Result Replication Report

Leyang Zhang, Wentao Chen, Zimin Dai
A14822724, A14964886, A14509602

1. INTRODUCTION

In the modern world, data has been highly valued by the society unprecedently. With an increasingly thorough data collecting and managing procedure, researchers are presented with large scales of data, which they are not used to. The term big data born under this trend, using to describe a field dealing with large or complex data sets. The appearance of big data leads to many high dimensional or large-scale data analyzing problems involve searching for signals among the large set of candidates. The range of such large-scale signal detection problems can be extremely broad, from medical field like detecting genes responsible for an increased risk of getting diabetes, to botany like detecting organic plants.

When dealing with such large-scale multiple testing problem, generating correct inference in terms of any of the strategies for thresholding is highly dependent on the test statistics of null distribution. The paper we are trying to replicate in neuroimaging (A. Schwartzman et al., 2009), shows example that demonstrate how the theoretical null distribution doesn’t fit well with the observed the distribution. Consequently, the empirical null distribution is thought to be the solution to this problem.

During this quarter, we go over some of the main methods of the paper and conduct several experimental simulations on these methods.

2. DATASET

Data sets suitable for conducting such multiple testing using empirical null distribution include observations from two groups in which we can detect a difference between them, and single observations of test statistics where we can use to detect significant signals directly. In this quarter, we applied the method to two data sets:

The first data set we use is the Jongho Bold dataset, from the research paper (A. Schwartzman et al., 2009), containing a single list of 15631 z-scores. These z-scores are collected from fMRI scans of brains when receiving a stimulus, after applying certain imaging processing techniques like spatial registration and normalization to every voxel in the search region. Z-score close to 0 means that part of the brain was not correlated to that stimulus, while large z-score, either positive or negative, shows strong correlation between that part of the brain and the stimulus.

Another dataset we deal with is called ‘Mootha’, (Mootha et al., 2003) being used as practice for us, contains the gene expressions collected from diabetes patients and control healthy group for comparison. In total of 10983 genes’ expression levels were presented, and for each gene, 34 samples were collected, half from patients with Type 2 diabetes mellitus, and the other half from the group with normal glucose tolerance.

3. METHODOLOGY

So far in our investigation, we have developed four main methods. These methods mostly focus on calculating the FDR of input dataset or random data generated from standard normal distribution, and also fitting data to the empirical null distribution.

3.1 False Discovery Rate and Threshold

The first method we have is used to explore the relationship between False Discovery Rate and threshold. In this case, the data we now have is the Mootha data that contains genetic expression levels measured from diabetes patients and control groups. By dividing the data into control and experimental group, we perform a two-sample t test on them to check any significant data point. The first decision we have to make is choosing the thresholds for the determination of significance. We pick a set of thresholds based on the p-value and t-value distributions. For example, we choose 0 as the minimum and 1 as the maximum with a step of 0.1 for p values, and -5 as the minimum and 5 as the maximum with a step of 0.1 for t statistics. By this threshold, we are able to cover most of the p-values and t-statistics generated from the two samples t test for the Mootha dataset. With these thresholds, we are able to calculate the expected counts and obtained count of t statistic and p values that are greater than the threshold. As the final step, which is the main purpose of this method, we plot the FDR rate as a function of threshold that provides a reference for how the FDR changes with different thresholds. Figure 3.1.1 shows FDR rate as a function of threshold for p values, and Figure 3.1.2 shows FDR rate as a function of threshold for t statistics.

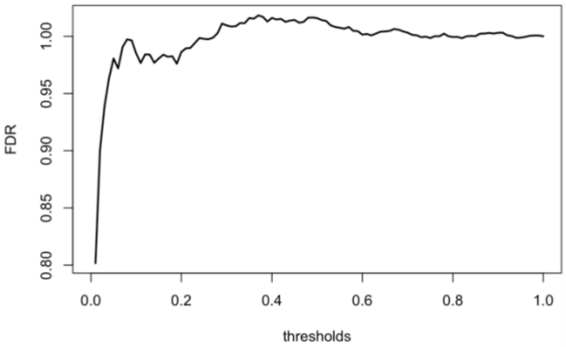

Figure 3.1.1

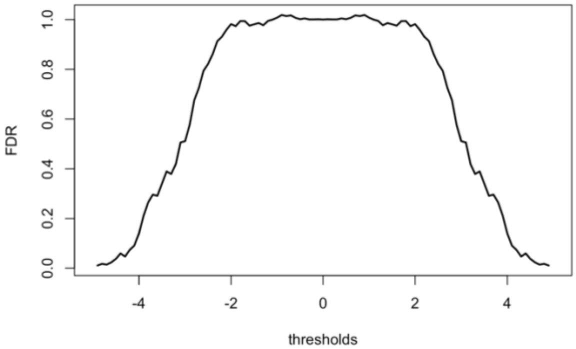

Figure 3.1.2

3.2 Theoretical vs Empirical Null Distribution on Matching Data

The second method we have is used to fit a dataset of z-scores with both theoretical and empirical null distribution, where we use the Jongho dataset. The first step in this method is to overlap the theoretic distribution, calculated by the midpoints of each bin with the dnorm() function, on top of the density of z-scores. This allows us to see how well the theoretical distribution fits the actual z-score distributions, as shown in Figure 3.2.1.

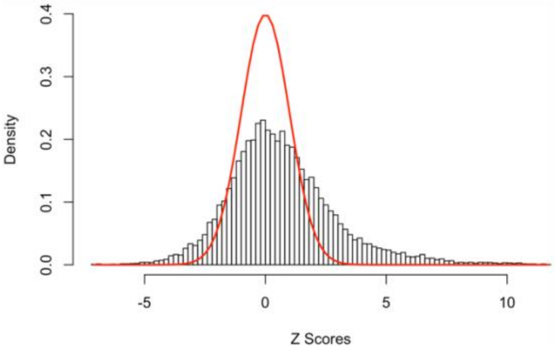

Figure 3.2.1

Then we make plot using formula lm(y~x + I(x^2)), where x is the midpoints of the histograms and y is the log density of each bin, to make prediction of the density under the empirical null. As a result, we plot the empirical null distribution on top of the transformed density to compare how it fits better than the theoretical distribution and then saves its coefficients. (Figure 3.2.2)

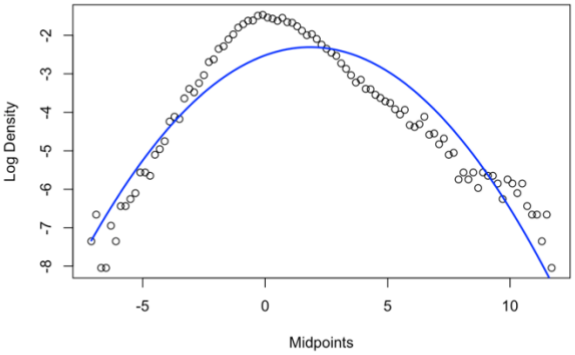

Figure 3.2.2

3.3 Exploration on Error Rates

The third method we have is called ‘Error Rate’ that is used to study the correlation between the TPR, FPR, and FDR of two distributions that contain randomly generated data from normal distribution. The first distribution is the null distribution, and the second distribution is the true distribution. We can also import data as the second distribution for experiment such as the Jongho or Mootha dataset. The threshold range we choose begins from the smaller mean minus 2 to the bigger mean plus 2 with steps of 0.1. Then we set anything greater than the threshold to be positive observation and negative observation otherwise to simulate model predictions. Eventually, we calculate the True Positive Rate (TPR), False Positive Rate (FPR), and False Discovery Rate (FDR) of each threshold chosen with formulas of TPR = TP/length of true distribution, FPR = FP/length of null distribution, FDR = FP/(FP+TP), where FP = sum(null distribution > threshold) and TP = sum(true distribution > threshold). Then we plot the TPR, FPR, and FDR together as a function of the thresholds (Figure 3.3.1),

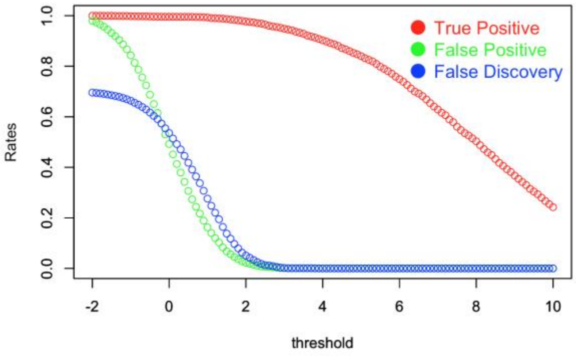

Figure 3.3.1

and TPR as functions of FPR and FDR separately (Figure 3.3.2 and Figure 3.3.3).

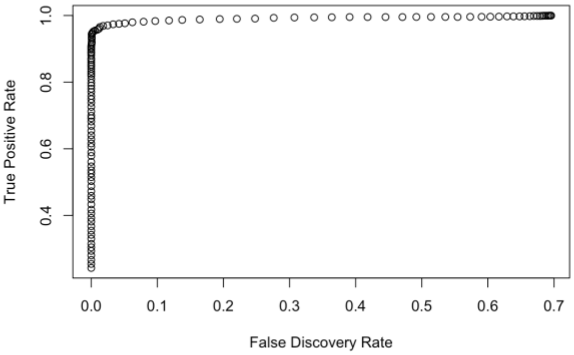

Figure 3.3.2

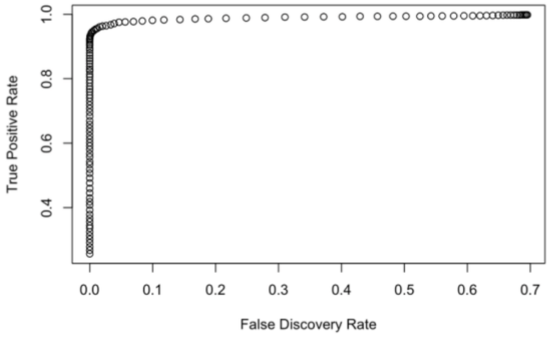

Figure 3.3.3

These three plots allow us to check the relationship between the three typical rates, which helps us with determining the threshold for the distribution by looking at a balance point.

3.4 Estimation of P0

Last but not least, we finally get closer and closer to develop a method that can separate mixed data from two different distributions — a two-class mixture model. By separating mixed data, we will be able find the significant data points that hide in the shadow. In order to do so, we first have to estimate p0 which is the fraction of null distribution (insignificant data) in the dataset. Thus, we created a method called “estimate_p0”. To simulate mixed data, we generate random data from two classes. The first class is the null distribution which is standard normal distribution, and the second class is the true distribution with different mean and standard deviation to the first class. The visualization of the mixed data is shown in Figure 3.4.1.

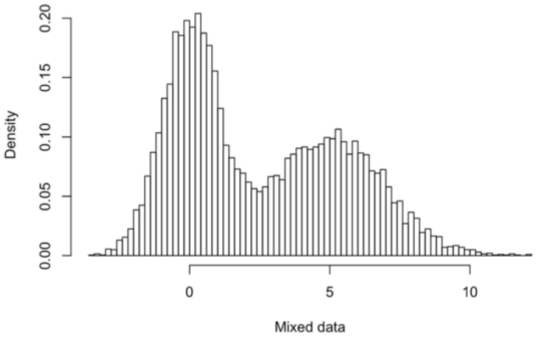

Figure 3.4.1

We choose upper limit of 1 and lower limit of -1 to be the range because one standard deviation away from the mean covers sixty-eight percent of the data which is sufficient at this point. Then we plot a histogram of the mixed data and find the index of midpoints of each cell that are within the upper and lower limit; these indices are the cells that we perform our estimation on. At the end, we calculate the ratio of density to the pdf of midpoints at these indices, and we take the mean of calculated ratios to be the final estimation of p0. The p0 fraction corresponds to the proportion of data under the green curve in Figure 2, which is about 0.5.

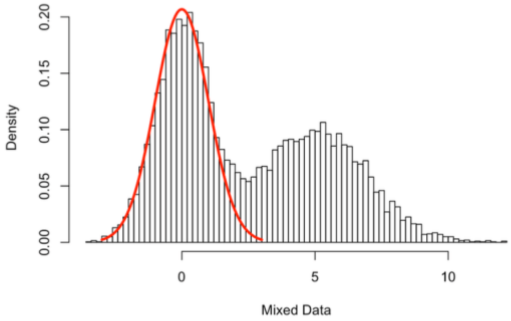

Figure 3.4.2

For now, we only take randomly generated data as input; but after some improvement later on, we should be able to input actual dataset into the function.

4. RESULT

In the first method, we explore the relationship between false discovery rate and chosen thresholds for the p-values and test statistics. From the Figure 3.1.1, we can see that false discovery rate will increase dramatically between a threshold of 0.01 and 0.1. The false discovery rate will be nearly 1 when threshold increases to 0.1, and the false discovery rate is still as high as 0.8 at a threshold of 0.01, which means that a more precise threshold is needed if we want to control the false discovery rate to a lower level. In the plot about false discovery rate and test statistics (Figure 3.1.2), it is obvious that any narrow ranges, less then -2 to 2 in this case, will result a false discovery rate of almost 1, while wider ranges, more than -4 to 4 in this case, can control the false discovery rate to a low level (less than 0.1). The result we find through the experiments matches our intuition on thresholds, that significant signals being found under strict criteria are more likely to be the correct ones.

For the second method about the matching of observed test statistics and null distributions, we overlap the test statistics with both theoretical and empirical null distribution. From figure 3.2.1, we can see that theoretical null distribution does not match the observed test statistics, that it’s much higher and thinner. While, in figure 3.2.2, we overlap the empirical null distribution on the overserved test statistics after transforming. It’s quite obvious that the blue line is a good match to the dots, which means empirical null distribution matches well to the observed test statistics.

In term of the Error Rate method, we’ve discovered several interesting results. For example, in the plot that contains TPR, FPR, and FDR as a function of threshold, the highest point of FDR rate is always the same as the p0 which is the fraction of null distribution; this is because the smallest threshold always treats every data point as positive. For practical application, to choose the best threshold that balance TPR and FPR, we should choose the point at the top left corner as the optimal point. In addition, if we input more separated means and standard deviations as parameters, the slope of the curve of each rate becomes steeper, since it’s much easier to distinguish more separated distributions.

Lastly, there are some surprising results of the Estimate_p0 method using various parameters. When the null and true distribution are well or somewhat divided, the estimation can be exactly right. However, as the two distributions become more clustered, the model estimation sometimes has margin of error of 0.2 or greater, which is quite large. Because of that, we will need some precise calibration of the upper and lower limit in the future. But for now, this method provides an important foundation for finding the empirical null distribution.

5. CONCLUSION

In this final report, we have demonstrated several methods to calculate error rates and estimate the null density from mixed data. Learned from these methods, we gain better understanding of how to choose more reasonable and accurate threshold based on the context of dataset. Even though we haven't got to the point of using the empirical null approach to discover significance, our methods are able to estimate the proportion of data that belongs to the null distribution for most of the time. As mentioned in the paper provided by the professor (A. Schwartzman et al., 2009), estimation of the null density is the most important aspect of using an empirical approach to FDR inference. Thus, with what we have developed so far, we are in somewhat good shape of getting into next quarter's project, when we go beyond medical field and try to apply these methods on a different data set in a different field.

6. REFERENCE

Schwartzman, A., Dougherty, R., Lee, J., Ghahremani, D., Taylor, J.,2009. Empirical null and false discovery rate analysis in neuroimaging. NeuroImage 44 (2009) 71−82

MOOTHA,V.K.,LINDGREN,C.M.,ERIKSSON,F.K.,SUBRAMANIAN,A.,SIHAG,S.,LEHAR,J.,PUIGSERVER,P.,CARLSSON,E.,RIDDERSTRAALE,M.,LAURILA,E.,ET AL.(2003). PGC-1-responsive genes involved in oxidativephosphorylation are coordinately downregulated in human diabetes.Nature Genetics34, 267–73.