Introduction

In many areas of study we encounter the problem of searching for signals that are correlated with our predictor among a large number of data entries. In gene expression studies, we search for certain genes correlated with a particular disease among many candidates; in voxel-based neuroimaging analysis, we search for certain brain regions that are responsive to a particular stimulation. Typically, to tackle those types of problems, we first choose a suitable test statistic, calculate it for every candidate (e.g. genes, brain voxels, etc.) based on a theoretical null distribution, and then select a threshold for that test statistic.

An issue with the common approach in using a theoretical null distribution is that the observed test statistic distributions are often skewed and do not match well with the theoretical null. In this paper, we will present two examples where we will be using multiple testing to search for correlated signals among a large pool of candidates. With the data collected by Mootha et al. (gene expression), we will show the issues with large-scale multiple testing on high dimensional data. One of the main problems with this approach using theoretical null is that as we try to quantify the effects of metrics such as the false discovery rate, we establish arbitrary correlation due to the high variance of the fdr. With the z-scores of fMRI data we will show how the use of an empirical null distribution, as opposed to a theoretical null, allows us to better control the false discovery rate of active/inactive voxels.

The first data example, collected by Mootha et al., consists of a gene expression dataset from a random sample of seventeen diabetes patients and seventeen healthy controls. It contains the gene expression of 10983 different genes among 34 different individuals. 17 of these individuals had type II diabetes and the rest were healthy control subjects. A genomic dataset such as this is one of the more commonly used types of data in large-scale multiple testing because of the large number of features (genes measured). We will calculate t-statistics using the comparison between the diabetes group and the control group. The dataset contains high-dimensional data in which through Multiple Testing and controlled thresholding, we can find which genes show high correlations with the patients having type II diabetes.

The second example with the fMRI data will contain a dataset with the z-scores of fMRI BOLD data that consists of the brain activity of a single subject when given standard stimulations. It contains 15611 observations of z-scores, with both positive and negative values. Previous study on the data set has shown that the histogram of the observed z-scores is substantially wider compared to the density of a theoretical null distribution (Schwartzman et al., 2008). Therefore, we will perform a similar Multiple Testing, instead using a computed empirical null distribution and controlling the false discovery rate, in order to determine which brain voxel shows correlation with the given stimulation.

Theory and Methods

The false discovery rate is found to be useful in many exploratory studies. Compared to other commonly used methods such as controlling the family-wise error rate (FWER), the FDR-controlling procedure is more tolerant of false positives. In our study, we aim to show that using an empirical null distribution, representative of the data, results in lower variances of false positives. This process allows some false positives, and therefore, we will be using the method of controlling the FDR rather than similar metrics such as the FWER, which is intolerant of false positives.

The false discovery rate is defined as the expected proportion of type I errors, where one falsely rejects the null hypothesis. Mathematically, this converts to the ratio between the number of false positives and the number of all predicted positives. In our example with the gene expressions dataset, we assume that there are N two-sided t-tests with each one yielding a t-statistic T_i. We also set the threshold to u, where the gene is significant if $T_i > u$. Furthermore, we use *FP(u)* to denote the number of false positives given a threshold u, and *TP(u)* to denote the number of true positives given u. Conclusively, we have:

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

We will control the false discovery rate by changing the value of u. Empirically, as we increase the threshold, we reduce the number of false positives; however, by doing so we also risk increasing the number of false negatives. Therefore, as we aim to minimize the FDR, we also want to keep the threshold low. In this case, we will adopt the following algorithm where given significance level α and a range of u, we will compute *FDR(u)* for each u and select the minimum u for which *FDR(u)* $\leq \alpha$. The formula is given as the following;

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

In exploring N two-sided t-tests of the gene data in type-II diabetes patients and controls, the p-values for each gene was uniformly distributed. Similarly, the histogram of the t-statistics was normally distributed with a slight right skew. When we plotted the expected number of p and t values greater than a threshold we found that they were extremely close to the expected values. We also created plots demonstrating the effects that the p and t values have on the false discovery rate. As the p-value increased the FDR increased dramatically, reaching 1 by about 0.2. With the t-statistics the closer they came to zero the higher the FDR got, where the FDR was 1 in between -2 and 2.

Reference

Mootha, Vamsi K, et al. "PGC-1α-Responsive Genes Involved in Oxidative Phosphorylation Are Coordinately Downregulated in Human Diabetes." *Nature Genetics*, vol. 34, no. 3, 2003, pp. 267–273., doi:10.1038/ng1180.

Schwartzman, A., Dougherty, R.F., Lee, J., Ghahremani, D., Taylor, J.E., 2008. Empirical null and false discovery rate analysis in neuroimaging. NeuroImage 44 (2009) 71–82.

Schwartzman, A., and X. Lin. "The Effect of Correlation in False Discovery Rate Estimation." *Biometrika*, vol. 98, no. 1, 2011, pp. 199–214., doi:10.1093/biomet/asq075.

Lee, J., Shahram, M., Schwartzman, A., Pauly, J.M., 2007. A complex data analysis in high-resolution SSFP fMRI. Magn. Reson. Med. 57, 905–917.