Identifying Autism From Brain Connectivity Patterns

Submitted to:
Piotr Zwiernik

Report Prepared By:
Seemal Sipra, Amy Mann

April 8, 2025

Introduction

a. Motivation

This report summarizes the results of statistical analysis performed on the Autism Brain Imaging Data Exchange (ABIDE) dataset. We have a strong interest in understanding how the brain operates and how these operations differ in different demographic groups. Hence, the ABIDE dataset strongly appealed to us as a way to learn more about systems neuroscience. For each of 47 subjects, this dataset contains multiple data variables: 110 brain regions, the subject's age at the time of scan, the subject's gender, and the subject's autism diagnosis. Furthermore, we are interested in exploring the relationships between the fMRI scan data and the various demographic variables. These characteristics and motives make the dataset highly suitable for multivariate analysis.

b. Research Question

Our research focuses on understanding brain connectivity and exploring how demographic factors affect brain connectivity patterns. In particular, we aim to answer the following questions:

1) How can we detect patterns of brain connectivity from these fMRI recordings?
2) Are there connectivity patterns that are specific to individuals with Autism but not present in the Control group?
3) Can we use this data to predict diagnosis based on brain activity alone?
4) How does age and other demographic data influence brain connectivity patterns?

c. Brief Summary of Methods

Since this dataset is high-dimensional, we would like to reduce it to a smaller set of variables in order to reduce noise and allow for easier visualization of important information while maintaining the same amount of variance in the data. Hence, we expect to use Principal Component Analysis (PCA). Furthermore, since we are interested in determining associations between brain connectivity patterns and demographic variables, we expect to use Canonical Correlation Analysis (CCA) to uncover pairs of linear combinations from these two sets that are maximally correlated. We also expect to employ Independent Component Analysis (ICA) to denoise the fMRI data by separating brain activation signals from other signals and unknown random noise.

Data and Preprocessing

a. Description of Dataset

The ABIDE dataset contains fMRI data on 47 children aged approximately 7 years to 17 years and 10 months old. Each subject has a matrix of fMRI scan data measuring brain activity in 110 regions of the brain at 196 time points.

b. Statistical Summary and Visualizations

Table 1 summarizes the demographic information of the subjects. Figure 1 displays histograms of the mean correlation values between different brain regions, with the histograms grouped by diagnosis, gender, and age group. These correlation values were averaged over time for each subject. It is interesting to note that the correlation values follow a normal-like distribution with slight skew in all 3 plots, and that this holds for all gender, age, and diagnosis groups.

	Male	Female	Total
Autism	Under 11: 3 11+ Years: 11 Total: 14	Under 11: 1 11+ Years: 6 Total: 7	Under 11: 4 11+ Years: 17 Total: 21
Control	Under 11: 8 11+ Years: 11 Total: 19	Under 11: 1 11+ Years: 6 Total: 7	Under 11: 9 11+ Years: 17 Total: 26
Total	Under 11: 11 11+ Years: 22 Total: 33	Under 11: 2 11+ Years: 12 Total: 14	Total: 47

Table 1.

Table 1. Subject counts by diagnosis, gender, and age group.

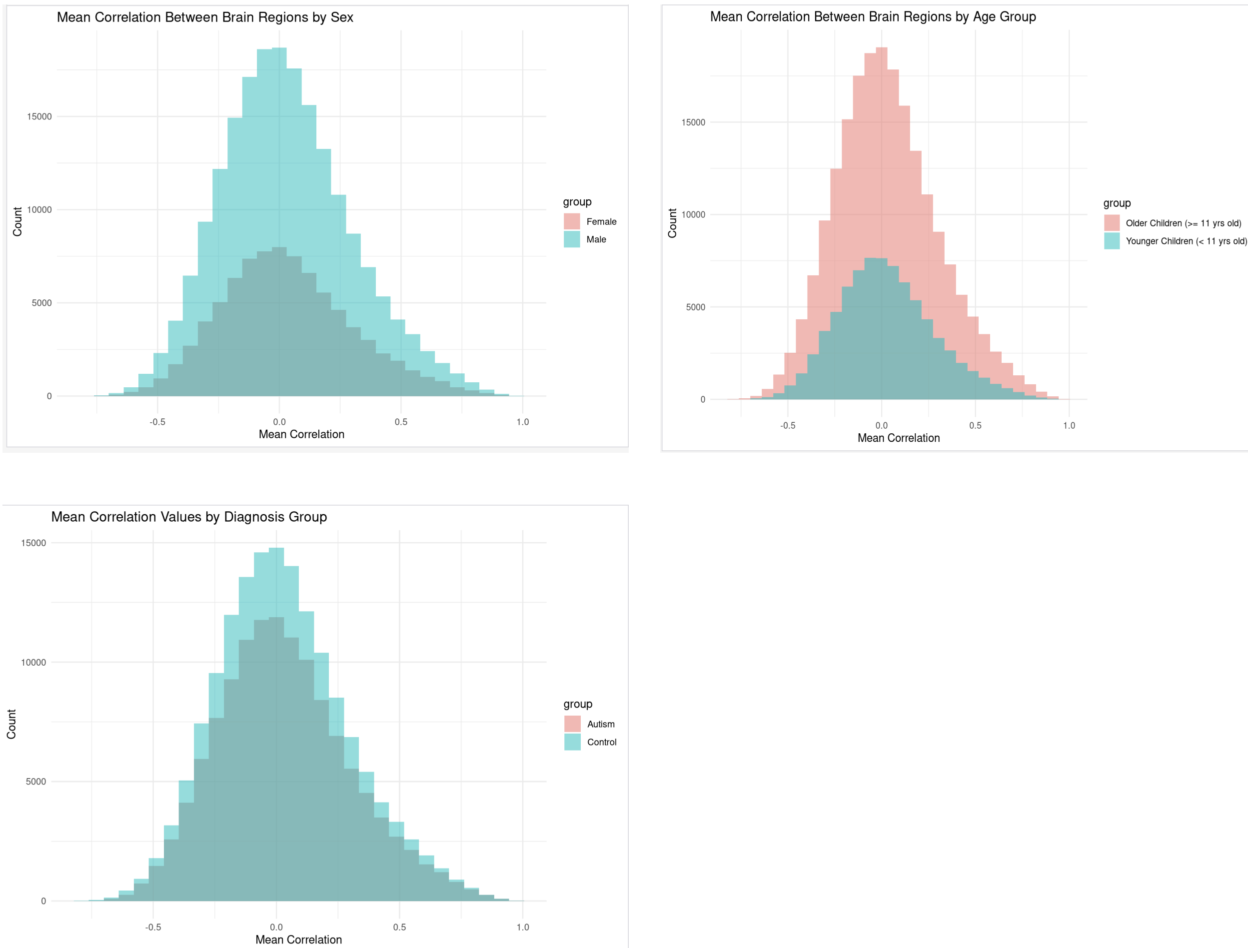

Figure 1. Count distribution of mean correlation values between different brain regions for each subject, filtered by sex, age group, and diagnosis group.

c. Handling of Missing Data

We manually searched the YALE_demo_var file of demographic data and found no issues or missing values in this dataset. Additionally, we searched the YALE_fmri file of fMRI data for non-numeric or missing values. We found no such issues.

d. Transformations

No normalization or log-scaling transformations were applied.

Methodology

Our first research question aims to uncover how brain connectivity patterns can be detected. If strong correlation exists between fMRI data in different brain regions, this suggests that the brain activity in those regions may be connected, revealing possible connectivity patterns. Hence, we initially plotted several visualizations, such as correlation heatmaps, to examine correlation values between different brain regions. However, as previously mentioned, high-dimensional data like the ABIDE dataset is not straightforward to interpret and is more prone to noise in the data from known sources (e.g. breathing and pulsation signals) and unknown sources. Thus, we performed PCA to determine which factors contribute most to the variance in the data. We then analyzed the scores and loadings of the PCA to determine which variables in the data contributed most to the principal components.

Connectivity matrices were calculated by correlating region-wise mean time series for each subject, resulting in 110×110 correlation matrices. To handle this high-dimensional data, we employed Principal Component Analysis (PCA) for dimensionality reduction, extracting key connectivity patterns. We then applied Canonical Correlation Analysis (CCA) to explore multivariate relationships between brain connectivity (via PCA components) and demographic variables (diagnosis, age, sex).

To examine direct connectivity differences between groups, we constructed Gaussian Graphical Models (GGMs) using the EBICglasso method, identifying sparse partial correlation networks separately for ASD and controls. We performed nonlinear dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) and applied k-means clustering to detect connectivity-based subtypes beyond traditional group labels. Hierarchical clustering with Ward's method on full connectivity matrices was performed for validation.

Finally, motivated by literature linking (Default Mode Network) DMN dysfunction to ASD, we examined within-DMN connectivity by averaging connectivity between DMN regions and compared it across diagnosis groups and age subgroups (under 11 vs. 11+ years) since this was common in the literature.

Results

To begin, we plotted the simply mean connectivity between the autism group and the control group. Figure X below. The autism group had a connectivity 0.0324 and the control group had a mean connectivity of 0.0279. We ran a simple t-test finding that the difference was non-significant with a 95% confidence interval for the mean difference of [-0.0023,0.0113].

We conducted principal component analysis and found that none of the ten principle components had a statistically significant relationship. This suggests that differences in connectivity between the autism and non-autism groups are more subtle and require the consideration of higher order terms. It also could reflect the presence of both hypoconnectivity and hyperconnectivity which may obscure the differences in connectivity in classic summary statistics. There are suggestions in the literature that the connectivity in certain regions of the brain are important to autism. If this is the case, it may also explain the lack of a statistically significant relationship with any of the principal components because the principal components reflect the highest variance vectors of overall connectivity, thus it would not pick up on differences that may be important to only a few specific regions.

Canonical correlation analysis (figure 4) similarly showed modest correlations (highest canonical correlation ~0.64) between connectivity and demographic factors, driven primarily by diagnosis, but results were not statistically significant (Wilks' lambda, $p = 0.26$). These findings suggest subtle differences in connectivity patterns rather than robust, group-level distinctions.

To test the hypothesis that perhaps the differences lay in a few specific regions of the brain, we performed an analysis on the Default Mode Network (DMN) region of the brain composed of (medial prefrontal cortex, posterior cingulate/precuneus, and lateral parietal regions) where the literature suggests there may be underconnectivity of those with ASD. We also split it by age as there is some indication that the connectivity patterns may differ between age groups (figure 3). There are marked differences between children under 11 and over 11. A higher variance in connectivity is seen within the children with autism over 11, while for those under 11 the overall patterns look quite similar.

DMN connectivity was lower on average in the ASD group compared to controls (0.78 vs. 0.82), though this difference was not statistically significant ($p = 0.16$). Age analysis showed a positive trend between age and DMN connectivity ($r = 0.27$, $p = 0.07$), suggesting developmental changes, with younger ASD participants showing mild hypo-connectivity and older adolescents exhibiting more variability, including hyper-connectivity outliers.

We also looked at sex differences in connectivity. Figure 2 illustrates that the overall connectivity is quite similar between male and female participants (the t-test found no statistically significant difference in the mean connectivity).

Age does seem important in the connectivity patterns. In Figure 5, each bar is a count of autistic participants within a Z-score bin, relative to the control group's mean connectivity. The dashed lines ($z=-1$ and $z=+1$) mark thresholds for hypo-connectivity (z less than or equal to 1) and hyper-connectivity (z greater than or equal to 1). Note that for the under 11 group most subjects fall into the hypo-connective or borderline typical range. No one in this group shows clear hyper-connectivity. This supports the literature suggesting reduced connectivity in younger children with ASD, particularly between networks.

In the 11 and older group there is a very broad range of z-scores. Some are hypo-connective, some are strongly hyper-connective, and many fall in the typical range. This variability suggests that older autistic individuals are more heterogeneous in mean connectivity.

Gaussian graphical modelling revealed no substantial differences in network structure between ASD and controls (see appendix). Both groups had similar overall network topologies, though the ASD network was slightly sparser, indicating subtly weaker inter-regional direct connections.

UMAP projection and k-means clustering identified three clusters based on connectivity (figure 6). These clusters did not significantly differ by diagnosis or age (chi-square tests $p > 0.55$). Nonetheless, the clusters indicated potential subgroups reflecting heterogeneous connectivity profiles: one typical connectivity cluster (mixed ASD and controls), one cluster indicative of global hyper-connectivity (dominated by ASD participants), and one cluster showing subtle hypo-connectivity.

Hierarchical clustering independently supported these findings, highlighting two clear outliers with distinct connectivity profiles (both autistic individuals), reinforcing the heterogeneity within ASD (table 2).

Clusters	1 (Autism)	2 (Control)
1	19	26
2	1	0
3	1	0

Table 2. Hierarchical clustering summary shows to clear outliers, both of whom have autism.

Our analyses indicate that functional connectivity differences in autism are subtle, complex, and highly heterogeneous. Rather than clear hypo- or hyper-connectivity profiles, we identified variability suggesting distinct connectivity subtypes within the autism population.

Nonetheless, we identified some connectivity differences between the control and autism group when examining the Default Mode Network as well as age differences in autism connectivity. Specifically, we found that hyperconnectivity is more common in children with autism under 11, whereas there is a lot of heterogeneity in children with autism over 11, several of which have significant hyperconnectivity.

Extreme outliers in the hierarchy clustering occurred only in children with autism. This suggests that brain connectivity can be an indication of autism, especially in cases of clear outliers, however brain connectivity is not sufficient to identify a person with autism alone. Indeed, most of the group-level differences were non-significant in our analysis including along ten principal components, suggesting that the connectivity differences are heterogeneous and subtle enough that it is difficult (and perhaps not possible) to identify a person with autism using brain connectivity alone.

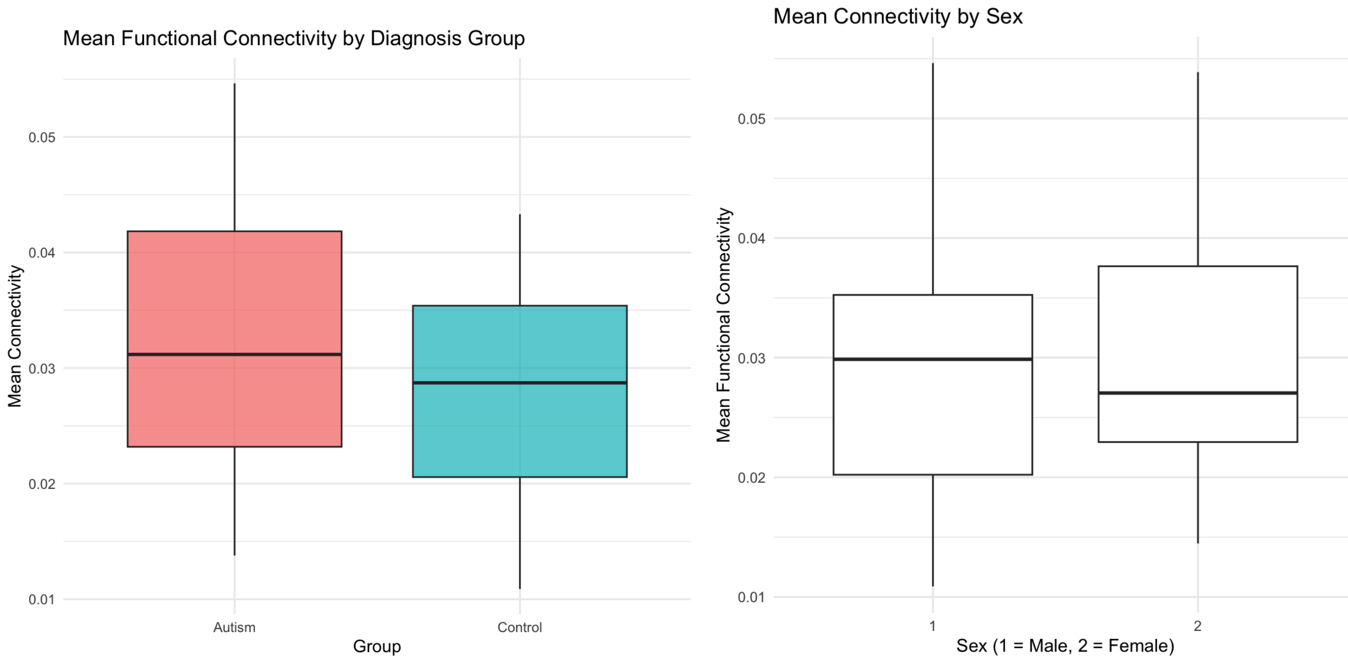

Figure 2. Mean Functional Connectivity by Diagnosis Group and by sex.

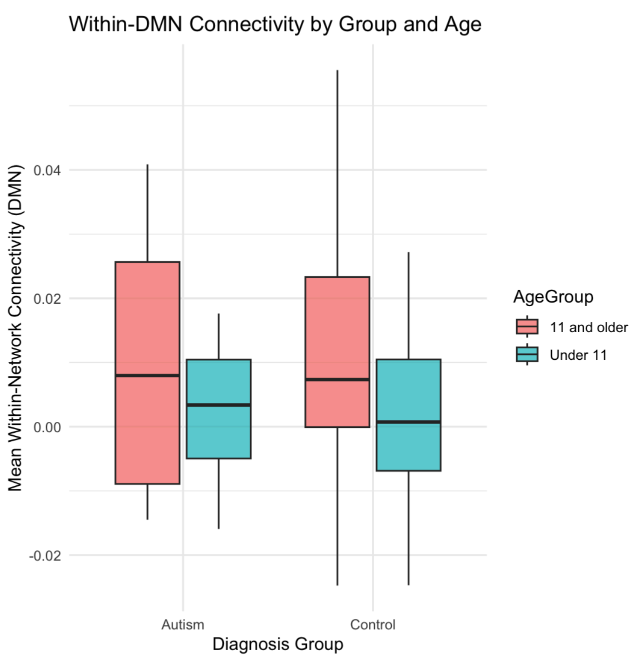

Figure 3. Mean functional connectivity by diagnosis group and age group within the Default Modal Network.

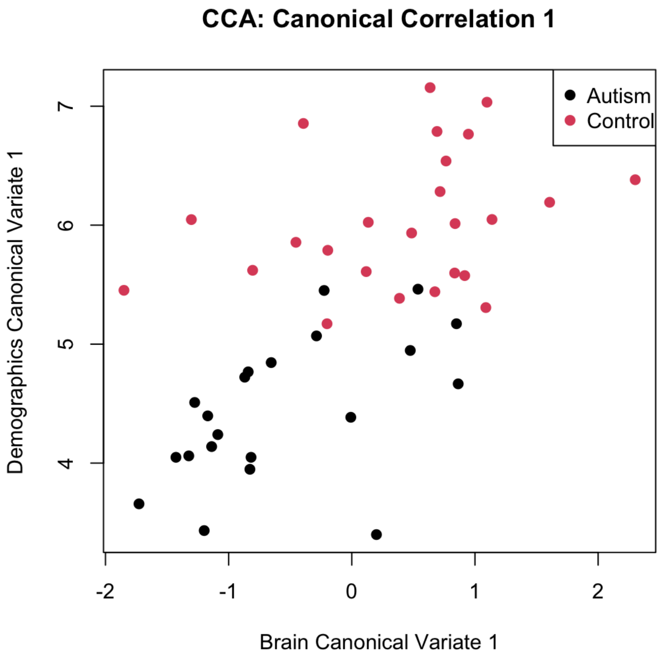

Figure 4. Canonical Correlation Analysis of with autism and control group.

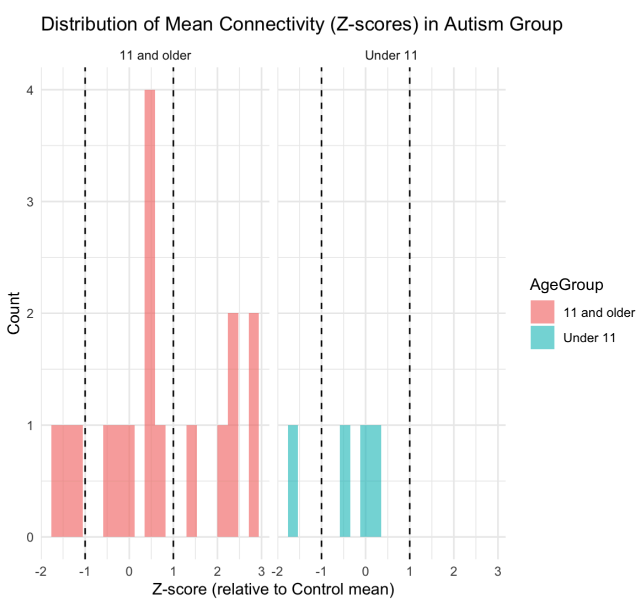

Figure 5. Each bar is a count of autistic participants within a Z-score bin, relative to the control group's mean connectivity. The dashed lines (z=-1 and z=+1) mark thresholds for hypo-connectivity (z less than or equal to 1) and hyper-connectivity (z greater than or equal to 1)..

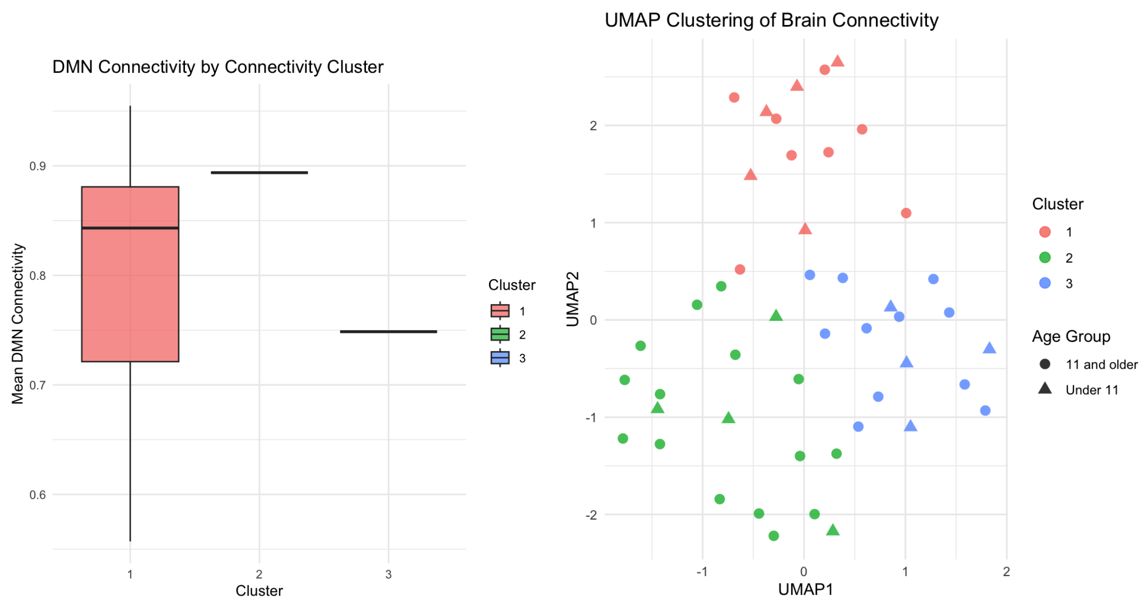

Figure 7.

Figure 6. Clustering analysis. The first image is the DMN connectivity by cluster group and the second shows the UMAP clustering where circles are those older than 11 and triangles are those under 11.

Appendix (Additional Figures)

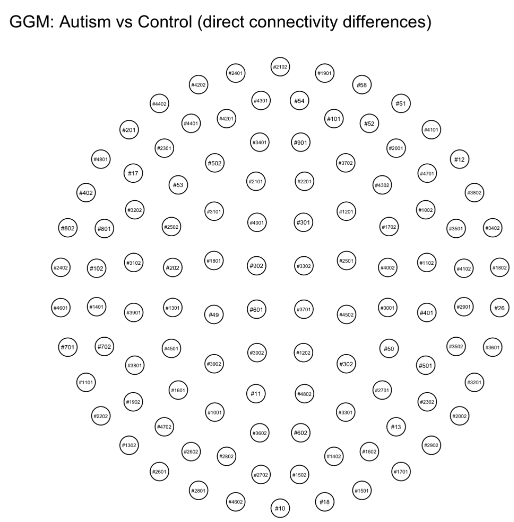

Figure 7. Gaussian Graphical Model. The graph is entirely disconnected, indicating that there are not any non-zero partial correlations.

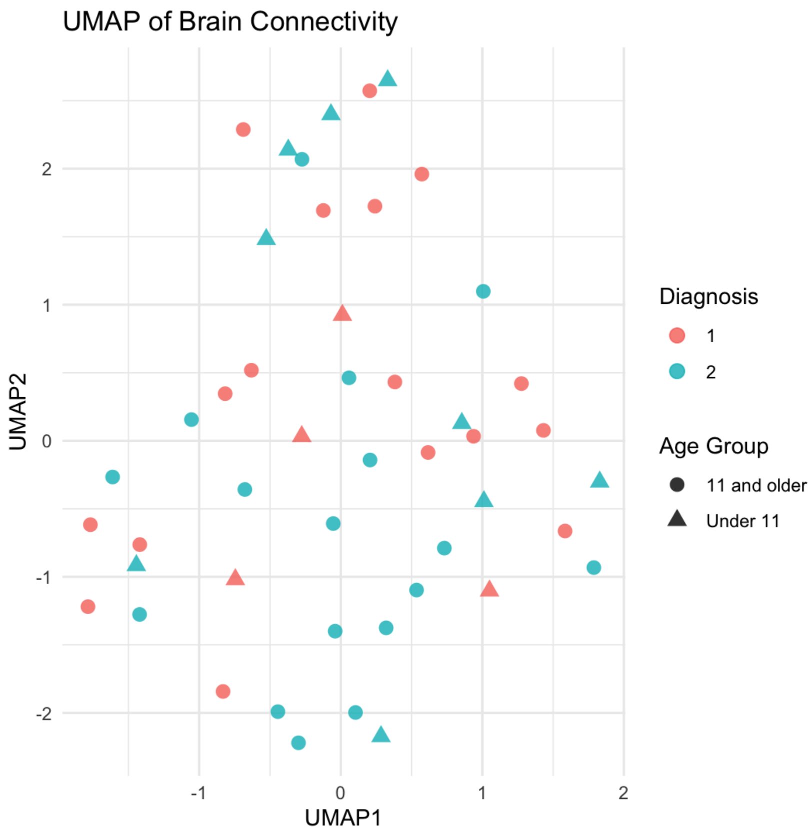

Figure 7.UMAP of Brain Connectivity with no clustering. Red represents autism diagnosis and blue represents the control group. There is no clear diagnostic separation, however age seems to be more structured because the triangles are less dispersed than the circles.

Code

```
library(dplyr)
library(ggplot2)

### loading data ######
data_path <- "/Users/amymann/Documents/STA437/ABIDE_YALE.RData"
data_path2 <- "/Users/amymann/Documents/STA437/ho_labels.csv"


load(data_path)
names <- read.csv(data_path2)
ls()

n_subjects <- length(YALE_fmri)
subject_ids <- paste0("Sub", seq_len(n_subjects))
YALE_demo_var$SubjectID <- subject_ids
head(YALE_demo_var)
```

```r
#### computing correlation matrices, assigning id, and splitting by diagnosis group ####
subject_corr_matrices <- lapply(YALE_fmri, function(mat) {
  cor(mat)
})
names(subject_corr_matrices) <- subject_ids
autism_ids  <- YALE_demo_var$SubjectID[YALE_demo_var$DX_GROUP == 1]
control_ids <- YALE_demo_var$SubjectID[YALE_demo_var$DX_GROUP == 2]
autism_corr_mats  <- subject_corr_matrices[names(subject_corr_matrices) %in% autism_ids]
control_corr_mats <- subject_corr_matrices[names(subject_corr_matrices) %in% control_ids]

average_corr_matrix <- function(mat_list) {
  mat_list <- Filter(function(x) !is.null(x) && is.matrix(x), mat_list)
  if (length(mat_list) == 0) stop("No valid matrices.")
  arr <- simplify2array(mat_list)
  apply(arr, c(1, 2), mean, na.rm = TRUE)
}

autism_mean_corr  <- average_corr_matrix(autism_corr_mats)
control_mean_corr <- average_corr_matrix(control_corr_mats)

# plotting distribution of correlation matrices
flatten_corr <- function(corr_mat) {
  corr_mat[upper.tri(corr_mat)]
}

autism_flattened  <- do.call(c, lapply(autism_corr_mats, flatten_corr))
control_flattened <- do.call(c, lapply(control_corr_mats, flatten_corr))

corr_df <- data.frame(
  corr_value = c(autism_flattened, control_flattened),
  group = c(rep("Autism", length(autism_flattened)),
            rep("Control", length(control_flattened)))
)

ggplot(corr_df, aes(x = corr_value, fill = group)) +
  geom_histogram(alpha = 0.5, position = "identity", bins = 30) +
  labs(title = "Distribution of Functional Connectivity",
       x = "Correlation", y = "Count") +
  theme_minimal()

### looking at mean connectivity and demographics ####
mean_connectivity <- sapply(subject_corr_matrices, function(corr_mat) {
  mean(flatten_corr(corr_mat), na.rm = TRUE)
```

```
})
connectivity_df <- data.frame(
  SubjectID = names(subject_corr_matrices),
  MeanConnectivity = mean_connectivity
)
analysis_df <- left_join(YALE_demo_var, connectivity_df, by = "SubjectID")
cor.test(analysis_df$AGE, analysis_df$MeanConnectivity)
ggplot(analysis_df, aes(x = factor(SEX), y = MeanConnectivity)) +
  geom_boxplot() +
  labs(title = "Mean Connectivity by Sex",
       x = "Sex (1 = Male, 2 = Female)",
       y = "Mean Functional Connectivity") +
  theme_minimal()
ggplot(corr_df, aes(x = corr_value, fill = group)) +
  geom_histogram(aes(y = ..density..), alpha = 0.5, position = "identity", bins = 30) +
  labs(title = "Density of Functional Connectivity Values",
       x = "Correlation", y = "Density") +
  theme_minimal()


### doing principal component analysis ###
pca_result <- prcomp(connectivity_matrix, scale. = TRUE)

# Get the first 2 principal components for plotting
pca_df <- data.frame(
  PC1 = pca_result$x[, 1],
  PC2 = pca_result$x[, 2],
  Group = factor(YALE_demo_var$DX_GROUP, labels = c("Autism", "Control"))
)
ggplot(pca_df, aes(x = PC1, y = PC2, color = Group)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(
    title = "PCA of Functional Connectivity Patterns",
    x = "Principal Component 1",
    y = "Principal Component 2"
  ) +
  theme_minimal() +
  scale_color_manual(values = c("steelblue", "tomato"))
t.test(PC2 ~ Group, data = pca_df)

#### doing CCA analysis ####
library(CCA)
```

```
# flattening corr matrices
flatten_corr <- function(mat) mat[upper.tri(mat)]
flat_corr_list <- lapply(subject_corr_matrices, flatten_corr)
connectivity_matrix <- do.call(rbind, flat_corr_list)

pca_df <- as.data.frame(pca_result$x[, 1:10])
colnames(pca_df) <- paste0("PC", 1:10)
pca_df$SubjectID <- paste0("Sub", 1:nrow(pca_df))

YALE_demo_var$SubjectID <- paste0("Sub", 1:nrow(YALE_demo_var))
cca_data <- merge(pca_df, YALE_demo_var, by = "SubjectID")
cca_data$SEX <- as.numeric(cca_data$SEX)
cca_data$DX_GROUP <- as.numeric(cca_data$DX_GROUP)
X <- as.matrix(cca_data[, paste0("PC", 1:10)])
Y <- as.matrix(cca_data[, c("AGE_AT_SCAN", "SEX", "DX_GROUP")])
cca_result <- cc(X, Y) #running CCA
print(cca_result$cor)

# canonical coefficients
print(cca_result$xcoef)  # for PCA components
print(cca_result$ycoef)  # for demographics
print(cca_result$cor)

# performing test for significance
can_corrs <- cca_result$cor
p.asym <- p.asym(rho = can_corrs, N = n, p = p, q = q, tstat = "Wilks")
print(p.asym)

# canonical variate scores
X_canon1 <- X %*% cca_result$xcoef[, 1]
Y_canon1 <- Y %*% cca_result$ycoef[, 1]

# plotting!
group_labels <- as.factor(cca_data$DX_GROUP)
plot(X_canon1, Y_canon1, col = group_labels,
   pch = 19, xlab = "Brain Canonical Variate 1",
   ylab = "Demographics Canonical Variate 1",
   main = "CCA: Canonical Correlation 1")
legend("topright", legend = c("Autism", "Control"),
     col = c(1, 2), pch = 19)


### trying to group by hyper and hypo connectivity (this corresponds to figure 5)
```

```
control_means <- analysis_df$MeanConnectivity[analysis_df$DX_GROUP == 2]
mean_ctrl <- mean(control_means, na.rm = TRUE)
sd_ctrl <- sd(control_means, na.rm = TRUE)
analysis_df$z_score <- (analysis_df$MeanConnectivity - mean_ctrl) / sd_ctrl
analysis_df$connectivity_type <- NA
analysis_df$connectivity_type[analysis_df$DX_GROUP == 1 & analysis_df$z_score <= -1] <- "Hypo"
analysis_df$connectivity_type[analysis_df$DX_GROUP == 1 & analysis_df$z_score >= 1] <- "Hyper"
analysis_df$connectivity_type[analysis_df$DX_GROUP == 1 & abs(analysis_df$z_score) < 1] <- "Typical"
analysis_df$AgeGroup <- ifelse(analysis_df$AGE_AT_SCAN < 11, "Under 11", "11 and older")


analysis_df %>%
  filter(DX_GROUP == 1) %>%
  group_by(AgeGroup, connectivity_type) %>%
  summarise(N = n())

ggplot(analysis_df[analysis_df$DX_GROUP == 1, ], aes(x = z_score, fill = AgeGroup)) +
  geom_histogram(position = "identity", alpha = 0.6, bins = 20) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed") +
  facet_wrap(~AgeGroup) +
  labs(title = "Distribution of Mean Connectivity (Z-scores) in Autism Group",
       x = "Z-score (relative to Control mean)", y = "Count") +
  theme_minimal()

t.test(
  MeanConnectivity ~ DX_GROUP,
  data = analysis_df[is.na(analysis_df$connectivity_type) | analysis_df$connectivity_type == "Typical", ]
)

labels <- read.csv("/Users/amymann/Documents/STA437/ho_labels.csv", header = FALSE, skip = 2)
colnames(labels) <- c("Index", "Region")
labels$Index <- as.numeric(as.character(labels$Index))


#### manually assign network label ####
labels$Network <- NA
labels$Network[grep("Prefrontal|Frontal", labels$Region, ignore.case = TRUE)] <- "Default"
labels$Network[grep("Cingulate", labels$Region, ignore.case = TRUE)] <- "Salience"
labels$Network[grep("Insula", labels$Region, ignore.case = TRUE)] <- "Salience"
labels$Network[grep("Temporal", labels$Region, ignore.case = TRUE)] <- "Social"
labels$Network[grep("Amygdala|Hippocampus|Parahippocampal", labels$Region, ignore.case = TRUE)] <-
"Limbic"
```

```
labels$Network[grep("Thalamus|Putamen|Caudate|Pallidum", labels$Region, ignore.case = TRUE)] <-
"Subcortical"
labels$Network[grep("Motor|Paracentral|Precentral|Postcentral", labels$Region, ignore.case = TRUE)] <-
"Sensorimotor"
labels$Network[grep("Occipital|Calcarine|Cuneus|Lingual", labels$Region, ignore.case = TRUE)] <- "Visual"

# within network connectivity
compute_network_connectivity <- function(corr_mat, labels) {
  n <- nrow(corr_mat)
  result <- data.frame(Within = numeric(0), Between = numeric(0), Network = character(0))

  for (net in unique(na.omit(labels$Network))) {
    idx <- which(labels$Network == net)

    within_vals <- corr_mat[idx, idx][upper.tri(corr_mat[idx, idx])]
    result <- rbind(result, data.frame(
      Network = net,
      Within = mean(within_vals, na.rm = TRUE),
      Between = NA
    ))
  }

  # between network connectivity
  nets <- unique(na.omit(labels$Network))
  between_results <- data.frame()

  for (i in 1:(length(nets)-1)) {
    for (j in (i+1):length(nets)) {
      idx1 <- which(labels$Network == nets[i])
      idx2 <- which(labels$Network == nets[j])
      between_vals <- corr_mat[idx1, idx2]
      between_results <- rbind(between_results, data.frame(
        Network1 = nets[i],
        Network2 = nets[j],
        Between = mean(between_vals, na.rm = TRUE)
      ))
    }
  }

  list(within = result, between = between_results)
}
network_results <- list()
```

```r
for (i in seq_along(subject_corr_matrices)) {
  result <- compute_network_connectivity(subject_corr_matrices[[i]], labels)
  network_results[[i]] <- result
}

within_df <- do.call(rbind, lapply(seq_along(network_results), function(i) {
  data.frame(
    SubjectID = paste0("Sub", i),
    DX_GROUP = YALE_demo_var$DX_GROUP[i],
    AGE = YALE_demo_var$AGE_AT_SCAN[i],
    SEX = YALE_demo_var$SEX[i],
    network_results[[i]]$within
  )
}))

within_dmn <- within_df[within_df$Network == "Default", ]

within_dmn$AgeGroup <- ifelse(within_dmn$AGE < 11, "Under 11", "11 and older")
within_dmn$Group <- factor(within_dmn$DX_GROUP, labels = c("Autism", "Control"))

# plotting!
ggplot(within_dmn, aes(x = Group, y = Within, fill = AgeGroup)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "Within-DMN Connectivity by Group and Age",
    x = "Diagnosis Group",
    y = "Mean Within-Network Connectivity (DMN)"
  ) +
  theme_minimal()

# extracting results and merging with earlier results
between_df <- do.call(rbind, lapply(seq_along(network_results), function(i) {
  between_vals <- network_results[[i]]$between
  dmn_sal <- subset(between_vals,
                  (Network1 == "Default" & Network2 == "Salience") |
                    (Network1 == "Salience" & Network2 == "Default"))
  data.frame(
    SubjectID = paste0("Sub", i),
    DX_GROUP = YALE_demo_var$DX_GROUP[i],
    AGE = YALE_demo_var$AGE_AT_SCAN[i],
    SEX = YALE_demo_var$SEX[i],
    DMN_Salience = dmn_sal$Between
  )
```

```r
}))

between_df <- merge(between_df, analysis_df[, c("SubjectID", "connectivity_type")], by = "SubjectID")
asds <- subset(between_df, DX_GROUP == 1)

asds_dmn <- subset(within_dmn, Group == "Autism")
asds_dmn$AgeGroup <- ifelse(asds_dmn$AGE < 11, "Under 11", "11 and older")
ggplot(asds_dmn, aes(x = AgeGroup, y = Within, fill = AgeGroup)) +
  geom_boxplot(alpha = 0.7) +
  labs(
    title = "Within-DMN Connectivity in Autism by Age Group",
    x = "Age Group",
    y = "Mean Within-Network Connectivity"
  ) +
  theme_minimal()

t.test(Within ~ AgeGroup, data = asds_dmn)


#### GAUSIAN GRAPHICAL MODELS #########
library(qgraph)

# mean correlation matrices per group
autism_mean <- average_corr_matrix(subject_corr_matrices[cca_data$DX_GROUP==1])
control_mean <- average_corr_matrix(subject_corr_matrices[cca_data$DX_GROUP==2])

# partial correlation networks
autism_ggm <- qgraph::EBICglasso(autism_mean, n=21)
control_ggm <- qgraph::EBICglasso(control_mean, n=26)

# plotting difference network
diff_network <- autism_ggm - control_ggm
qgraph(diff_network, layout="spring", title="GGM: Autism vs Control (direct connectivity differences)")

##### UMAP #####
library(uwot)

# flattening correlation matrices
flatten_corr <- function(mat) mat[upper.tri(mat)]
flat_data <- do.call(rbind, lapply(subject_corr_matrices, flatten_corr))

umap_results <- umap(flat_data, n_neighbors=10, min_dist=0.1, metric="euclidean")
umap_df <- data.frame(UMAP1=umap_results[,1],
```

```r
                    UMAP2=umap_results[,2],
                    DX_GROUP=factor(cca_data$DX_GROUP),
                    AgeGroup=ifelse(cca_data$AGE_AT_SCAN<11, "Under 11","11 and older"))

ggplot(umap_df, aes(x=UMAP1, y=UMAP2, color=DX_GROUP, shape=AgeGroup)) +
  geom_point(size=3, alpha=0.8) +
  labs(title="UMAP of Brain Connectivity",
       color="Diagnosis", shape="Age Group") +
  theme_minimal()

write.csv(umap_df, "umap_points.csv", row.names = FALSE)
umap_df <- read.csv("umap_points.csv")

#  k-means clustering
set.seed(42)
kmeans_result <- kmeans(umap_df[, c("UMAP1", "UMAP2")], centers = 3)

umap_df$Cluster <- as.factor(kmeans_result$cluster)
ggplot(umap_df, aes(x = UMAP1, y = UMAP2, color = Cluster, shape = AgeGroup)) +
  geom_point(size = 3, alpha = 0.8) +
  labs(
    title = "UMAP Clustering of Brain Connectivity",
    x = "UMAP1",
    y = "UMAP2",
    color = "Cluster",
    shape = "Age Group"
  ) +
  theme_minimal()

umap_df$Diagnosis <- cca_data$DX_GROUP
umap_df$AgeGroup <- ifelse(cca_data$AGE_AT_SCAN < 11, "Under 11", "11 and older")
table(umap_df$Cluster, umap_df$Diagnosis)
chisq.test(table(umap_df$Cluster, umap_df$Diagnosis))
chisq.test(table(umap_df$Cluster, umap_df$AgeGroup))


###### HIERARCHICAL CLUSTERING #########
library(pheatmap)

flat_data <- do.call(rbind, lapply(subject_corr_matrices, function(mat) mat[upper.tri(mat)]))
hc <- hclust(dist(flat_data), method = "ward.D2")
plot(hc, labels = cca_data$DX_GROUP, main = "Hierarchical Clustering by Connectivity")
clusters <- cutree(hc, k = 3)
```

```r
table(clusters, cca_data$DX_GROUP)
which(clusters == 2)
which(clusters == 3)

dmn_df$Cluster <- as.factor(clusters)

aov_result <- aov(DMN_connectivity ~ Cluster, data = dmn_df)
summary(aov_result)

ggplot(dmn_df, aes(x = Cluster, y = DMN_connectivity, fill = Cluster)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "DMN Connectivity by Connectivity Cluster",
       y = "Mean DMN Connectivity") +
  theme_minimal()
```