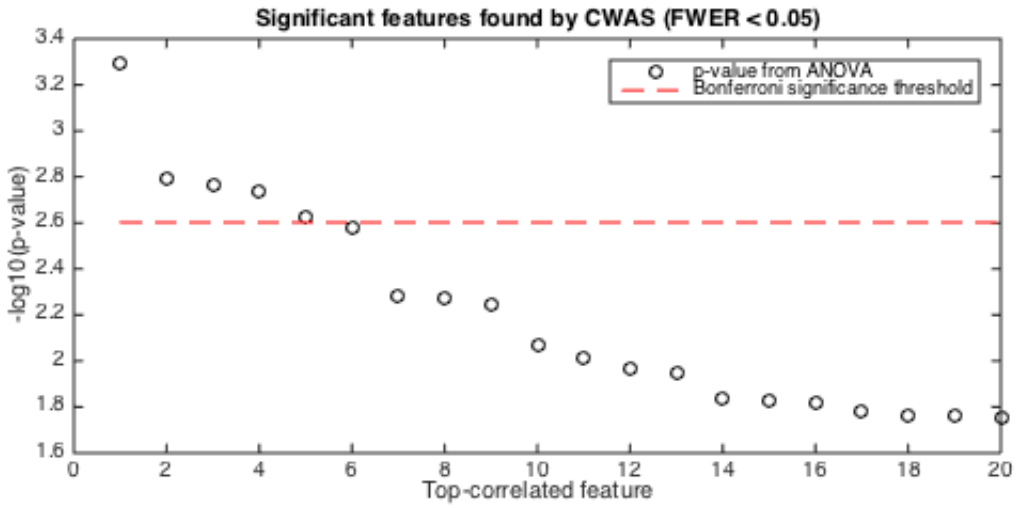
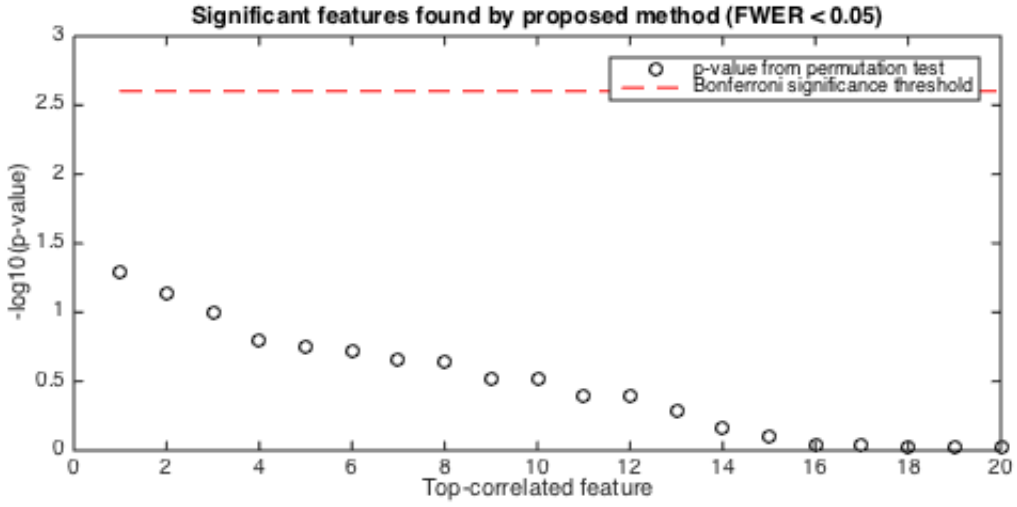
EN.580.694: Statistical Connectomics
Final Project Report

Jiarui Wang · May 7, 2015

A Statistical Sandbox for Validating Connectome-Wide Association Studies (CWAS).





Opportunity As CWAS grows as a field, there will be a need for a statistical framework for validating prognosis/diagnosis classification. As more data is generated at higher fidelity, there will be an increasing interest in trying to predict phenotypes based on connectomes. Many classifiers will be created, but there is currently no good way to validate whether or not these classifiers will perform accurately in a real-world setting. A statistical sandbox in which data are simulated with random disease labels independent of voxels can be useful for validating methodologies theoretically before they are put to real-world data. This crucial pre-validation step is often overlooked, but is critically important to ensure the validity of any proposed classification method for CWAS.

Challenge There is a physical limit to how much data at high enough resolution that we can collect. It is important to first establish that under the null case where voxels are independent of the disease labels that the classifier has a reasonably low false discovery rate. This kind of analysis was not done in a recent high-impact CWAS paper [1]. Current statistical methods rely on parametric tests and asymptotic assumptions, and non-parametric permutation testing have had limited application in the current literature. A main challenge is that there does not exist sufficient data for proper application of these statistical methods in terms of sample quantity and quality.

Action A data generating model was assumed where the connectomes and phenotypes are actually independent. A total of 100 patients were sampled from a Stochastic Block Model (SBM) with 1500 nodes, 20 clusters, and an inter-cluster probability of 0.3. Both data quantity and quality can be fully controlled by tweaking the SBM parameters. The first method tested with this generated data was the method used in the Shehzad paper in which the data is first filtered for the features that are most correlated with the disease labels then tested for significance by ANOVA. The p-values are then corrected for multiple comparisons by the Bonferroni correction, which was more conservative than the methods employed in the paper. The exact method could not be replicated since the code they provided failed to run. It should be noted however that in the process of re-creating their methods, extreme care was taken to ensure that the substitute techniques were either more or as conservative as the methods employed originally. An alternative proposed method is to use a permutation test with 10000 permutations with Bonferroni correction to evaluate p-values and to use a uniform random sample of all the features as a filter so as to reduce bias. The resulting analysis produced the figure above. Top 20 most correlated features were picked out of the 1500 possible. Higher values are more significant. Features above the Bonferroni threshold are classified as significant.

Resolution We established a sandbox statistical environment for which we may control all the hidden parameters of the system for testing classifiers previously built. We observed that even under the most conservative conditions, the methods used in the Shehzad paper had a false discovery rate of 3 to 5 percent, while the proposed alternative method reported a zero false discovery rate. This is alarming given that the substitutions made to replicate their results were more conservative, indicating that the actual false discovery rate of their method is even higher. The proposed non-parametric permutation testing using the absolute

difference of the means as the test statistic with random feature picks ensured that the false discovery rate was zero for data artificially generated to have disease labels independent of the features. This alternative method does not rely on asymptotic assumptions like ANOVA and does not bias the data processing for false positives.

Future Work This method can be applied to other existing classifiers from the literature in order to assess their correctness. The false discovery rate for these classifiers under the null hypothesis can then be calculated to assess their effectiveness. Hopefully, future classifiers can be developed with this sandbox framework in mind so that all new classifiers can pass this test.

References

[1] Zarrar Shehzad, Clare Kelly, Philip T Reiss, R Cameron Craddock, John W Emerson, Katie McMahon, David A Copland, Xavier Castellanos, and Michael P Milham. A multivariate distance-based analytic framework for connectome-wide association studies. *NeuroImage*, pages 74–94, February 2014.