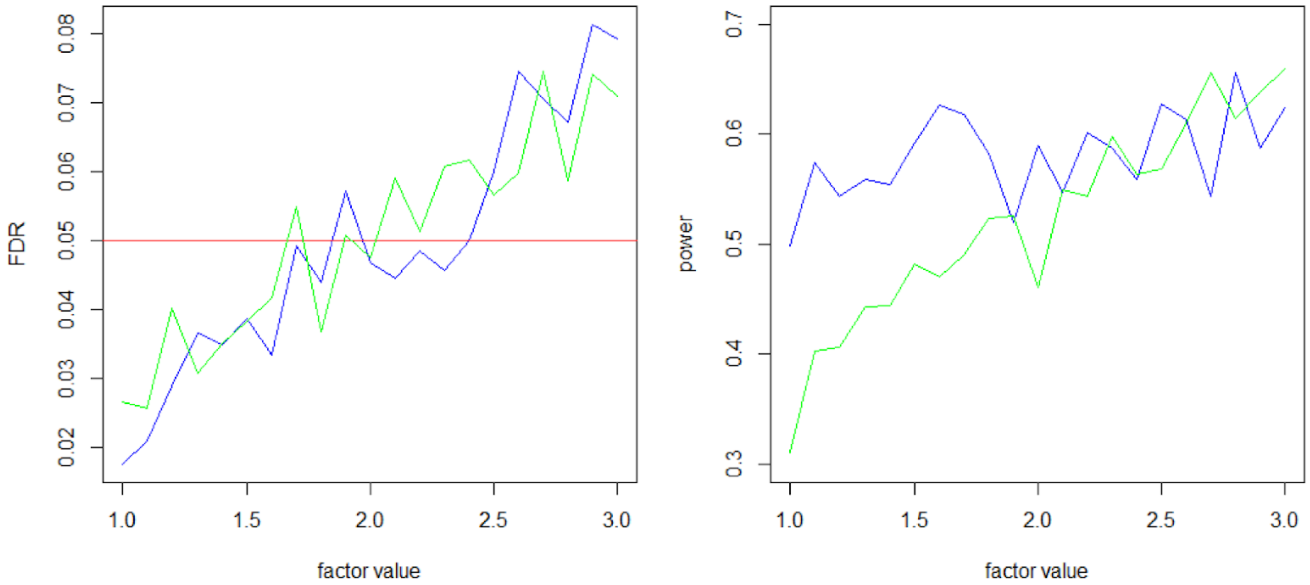

Figure 2(a) & (b): FDR and power from MRI data covariance (blue) and spatial correlation (green) simulated data verse value of c from data determined modified B-H procedure. Red horizontal line in **2(a)** indicates the value of **α** we set for our experiments.

Table 3: Resulting FDR and power from varying values of c for modified Benjamini-Hochberg procedure on data simulated from the overall fMRI data covariance and spatial correlation.

	Overall fMRI data covariance		***Spatial correlation***	
Value of c	***FDR***	***Power***	***FDR***	***Power***
1.0	0.018	0.499	0.027	0.311
1.1	0.021	0.574	0.026	0.403
1.2	0.029	0.543	0.040	0.407
1.3	0.037	0.559	0.031	0.444
1.4	0.035	0.554	0.035	0.444
1.5	0.039	0.592	0.038	0.482
1.6	0.034	0.626	0.042	0.470
1.7	0.049	0.618	0.055	0.490
1.8	0.044	0.583	0.037	0.523
1.9	0.057	0.520	0.051	0.526
2.0	0.047	0.590	0.047	0.461
2.1	0.045	0.547	0.059	0.549
2.2	0.048	0.602	0.051	0.544
2.3	0.046	0.588	0.061	0.599
2.4	0.050	0.559	0.062	0.564
2.5	0.060	0.627	0.056	0.569
2.6	0.075	0.613	0.060	0.610
2.7	0.071	0.544	0.074	0.656
2.8	0.067	0.656	0.059	0.615
2.9	0.081	0.587	0.074	0.638
3.0	0.079	0.625	0.071	0.659

Reduced dimension by method of clustering

The final method we explored also involved using the voxel coordinates. Since we assume nearby voxels activate similarly, we designed a method to reduce the dimensions of our dataset by clustering small groups of voxels. We assume that clusters of voxels are mainly influenced by a single more significant voxel, and that the clusters are independent from one another. Specifically, voxels within 3 voxels away from the center voxel were considered as a cluster and we then took the smallest voxel p-value within that cluster to be the representative p-value of the cluster. This allows us to reduce the dimensions of the data from 4698 individual voxels to 1824 clusters. Assuming that each cluster is independent or only weakly correlated, we can then modify the original BH procedure to $\frac{\alpha k}{n}$, where n is now 1824, the number of clusters.

Using the simulation based on the real data's covariance, we tested how our modified BH procedure works on our reduced-dimension dataset. The performance was quite solid when the proportion of nulls were 0.5, with an FDR of 0.04 and power of 0.72. While the FDR and power seemed somewhat satisfactory, we have to be cautious as this method has discarded a large amount of data during the process of dimension reduction and hence could miss important discoveries in the real dataset if the experiment aim to identity significant brain regions.

On the other hand, because the FDR and power were comparable to the ones we obtained from our factor modified BH procedure, the clustering method could offer some intriguing validating comparisons to our other method of choice when applying to the real dataset.

Real data analysis

After exploring different methods of applying the Benjamini-Hochberg procedure, we decided to use the BH procedure with a modifier of c = 1.6 and the cluster dimension-reducing method on the real fMRI data as they performed the best on our various simulations.

We decided to analyze the difference in brain reactions caused by the order of the stimuli presentation. We used a paired t-test to compare the voxels' activities between "picture first" presentation and "sentence first" presentation for the first trial, resulting in 4698 p-values. We then applied the modified BH procedure and made 65 discoveries. The cluster dimension-reducing method gave us 73 discoveries.

Conclusion and Limitations

The results of our real data analysis gave us a very low number of discoveries relative to the total number of voxels, which could signal that our methods were too conservative. However, the result could still be valid considering the condition of "picture first" and "sentence first" are quite subtle and there simply might not be as many brain regions that react differently according to the order of stimuli.

During our exploration of different adjustment for the multiple testing problems, we have encountered major computing challenges. Due the the size of the covariance matrices (4698x4698), each simulation and iteration takes a significant amount of time. Therefore, unless specified, the results we obtained on investigating the performance of these methods were generated by only 1 iteration. This could obviously lead to some outlier values in our report. However, we could offer some comfort in stating that throughout the process of creating this report, we have rerun the code a few times for testing or due to system crashes and we have reason to believe that the numbers we have computed are not far from the norm.

Similarly, for comparing different values of the c modifier, we used an arbitrary 0.5 as the proportion of nulls. For deeper understanding of the performance for these methods, we would have liked to repeat the methods using various proportions of nulls. However, running one loop testing 21 values of c took about 5 hours, hindering us from carrying out further examinations.

In conclusion, we have presented a few methods of accounting for the multiple testing problem in an fMRI data set, and showed how they perform on various simulated data structures that each represent different characteristics of the highly correlated fMRI data. It is clear that the group-adaptive Benjamini-Hochberg procedure has shortcomings and is not ideal in a highly correlated setting. The Benjamini-Hochberg procedure with an added factor modification could be a better alternative and further investigation of its performance would be valuable.