Introduction

In many areas of study we encounter the problem of searching for signals that are correlated with our independent variable 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 that large-scale multiple testing on high dimensional data leads to issues with arbitrary correlation attributing to higher variance of false discoveries and the discovery rate estimators by attempting to quantify the effects of correlation on these metrics. 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 example 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. In the gene data, a majority of the gene expressions are positive values. 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 our independent variable.

The second example 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.

Reference

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.