Topological Data Analysis Reveals Meaningful Subgroups in ASD Research Data Based on Neural Responsivity and Behavioral Measures

T. McAllister, A. Naples, A. Chang, S. Hasselmo, M. Rolison, T. Day , T. Halligan, S. Malak, K. McNaughton, K. Stinson, J. Trapani, K. Ellison, E. Jarzabek, B. Lewis, J. Wolf, J. McPartland

McPartland Lab, Yale Child Study Center, New Haven, CT

Background:

- Modern neuroscience research increasingly collects vast quantities of rich, multivariable data
- Without more sophisticated tools, the full richness of the data will go unutilized
- Clinical, eye tracking, and electroencephalography (EEG) datasets from Autism Spectrum Disorder (ASD) research often contain hundreds of variables
- Topological Data Analysis (TDA) is a method of visualizing and exploring high-dimensional datasets, separate from statistical analysis
- The Mapper Algorithm[1] reduces dimensionality while maintaining structural features by generating clusters in the full high-dimensional space
- Cluster visualizations offer insights that can direct statistical investigation, support current methods, and foster understanding of complex interrelations between variables

Objectives:

Implement the Mapper Algorithm to

1. Visualize EEG, eye tracking, and clinical characterization data from a sample of individuals with ASD and typical developing controls (TD)
2. Assess the ability of simple methods to differentiate subgroups
3. Assess the utility of TDA for high-dimensional clinical neuroscience datasets

Methods:

- The Mapper algorithm was implemented using Javascript and Python
- Dataset from individuals with ASD and TD controls (ASD: n=81, mean age=13.71; TD: n=42, mean age=13.38)
- Clinical variables included the Child Behavior Checklist, Differential Ability Scales-II, and Vineland Adaptive Behavior Scales-II
- Event Related Potential (ERP) variables included amplitude and latency at the P100 and N170 in response to dynamic faces
- Eye tracking variables included dwell time and proportional dwell time in areas of interest, such as eyes and mouth, in response to dynamic faces
- Results were visualized as 2D force-directed graphs. Each node represents a group of individuals who cluster within a multidimensional space
- Binning (allocating observations of a variable within the same or different subgroups) was done with overlapping windows along a singular axis of metadata such as Distance to Measure (average distance to the k closest neighbors). This leaves us with more manageable groups of data to cluster
- Naive clustering was based on Euclidean distance as a measure of similarity in a high-dimensional space and done within each individual bin
- Edges were created between clusters with shared subjects, e.g., if an individual could belong to multiple clusters the clusters would be joined
- Subsets of data with no missing variables were examined
- Compared results between data with and without normalized measures (scaled such that mean=0, sd=1)
- Subsets of variables (e.g., only ERP measures) were used for both binning and clustering

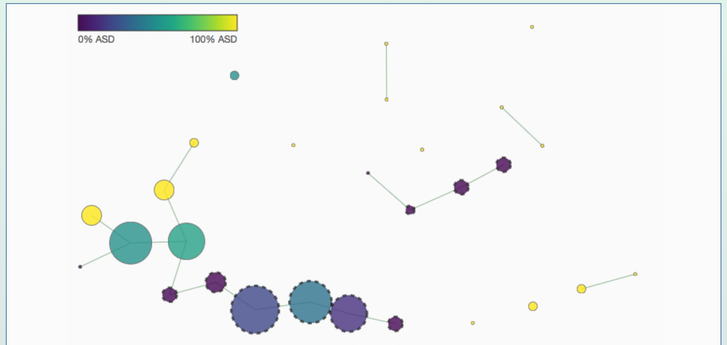

Figure 1 : Clusters of ERP, eye tracking, and behavioral data visualized with the Distance to Measure filter (k=10) and rectangular bins (resolution=15, overlap=.5). Rows with missing ERP or behavioral data were omitted, and normalized measures were used in clustering. Radius reflects the number of subjects in a node. Nodes with dotted outlines collectively contain 80% of the TD subjects (n=30), and only 5 subjects with ASD. Subjects with ASD appear more heterogeneous, with more numerous and smaller clusters.

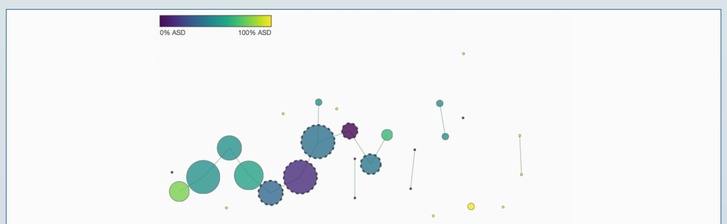

Figure 2 : The same data were visualized without normalizing the measures used in clustering. The same parameters for visualization were used. Differentiation between diagnosis in the resulting clusters is lower than before (with normalized variables), indicating that certain measures that are important to differentiation are being overshadowed when we do not normalize. Nodes with a dotted outline only contain 60% of the TD subjects, and also 7 subjects with ASD.

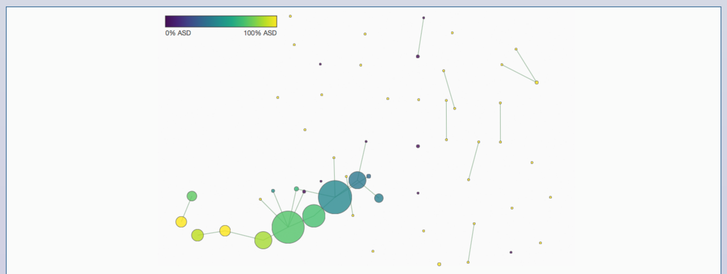

Figure 3 : The same data were visualized without normalized measures, or removing rows with missing data. The same parameters for visualization were used. When including subjects with missing data, outliers are more common, and skew measures of relative similarity between all subjects. Subject N sizes also differ considerably (ASD:n=81, TD:n=42), accounting for the increase of nodes with high percentage of ASD subjects..

References:

[1] Singh, Gurjeet, Facundo Mémoli, and Gunnar E. Carlsson. "Topological methods for the analysis of high dimensional data sets and 3d object recognition." *SPBG*. 2007.

Results:

- Created visualizations for behavioral data, eye tracking data, ERP data, and combinations of data from different modalities
- Resulting structures included areas of diagnostic similarity, suggesting that high-dimensional clustering can successfully differentiate groups in a data-driven manner
- Areas of heterogeneous data suggest more fine-tuned metrics and clustering algorithms should be explored
- *Figure 1* demonstrates a strong differentiation of diagnosis in a visualization based on 306 variables of ERP, eye tracking, and behavioral data. The nodes with a dotted outline contain 80% (n=24) of the TD population (n=30), and only 15.2% (n=5) of the ASD population (n=33)
- *Figure 2* illustrates how lack of normalization can affect clustering by altering the weights of measures relative to one another when calculating similarity. Here, lack of normalization decreases the differentiation of diagnosis
- *Figure 3* shows how the missing measures in data can create outliers and thus affect estimates of how relatively similar two subjects are compared to others. Patterns of missing data result in reduced similarity among individuals

Conclusions:

- Initial results indicated subgroups of participants that are diagnostically well-differentiated by high-dimensional neural data and other subgroups which appear more heterogeneous
- Ongoing work seeks to further analyze which measures discriminate most between subgroups, and thus improve predictive models of differences in clinical phenotype
- Normalization can help to reveal which measures are most important for creating clusters that are mostly ASD or mostly TD
- Missing data hinders proper clustering of data by introducing both outliers and unintended similarity due to missing data on overlapping measures
- TD subjects appear to be more similar to one another, as identified by larger joined clusters, promoting stronger clustering than between subjects with ASD
- Further work will examine detailed differences in subgroups for stratifying samples in clinical trials and whether smaller subgroups within diagnoses differ meaningfully
- These visualizations of latent structure within our data are a novel and valuable tool for exploring clinical datasets and building unique insights that can be applied in future research

Acknowledgements:

NIMH R01 MH100173 (McPartland), UL1 RR024139 (McPartland), NIMH R03MH079908 (McPartland), NARSAD Atherton Young Investigator Award (McPartland), NIMH R21 MH091309 (McPartland), Autism Speaks Translational Postdoctoral Fellowship (Naples), Waterloo Foundation 1167-1684 (McPartland), Patterson Trust 13-002909 (McPartland), NIMH R01 MH107426 (McPartland, Srihari), NIMH R01 MH111629-01 (Hirsch, McPartland)