# Clustering Autism Cases on Social Functioning

Nelson Ray and Praveen Bommannavar

## 1 Introduction

Autism is a highly heterogeneous disorder with wide variability in social functioning. Many diagnostic and neuropsychological tests exist which assess social functioning, but thus far few attempts have been made to classify distinct social phenotypes in autism, thereby reducing phenotypic heterogeneity and improving efforts to identify candidate genes. In this project we will use clustering procedures to stratify social phenotypes based on social functioning data available in the Autism Genetic Resource Exchange (AGRE) dataset. The next stage in the analysis would be to determine whether candidate genes involved in social functioning are associated with distinct social phenotypes.

## 2 IQ

In the Autism Spectrum Disorder (ASD) literature, IQ is consistently cited as one of the primary aspects of heterogeneity in autism.[1] An overly coarse stratification based solely on IQ would roughly identify low and high functioning groups. We examine the distribution of various IQ measures in both the autistic group and the control group in order to quantify the intelligence based heterogeneity and stratify the autistic population. There are some missing values for various IQ measures, but due to the small number of predictors in consideration we consider only the complete cases for sample sizes of 123 for the control group and 312 for the autistic group.

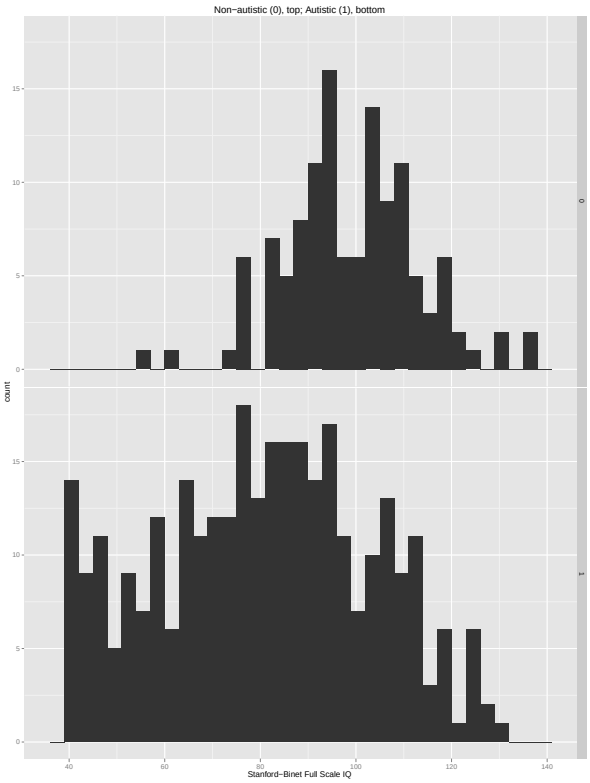
In figure 1 we see the unimodal IQ distribution for the control group and bimodal distribution for the autistic group compared with the fitted density for the autistic group, determined from fitting a mixture of Gaussians with the E-M algorithm. The number of components was chosen using the Bayesian information criterion (BIC) to balance parsimony with fit and resulted in a two component model.[2][3]

In figure 2a we see the results of PCA on all 5 IQ measures. The data were centered but not standardized because the IQ measures were already standardized. All measures are negatively correlated with the first principal component, so we think of the first principal component as being a measure of general unintelligence. Note that the full scale IQ score is highly correlated with the first principal component. As indicated by the scree plot, the first principal component explains much of the variance. This suggests that using full scale IQ alone is quite informative. The second principal component highlights the contrast between the nonverbal tests (Stanford-Binet nonverbal and Raven's Progressive Matrices) and the verbal tests (Stanford-Binet verbal and Peabody Picture Vocabulary Test).

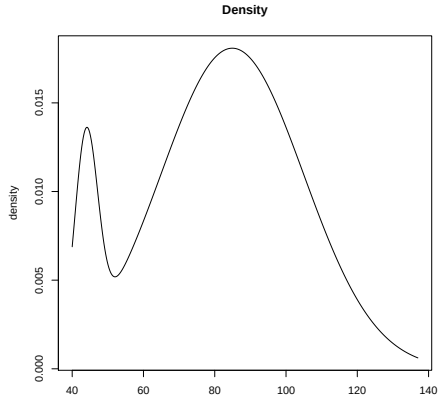
In figure 2b we fit a mixture of Gaussians to the 4 IQ measures (not counting full scale IQ) with BIC choosing a two components.[2][3]

| | 1 | 2 |
|---|---|---|
| 1 | 33 | 1 |
| 2 | 2 | 276 |

Table 1: Rows: Classifications for FSIQ score; Columns: Classifications for 4 IQ measures.




(a) The full scale IQ distributions for the control (top, unimodal) and autistic (bottom, bimodal) populations.




(b) The fitted density of full scale IQ scores for the autistic population.

Figure 1: BIC chooses the bimodal model in (b), providing a mathematical justification for what we see graphically in (a).

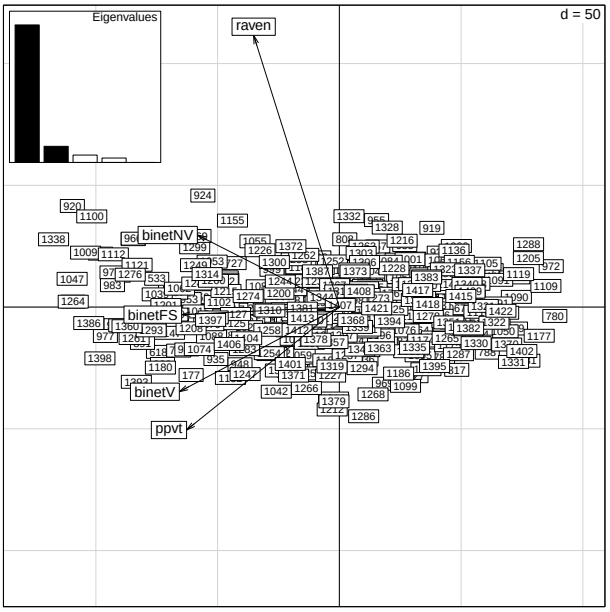
We compare the classifications of the univariate (FSIQ only) and 4 variable models. They agree on assigning 33 autistic children to the low functioning group and 276 to the high functioning group. Three are assigned to different classes. In comparing the congruence of the classifications, we find that the Rand coefficient[4] is .98 and the adjusted Rand coefficient[5] is .94, agreeing with table 1. We conclude that a full scale IQ score cutoff produces nearly identical groups to a four variable model and prefer it for simplicity. Now that we have confirmed the IQ heterogeneity described in the ASD literature and understand the relationships among the various IQ measures, we add in social features in hopes of being able to describe groups like "group 1 is low functioning and measures low on eye contact but high on peer interaction."

# 3 Clustering on Social Features

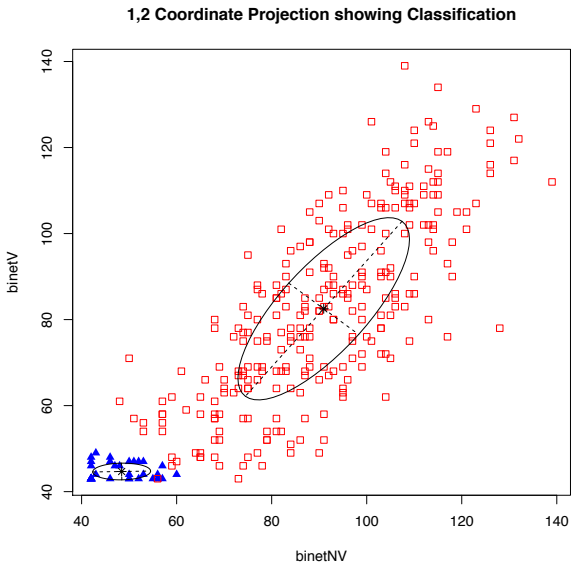
We consider responses relating to social functioning drawn from the Autism Diagnostic Interview-Revised (ADI-R), Autism Diagnostic Observation Schedule (ADOS), and Social Responsiveness Scale (SRS). Typically, questions are coded are on a 0 to 3 scale (from "behavior not present" to "extreme severity of specified behavior"). Example questions include "recognizes when something is unfair," "showing and directing attention," and "group play with peers." Scores on these tests are used in diagnosing ASD in children.

We ended up with 1422 observations of autistic and normal children with 91 social and IQ variables of interest. Fewer than 5% of the entries were missing, so we used median imputation for simplicity.

In figure 3, we see from the hierarchical clustering with average linkage on the social features that the




(a) PCA on all 5 IQ measures with "general unintelligence" as the first principal component and the contrast between verbal and nonverbal intelligence as the second.




(b) Gaussian mixture fit on 4 IQ measures with BIC choosing two components, projected down to 2 dimensions for plotting.

Figure 2: PCA and classification plots for 4 and 5 IQ measures.

| Method | k-means | Gaussian mixtures | spectral |
|---|---|---|---|
| k-means | 1 | .34 | .66 |
| Gaussian mixtures | .34 | 1 | .47 |
| spectral | .66 | .34 | 1 |

Table 2: Adjusted Rand index between various clusterings.

majority of the healthy individuals with autistic siblings are similar to one another. Due to small sample sizes of other sibling groups, no other apparent conclusions can be drawn.

There are many clustering algorithms and often no one correct technique for the data at hand. Therefore, our strategy is to apply three clustering algorithms, k-means, Gaussian mixtures, and spectral clustering, and only decide on a final stratification if there is high congruence between two or more algorithms. That is, if several orthogonal techniques all agree on the same clustering, then that is pretty good evidence for the validity of those clusters. We don't use hierarchical clustering here due to the difficulty in cutting the dendrogram correctly and just use it for insights that can be drawn from the informative visualization.

The data were standardized to put the variables on equal footing; an IQ score in the hundreds does not compare with social responses in the single digits. All clustering was done on only the autistic children since our final goal is to stratify them based on these variables. BIC chose a 3 component model for the Gaussian mixtures, and a within groups sum of squares plot on $k$ for k-means showed an "elbow" at $k = 3$ or 4. We considered spectral clustering with a radial basis kernel function and 3 clusters.

From table 2 we see that the largest "corrected-for-chance" version of the Rand index between two clusterings is .66 and the smallest is .34. This indicates that the clusterings are reasonably stable, but our highest confidence is in the clustering returned by either k-means or spectral clustering. The k-means clustering returned clusters of size 30, 67, and 112. With 91 variables, it is a bit difficult to qualify the exact differences between the groups. However, the search for the genetic factors driving the stratification can proceed with only the clusterings. To shed light on the latent structure underlying the observed social

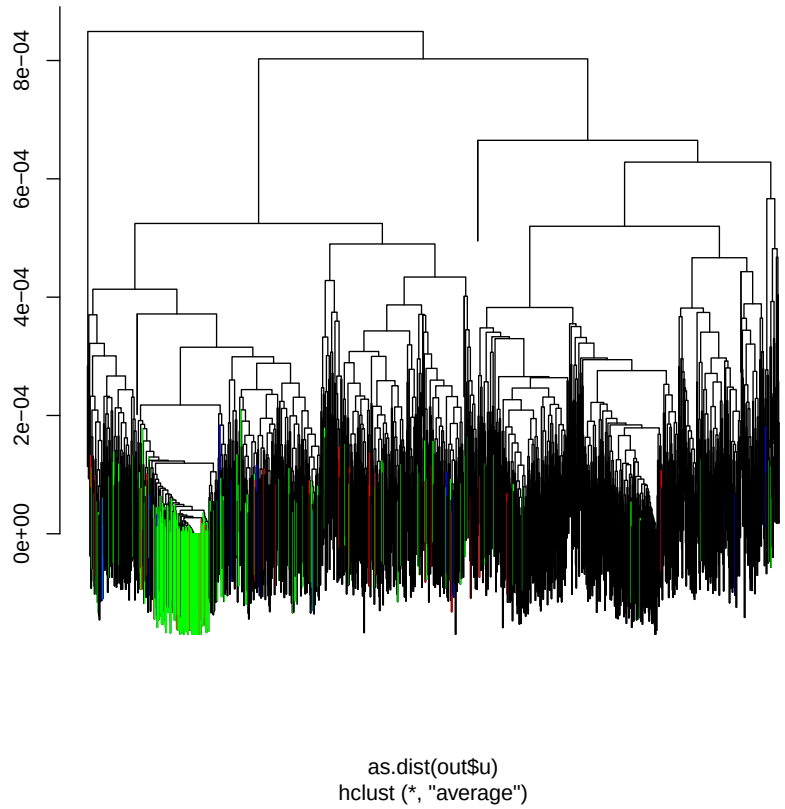



Figure 3: Hierarchical clustering with average linkage on social features. Black: Autistic; Green: Healthy with Autistic Sibling; Red: Healthy with Healthy Sibling; Blue: Healthy with no Sibling

phenotypes, we apply factor analysis.

# 4 Factor Analysis

```
Loadings:
          Factor1 Factor2 Factor3 Factor4 Factor5
CogQ17    -0.68
CommQ22   -0.68
CommQ35    0.63
CommQ37    0.63
CogQ40    -0.62
CogQ44     0.62
CogQ48    -0.64
MannerQ31          0.61
PEERPL5                    0.66
SHOW5                      0.67
QRESP5                     0.70
```

```
GRPLAY5                    0.67
binetFS                           0.97
binetNV                           0.96
binetV                            0.90
               Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings      14.20    9.82    8.96    3.40    0.26
Cumulative Var    0.16    0.27    0.37    0.40    0.41
```

For parsimony, we chose 5 factors and employed a varimax rotation to increase correlations between variables and factors. In the interest of space, we have also omitted variables with small correlations with the factors. We note that the factors explain 41% of the total variance. Drawing from the codebook for these psychological tests, we characterize factor 1 as not recognizing unfairness, not playing appropriately with children his/her age, having trouble keeping up with the flow of a normal conversation, having difficulty relating to peers, etc... Factor 2 is described by not being able to get his/her mind off something once he/she starts thinking about it. Factor 3 is imaginative play with peers, showing and directing attention, appropriateness of social responses, and group play with peers. Factor 4 is simply highly correlated with three intelligence measures.

Given the descriptions of the factors, it seems like factor 1 encompasses many antisocial tendencies, factor 2 obsessive-compulsiveness, factor 3 sociability and empathy, and factor 4 intelligence.

# 5 Conclusion

We have examined the relationships among various measures of intelligence and their role in heterogeneity in ASD populations. We have also come up with a reasonably stable stratification of ASD groups into three subgroups that are more homogeneous as measured by social phenotypes. Although not sufficient to characterize the differences between these subgroups, we have attempted to shed light on the underlying latent variables that drive the social variability.

Future work includes a genetic analysis of the subpopulations in order to find potential driving SNPs. Further qualifying the differences among the subgroups is also desirable.

# References


[1] J. Munson, G. Dawson, L. Sterling, T. Beauchaine, A. Zhou, E. Koehler, C. Lord, S. Rogers, M. Sigman, A. Estes, *et al.*, "Evidence for latent classes of IQ in young children with autism spectrum disorder," *Journal Information*, vol. 113, no. 6, 2008.

[2] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[3] C. Fraley and A. E. Raftery, *MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering*, 2006. (revised in 2009).

[4] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[5] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.