Multiple Testing Method with Empirical Null Distribution in Leukemia Studies

Jacob Benson and Raymond Wang

Abstract

In genomics we are often faced with the task to identify genes correlated with a specific disease among a large number of candidate gene pools. A naive approach is to apply a hypothesis test to every individual gene. This method ignores confounding factors in the data and does not adjust for the additional variance. In this paper we will introduce a much more robust method primarily using estimations of the empirical null distribution and the false discovery rate (FDR). A leukemia dataset is used to demonstrate that the empirical null distribution, one estimated from observing the data first, provides a better fit of the theoretical null distribution. Furthermore, we will compare and contrast the result with unsupervised classification methods such as k-Means and the Gaussian Mixture Model.

Introduction

Multiple testing analysis is a commonly used statistical practice in genetic research, often used to separate significant genes from a large pool of candidates. For example, in genome-wide association studies (GWAS), the multiple testing approach can be used to identify particular genes responsible for a certain disease. This is achieved through simultaneously conducting a large number of hypothesis tests between the control and experimental groups on each individual gene. With each hypothesis test's t-statistics recorded, we can search through the data for anomalies and identify genes that might be correlated with the disease of our interest.

In this paper, we will apply this method with a dataset on leukemia studies. This dataset, collected by Harvard Professor of Pediatrics, Todd Golub, consists of the gene expressions of more than six thousand different genes from a mixture sample of seventy-two patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). Our task is to identify genes that differentiate the two types of leukemia. The original dataset was separated into a training set and a testing set for model building purposes. We will use both an empirical null distribution as well as a theoretical null distribution; however, we hypothesize that the empirical

null distribution would provide more accurate results than the theoretical null distribution. This is because we believe that our data can be skewed so that a theoretical null distribution will not represent our data well; instead an empirical null distribution, one that is obtained from first observing the data and estimating the fraction of each component of the mixture model's distribution, can better represent our data and give more accurate results. Furthermore, after computing the t-statistics, we will threshold by controlling the false discovery rate (FDR).

To test the effects of the multiple testing analysis, we will also apply unsupervised clustering methods such as k-means clustering and the Gaussian mixture model to the same dataset. We will compare and contrast the results from the clustering models to those obtained from multiple testing analysis. We hypothesize that naive classification models fall short when using large scale data.

Theory and Methods

Quantile Transformation

After some exploratory data analysis on the leukemia dataset, we discovered that the two groups, patients with AML and patients with ALL, have different variances across the dataset. This appears to be an issue for our hypothesis testing procedure, as the t-test function in R applies Welch's t-test using approximation to the degrees of freedom in case of unequal variances between two groups. Since the degrees of freedom varies among samples, we must apply a quantile transformation on the t-statistics in order to obtain uniform distribution across our data.

To do this we found the areas (or probabilities) of each t statistic and their respective z scores in a normal distribution. We used the R function *pt()* to get the areas. The *pt()* function takes in a t-statistic and the number of degrees of freedom and returns the area below that t-value to the upper-tail. We wanted to find the lower-tail areas so we used *1 - pt()* and input our set of t-scores and respective degrees of freedom for each score. Then we used the *qnorm* function to get the corresponding z scores for each area. The *qnorm* function takes the lower-tailed area and returns the corresponding z-value, hence why we wanted the lower-tailed areas from *pt*. After we input our areas we are now left with a normal distribution of z-scores that will allow us to use the empirical null distribution.

Empirical Null Distribution

We will estimate the empirical null distribution by estimating the fraction of true negatives. To do so, we separate the data into two sets, S_0 and S_A, where S_0 denotes all true negatives and S_A denotes all true positives. In our dataset, S_0 represents all genes that are uncorrelated with differentiating AML from ALL while S_A represents those genes that correlate. Furthermore, we define $f_0(t)$ and $f_A(t)$ to be the probability distribution of S_0 and S_A respectively. We will estimate the fraction p_0 of the number of true negatives over the number of samples so that the data distribution:

$$f(t) = p_0 f_0(t) + (1 - p_0) f_A(t) \quad (1)$$

is close to the the scaled density, $p_0 f_0(t)$ within an interval l_0 around the null distribution mean, μ_0.

Notably, Efron et al. (2001) discusses that such application of the empirical null distribution must reach the following two requirements:

1. The number of tests, N, must be large.
2. A large majority of the tests must be within the null set, S_0.

False Discovery Rate

We will use false discovery rate as our threshold metric for the statistical parametric map. The false discovery rate is defined as the rate of type I errors, the ratio between the number of false positives and the number of predicted positives. For our study, we set the level of threshold as *u* and define *FP(u)* and *TP(u)* to be the number of false positives and the number of true positives respectively under the threshold *u*. We will compute the false discovery rate under threshold *u* as the following:

$$FDR(u) = E\left[\frac{FP(u)}{FP(u) + TP(u)}\right]. \quad (2)$$

We aim to control the threshold *u* to lower the false discovery rate. Notably, when *u* is large, we have fewer number of false positives, leading to a lowered FDR; however, increasing the threshold *u* also leads to increased false negatives. Ideally, we want to achieve a set FDR with the minimum level of threshold, *u.* To do this, we will first set a significance level α for the FDR and then calculate the FDR for different levels of *u* and select the minimum *u* that satisfies the significance level of FDR. To put this in formula, we want to calculate

$$min_u FDR(u) \leq \alpha. \quad (3)$$

Compared to other thresholding criteria such as the family-wise error rate (FWER), which controls the rate of false positives among all samples, the FDR only controls the rate of false positives among the positives and therefore is more permissive of false positives by definition (Schwartzman et al., 2009). Since our goal is to identify the genes correlated with differentiating the AML from ALL, FDR is preferable as it allows some false positives as long as there are way more true positives.

Data examples

To illustrate the effect of different null distributions on multiple testing methods, we will use leukemia collected by Harard Professor of Pediatrics, Todd Golub. The dataset consists of the gene expressions of seventy-two leukemia patients among which forty-seven of them have ALL and twenty-five of them have AML. To compare the differences between the two groups of patients, we first conduct a t-test between the ALL patients and the AML patients.

Fig.1 shows the histogram of 2,185 t-statistics and the red line indicates the theoretical null distribution N(0,1). The figure shows that the t-statistics are much wider and shorter than the theoretical null distribution. We also observed that the t-statistics have a very slight right skew. To obtain a better fit, we converted the t-statistics into z-scores and performed quantile transformation.

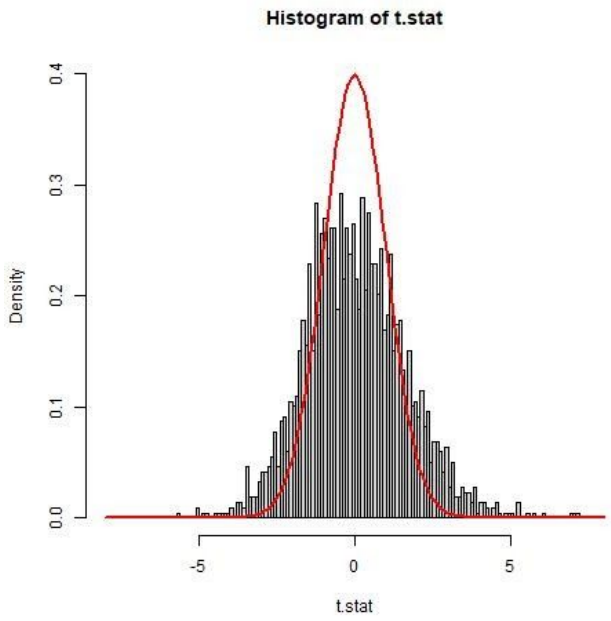

Fig.1. Histogram of the N=2,185 t-statistics (gray) and the theoretical null distribution (red line).

In Fig.2, we observe that standardization, the histogram of the z-scores are still wider and much shorter than the theoretical null distribution N(0,1). To obtain a better fitted null distribution, we use the $\widehat{p_0}$ estimation method mentioned in the section above using an interval of [-0.2, 0.2]. We compute the estimated parameter $\widehat{p_0} = 0.595$. This confirms our previous observation that our histogram is much lower than the theoretical null and estimates that around 40.5% of the genes are expressed differently in the two types of leukemia patients.

After scaling the theoretical null distribution N(0,1) with $\widehat{p_0}$, we obtain a distribution that better fits the z-scores. However, this approach still does not address the issue of the higher variance and the skewness of the data. Therefore, we estimate an empirical null distribution using the median of the z-scores and an estimated standard-deviation. Using the interquartile range (IQR), we estimate the standard deviation $\widehat{\sigma} = IQR/1.349$. Again we estimate the parameter $\widehat{p_0} = 0.933$ using the method above. This new estimation suggests that about 6.7% of the genes are expressed differently between the two groups, much lower than the 40.5% obtained above. Combine all the result, we have empirical null distribution estimated parameters $\widehat{p_0} = 0.933$, $\widehat{\mu} = -0.066$, and $\widehat{\sigma} = 1.534$. The resulting distribution, observed as the red solid line in Fig.2, provides a much better fit of the data than both the theoretical and scaled theoretical null distribution. The empirical null distribution adjusts to the histogram's additional variance due to confounding factors and will provide more realistic results and estimations of error rates.

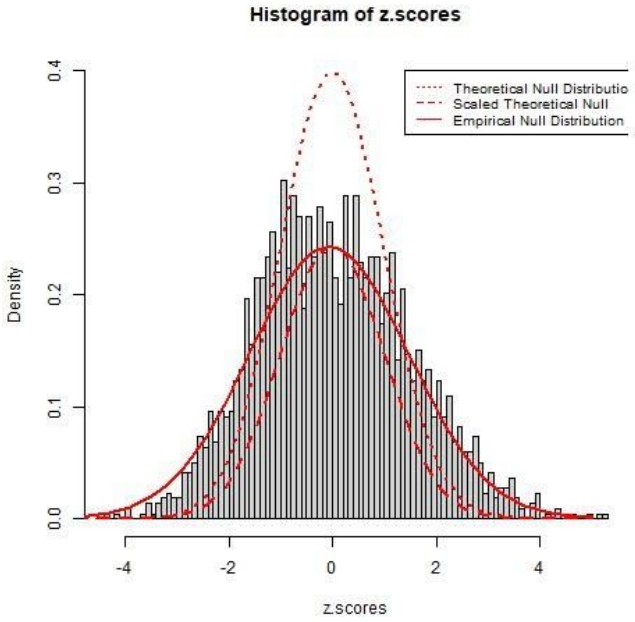

Fig.2. Histogram of the N=2,185 transformed z-scores (gray), the theoretical null distribution (red dotted line), the theoretical null distribution scaled by $\widehat{p_0}$ = 0.595 (red dashed line), and the empirical null distribution scaled by $\widehat{p_0}$ = 0.933.

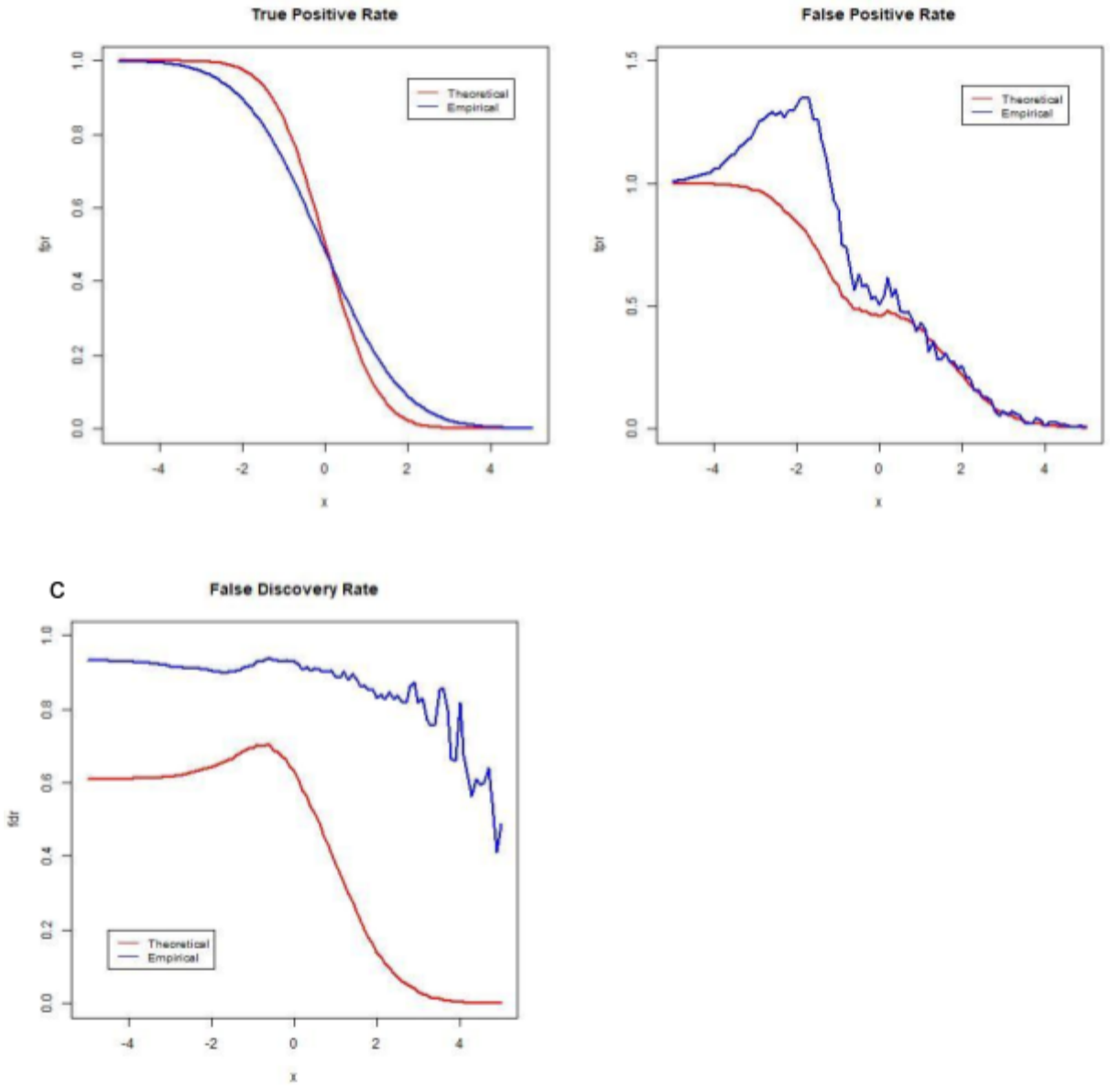

Fig.3. Error rate metrics of theoretical and empirical null distribution: (a) Right tail True Positive Rate (TPR). (b) Right tail False Positive Rate (FPR). (c) Right tail False Discovery Rate (FDR).

Fig.3 shows the right tail True Positive Rate (TPR), False Positive Rate (FPR), and False Discovery Rate (FDR) of the theoretical null distribution and the empirical null distribution. The theoretical null distribution is obtained from scaling the N(0,1) with estimated $\widehat{p_0} = 0.595$. The empirical null distribution is obtained from scaling N($\widehat{\mu}$, $\widehat{\sigma}$) whereas $\widehat{\mu} = -0.066$ and $\widehat{\sigma} = 1.534$ with $\widehat{p_0} = 0.933$. The TPR is calculated as the rate between the number of true

positives (TP) and the number of true positives plus the number of false negatives (FN). It indicates the likelihood that an actual positive sample produces a positive testing result. The FPR is calculated as the rate between the number of false positives (FP) and the number of false positives and true negatives (TN). It indicates the likelihood of false positives. Lastly, the FDR is computed in equation (2) as the rate between the number of FP and the number of total tested positives and it computes the rate of type I error. Fig.3c shows that the theoretical null distribution yields a FDR curve that converges to 0 as the threshold x increases. Meanwhile, the empirical null distribution produces a much higher level FDR than the theoretical null distribution given the same level of threshold. Moreover, the FDR curve yielded by the empirical null distribution fails to converge as x increases and is approximately 0.44 when the threshold is the largest at x = 5.

Discussion

At first, the result from Fig.3c seems to conflict with our hypothesis that the empirical null distribution will produce better results than the theoretical null distribution. While our result shows that the theoretical null distribution yields a lower FDR at the same level of threshold than the empirical null distribution, this does not necessarily mean that the theoretical null provides a more accurate result. Since we cannot know the actual label of the data, there is no way for us to actually test the accuracy of our result. What we have is an estimation of the error rate, which allows us to gain insight into the likelihood of false positives at each level of the threshold. The theoretical null distribution yields an apparent lower error rate, but this result may be incorrect. This is because that as we have observed, additional variance in the histogram, due to confounding factors, will skew the results of the theoretical null. In contrast, the empirical null distribution adjusts for those unknown confounding factors empirically. This results in lower detection power, as observed from the difference in the estimated parameter $\widehat{p_0}$, but the error rate may be more realistic, meaning it is closer to the true error rate.

References

Efron, B., Tibshirani, R., Storey, J.D., Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96 (456), 1151-1160.

Schwartzman,A., Dougherty, R., Lee, J., Ghahremani, D., & Taylor, J. (2009). Empirical null and false discovery rate analysis in neuroimaging. *NeuroImage*, *44*(1), 71–82.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-7. doi: 10.1126/science.286.5439.531. PMID: 10521349.