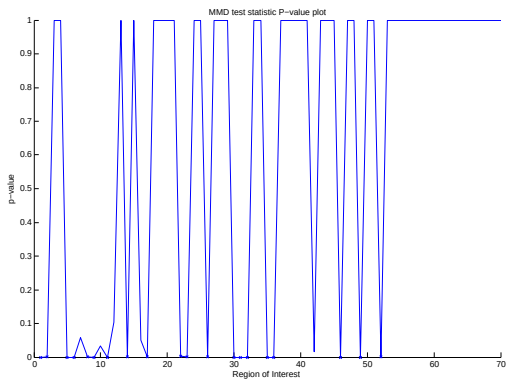
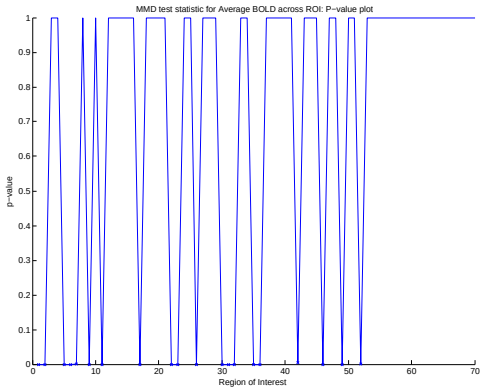
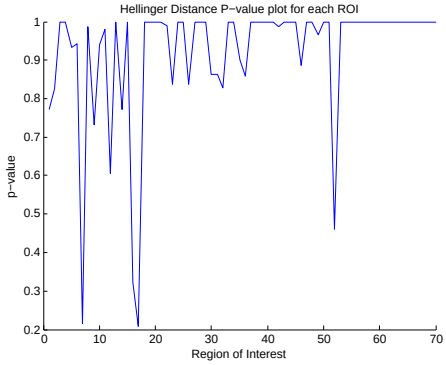
EN.580.694: Statistical Connectomics
Final Project Report

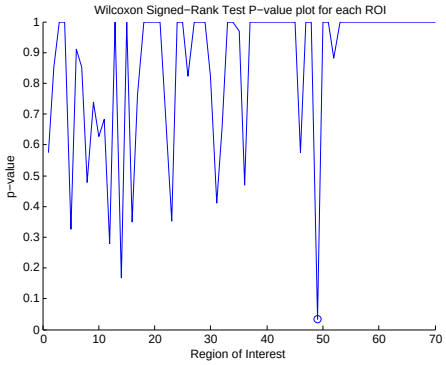
Heather G. Patsolic · May 14, 2015

Applying kernel based (and other) hypothesis testing techniques to a connectome-wide association study









Opportunity Many diseases seem to be correlated with functional and structural patterns in the brain. Using ideas from Shezhad et al., it would be nice to be able to use Blood Oxygen Level Dependent (BOLD) values in certain regions of the brain to determine whether or not a particular individual may have a particular phenotype, in this case ADHD. For example, can we say that beyond a certain threshold, a particular BOLD level may indicate that a person has ADHD? As was studied in Shehzad et al, Multivariate Distance Matrix Regression (MDMR) (see [3]) could be used; this requires a distance matrix, similar to using a kernel, but without actually using kernel-based hypothesis testing.

Challenge The authors of [3] used statistics based on a correlation matrix which forces the random variables to be dependent rather than independent. They then use the F-test which assumes independent identically distributed samples. Thus, the resulting p-values they find are related to one another, which makes their results confusing. There are also many calculations required, which increases computation time. We need to find a way to address some of these issues in determining which voxels or regions of interest may be associated with ADHD. The goal was to explore the use of kernels and kernel-based hypothesis testing and get rid of dependent random variables (such as correlation coefficients) in the analysis.

Action In order to reduce computation time, I averaged all the BOLD values over time for each voxel, and then summed all BOLD levels for each voxel in a region in order to create a vector for each participant of BOLD values over the various regions of interest (ROI) in the brain based on the Desikan scale (see [1]). I also created an average BOLD level across region vector for each participant (taking the sum of BOLD levels and dividing by the number of voxels in the region). I then performed three different analyses to compare the BOLD values for each region, each of which has pros and cons as described below. Correlation coefficients were not calculated for any of the analyses that follow. For each, the null hypothesis is that there is no correlation between BOLD values and presence/absence of ADHD, with an alternative hypothesis that these two are related. In other words, the null hypothesis is that the BOLD values for ADHD participants and the BOLD values for TDC participants come from the same distribution, while the alternative is that they come from different distributions.

1. Using the Gaussian kernel map to determine closeness of connectivity patterns for ROI across individuals, the Maximum Mean Discrepancy (MMD) test statistic is used, as described in [2]. For starters, I used the sum of the average BOLD levels for each voxel in a region and computed the test statistic for MMD using these values. However, these proved to be too large and resulted in values of all 0's and 1's. To correct this, I tried two things: (1) I got a normalized vector of BOLD levels across regions for each participant, (2) I recalculated the sum of BOLD values matrix to get an average BOLD level across each region for each participant. The first image above shows the results when I normalized the rows of the matrix and the image beside it shows the results for (2). A null distribution is created based on permutation testing, and then the p-value associated with the MMD test statistic calculated with the true labels is determined for each ROI.

2. Also performed is a series of Hellinger-Distances along with a kernel density map to test similarity. The Hellinger distance is another method used to determine if two samples come from the same probability distribution (see [4] for more details on using the Hellinger distance). The null distribution is created via permutation testing and a p-value found associated with the Hellinger distance between the two samples (ADHD and TDC) for each ROI.

3. Finally, I noticed that both the kernel-based test and the test using the Hellinger distance as a similarity measure require the use of independent samples; however, in the Shezhad et al paper the data is obtained via a matched sample where participants are paired based on age and sex. Therefore, I felt it was appropriate to use a Wilcoxon signed-rank test.

Resolution The methods employed take much less time than multivariate distance matrix regression, because fewer matrices need to be created. Using kernels allows us to do computations for analysis more easily, by allowing us the use of dot-products as a measure of distance along with norms that are easily computed. We also have the advantage of being able to use kernel-based methods for hypothesis testing as described by Harchaoui et al (2013) [2], such as the maximum mean discrepancy and the Gaussian kernel trick (which have the same result). Hypothesis testing using the Gaussian kernel required normalizing the vector of ROI for each participant in order to get anything conclusive or using an average BOLD value matrix. Once either of these was done, most ROI (that data existed for) were found to be potentially associated with ADHD with p-values < 0.01. Furthermore, while p-values using hypothesis testing based on the Hellinger distance found none of the ROI to be of interest, the three with the smallest p-values were 7, 17, and 52, the last two of which had a p-value $< .01$ when using the Gaussian kernel as a measure of similarity. Lastly, using the Wilcoxon signed-rank test (function courtesy of MATLAB), I found that no region had a p-value < 0.01 and one region, region 49, had a p-value < 0.05. Region 49 was also significant when using MMD.

Future For starters, there are some bugs. For example, the mapping into the ROI using the Desikan system (courtesy of Greg Kiar), had many regions which contained no voxels, thus resulting in what appeared in the matrices as regions where BOLD was 0. This may have occured for any number of reasons, for example: a mix-up in the matching of regions, the removal of voxels that did not align across participants, or a different system for labeling vertices being used than we thought. This causes many p-values to be 1 for the sheer fact that there are no recorded BOLD values for these regions. Also, when reconfiguring the matrix used for MMD (using the matrix of average BOLD value across each region), I had to ignore voxels in regions marked 0.

I think further tests would need to be done to establish consistency in determining which ROI truly are of interest with regard to determining whether someone may have ADHD. It's unclear whether the regions found using the MMD methods truly are significant or if there is some error. It would be necessary to compare these results to similar problems to determine whether these conclusions are the same. Before using these methods further in this area, we should go back and see why not all of the regions were significant across the

different methods. It would be interesting to see what is going on with the regions mentioned above (for example, looking at these regions at the voxel level). In the future, that is once we resolve bugs, we could test these methods with other data sets to determine associations between functional connectivity and other diseases or phenotypes. We could also determine associations between phenotypes and structural connectivity. I have been told that often we can apply similar methods from neural network data to social networks, so this could also be a future endeavor. We can also test whether kernels work better for connectome-wide studies or for more localized region of interest studies.

References

[1] R. S. Desikan, F. Ségonne, B. Fischl, B. T. Quinn, B. C. Dickerson, D. Blacker, R. L. Buckner, A. M. Dale, R. P. Maguire, B. T. Hyman, M. S. Albert, and R. J. Killiany. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31(3)** (2006 July), 968-980.

[2] Z. Harchaoui, F. Bach, O. Cappé and É Moulines. Kernel-Based Methods for Hypothesis Testing. *IEEE Signal Processing Magazine* (2013 July), 87-97.

[3] Z. Shehzad, C. Kelly, P. T. Reiss, R. C. Craddock, J. W. Emerson, K. McMahon, D. A. Copland, F. X. Castellanos, and M. P. Milham. An Multivariate Distance-Based Analytic Framework for Connectome-Wide Association Studies. *Neuroimage* **93 Pt 1** (2014 June), 74-94.

[4] R. Tamura and D. D. Boos. Minimum Hellinger Distance Estimation for Multivariate Location and Covariance. *Journal of the American Statistical Association* **Vol. 81, No. 393** (1986 March), 223-229.