STA437H1 - Project

Introduction

Autism spectrum disorder (ASD) is a widely documented condition arising from complex genetic and environmental factors that impact the development of the brain (Gershwin, 2009). Notably, the condition has been found to affect brain connectivity patterns, resulting in executive function deficits in affected individuals (Belmonte et al., 2004). Inspired by this connection, I find the ABIDE dataset interesting as it can help indicate the variables (including ASD diagnosis) that are associated with the brain's functionality. The dataset, which contains matrices of fMRI brain activity recordings for 47 study participants, is suitable for multivariate analysis since it has a large number of dimensions and multiple potential predictor variables.

In view of the above, the guiding question for this project is, "What variables might relate to brain activity and connectivity patterns that are averaged over time?" In my analysis:

- The main variables to look at are the diagnosis, age, and sex of an individual, and any latent factors (hypothetical variables).
- The methods are exploratory data analysis, linear and non-linear dimensionality reduction, graphical models, CCA, and factor analysis. If successfully implemented, they can help provide a broad guiding overview of the data, find clustering patterns in low dimensional spaces, determine any differences in brain connectivity between various groups, and identify any latent factors, among other tasks.

In this report, I first describe the steps of data preprocessing and findings from exploratory data analysis. Then, I detail the multivariate methods, justify their relevant, and show their results. Finally, I conclude with a broad summary of my findings and provide code in the appendix.

Data preprocessing and exploratory data analysis

First, I load the required packages and the files containing the data, rename variables as necessary, and check for missing values.

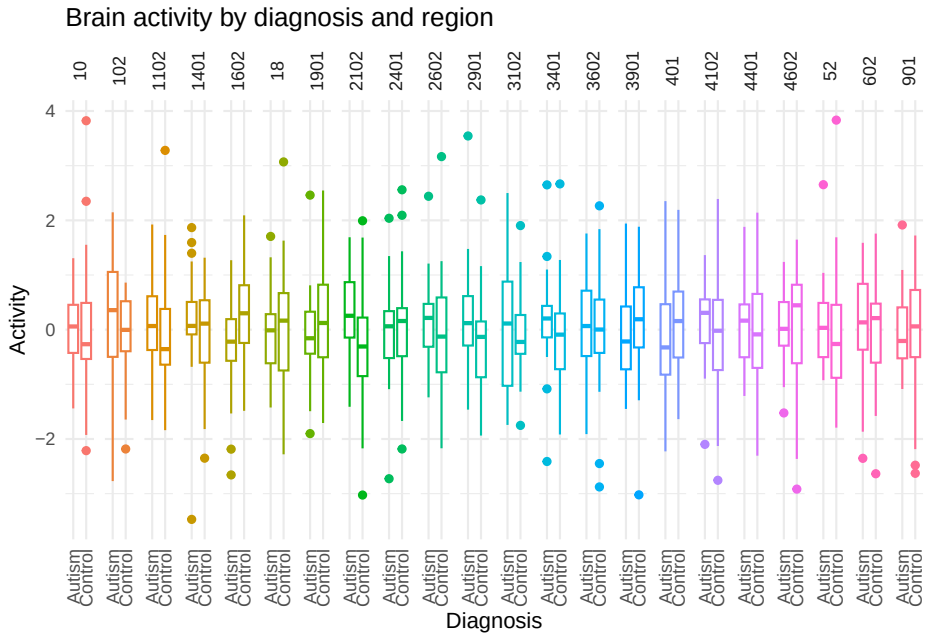
```
> Number of missing values in the fMRI data: 0.
```

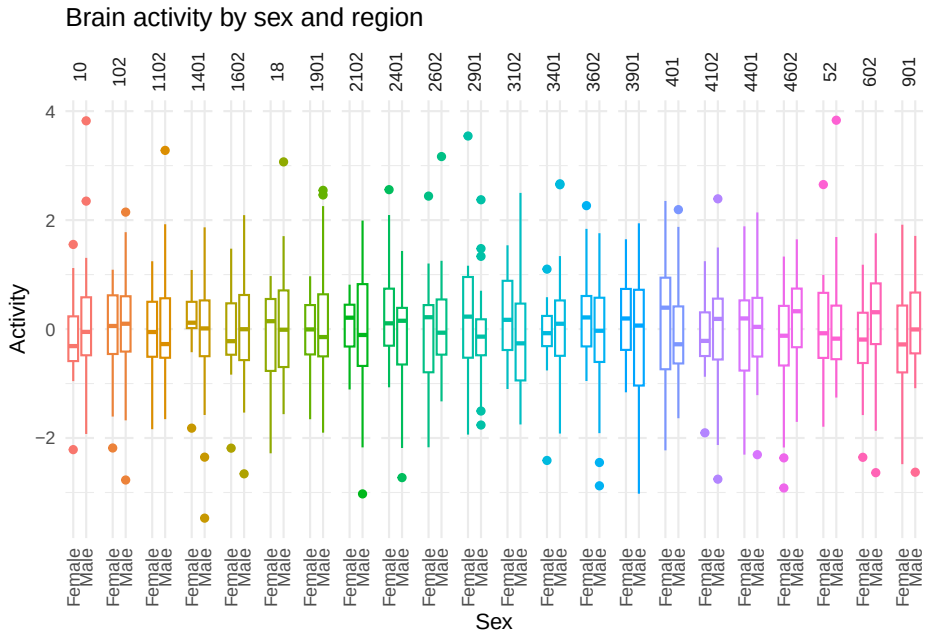
```
> Number of missing values in the demographic data: 0.
```

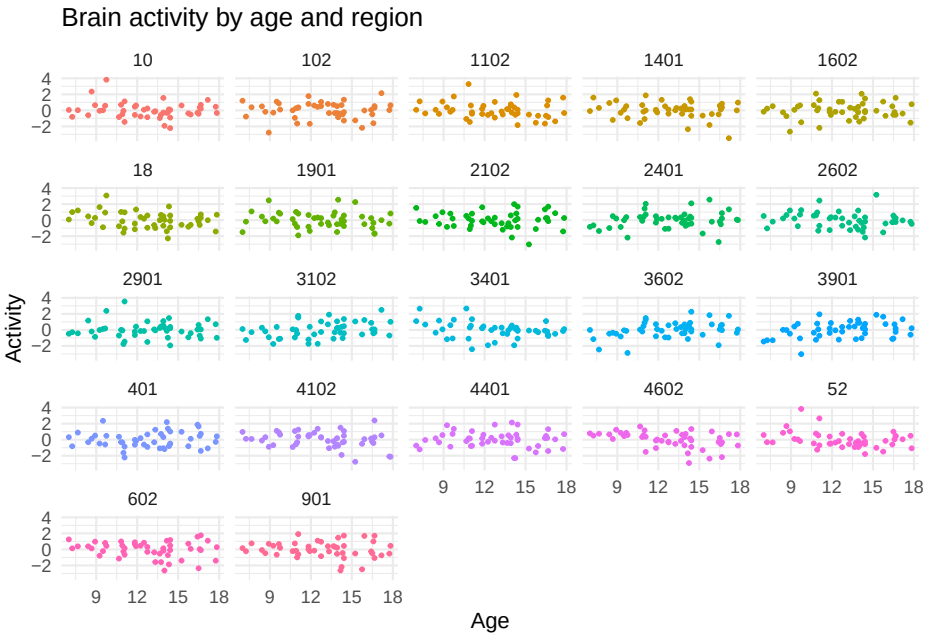
Since the data does not have any missing values, I move on to exploratory data analysis, where I summarize the demographic variables of the study participants.

```
>      group         age            sex
>  Autism :21   Min.   : 7.00   Female:14
>  Control:26   1st Qu.:10.88   Male  :33
>               Median :13.25
>               Mean   :12.80
>               3rd Qu.:14.42
>               Max.   :17.83
```

Out of the 47 participants, there are more non-autistic than autistic individuals and more than twice the number of males than females. These class imbalances, though noticeable, are not particularly extreme. In addition, the participants are all young, with the mean age being 12.8 and the maximum age being 17.8. Next, I average the brain activity recordings over time and normalize the measurements, resulting in a 47 by 110 matrix. To plot the activity with respect to the demographic variables, I only select every 5 brain regions due to space constraints and create boxplots and scatterplots faceted by region.







The plots show that there are differences in brain activity in each selected region across diagnosis, age, and sex. Thus, these variables might relate to brain activity levels.

To explore patterns of brain connectivity, I assume that regions with highly correlated brain activity patterns are more connected. I then create a heatmap corresponding to the correlation matrix between regions. Note that only every third row and column is labeled.

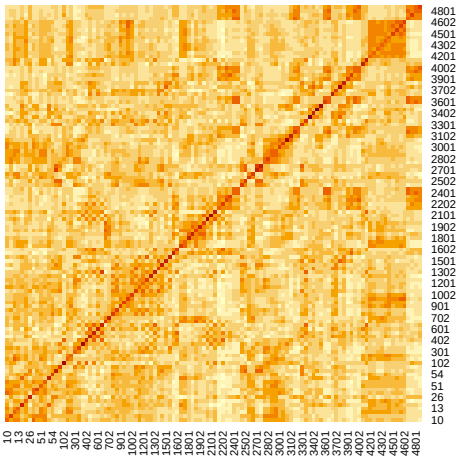


Figure 1: Heatmap of brain activity correlations across regions

The heatmap shows that patterns of brain connectivity are distinct by region. For instance, I notice that certain subsets of regions have higher connectivity; this can be clearly seen with the regions near 4201 and 4601, which are the central opercular cortex and the planum temporale respectively. Moreover, these patterns appear to differ between diagnoses, as can be seen from the following heatmaps of brain connectivity for the autism and control groups respectively.

A similar phenomenon occurs with the heatmaps for each sex (not shown). These findings support the need to further investigate the relations between the demographic variables and brain activity and connectivity.

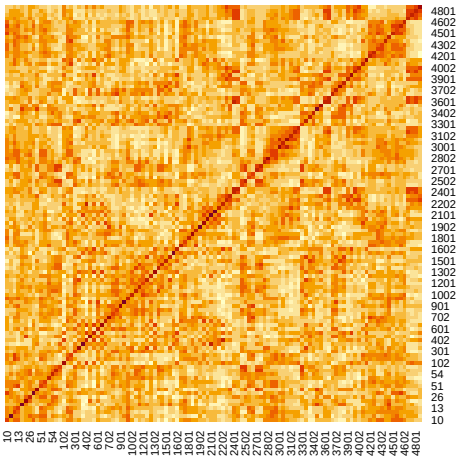


Figure 2: Heatmap of brain activity correlations across regions for the autism group

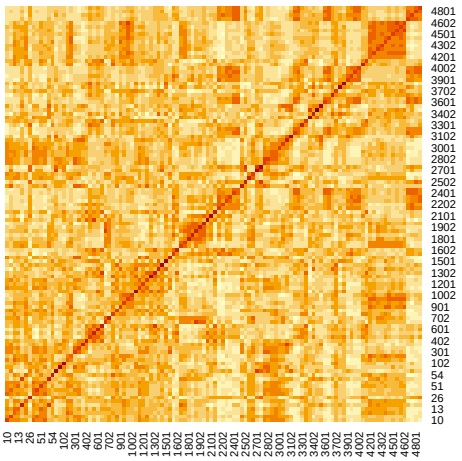
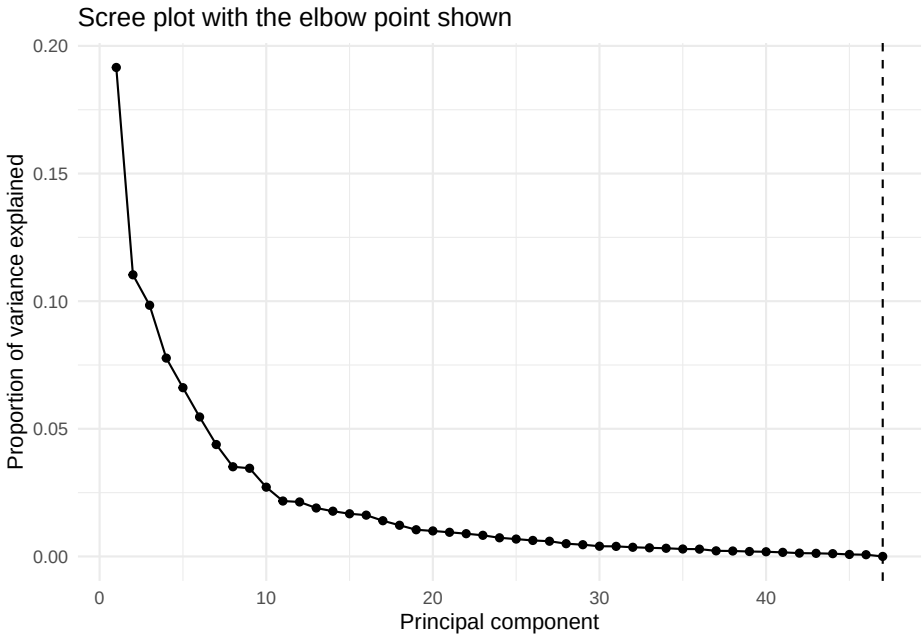


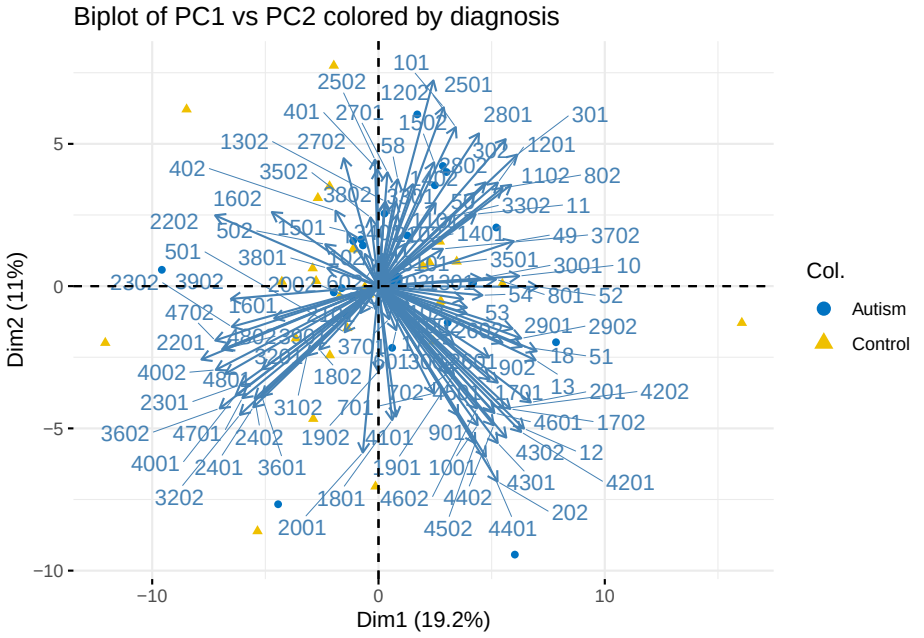
Figure 3: Heatmap of brain activity correlations across regions for the control group

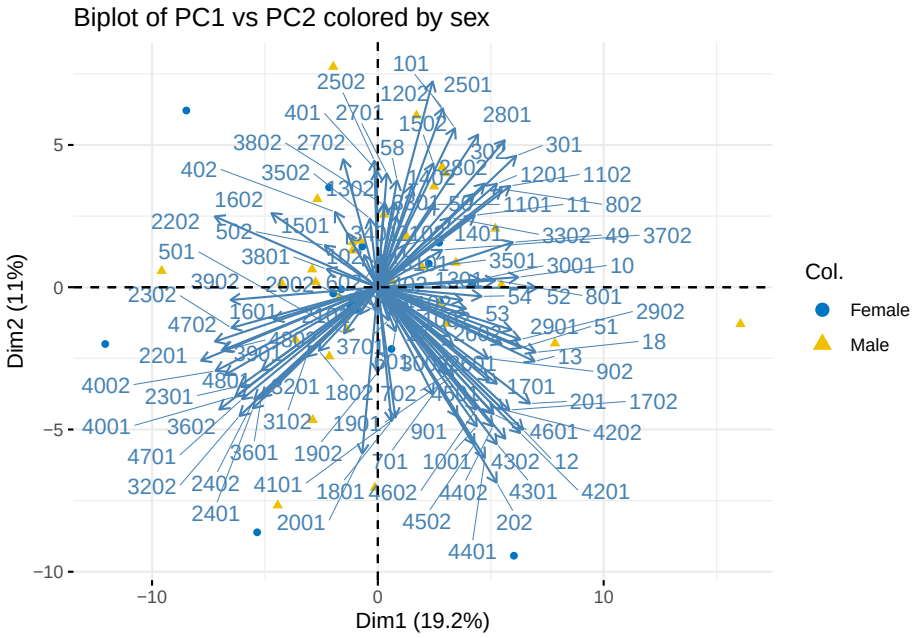
Multivariate methods and results

Dimensionality reduction

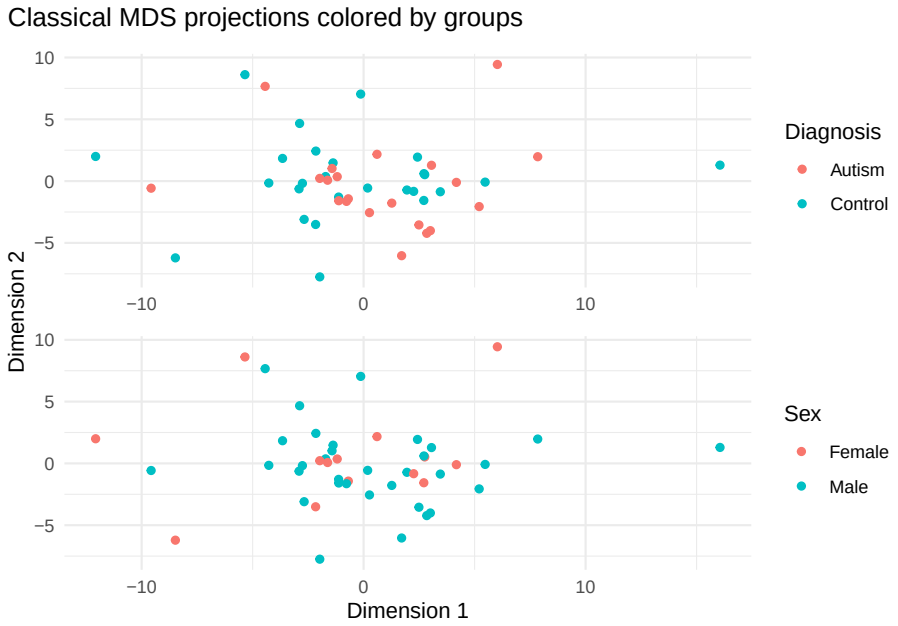
To start off the multivariate analysis, I treat the brain regions as dimensions and perform PCA to see if there is a low dimensional linear subspace with distinct clusters of demographic groups. This relates to my research question because I hypothesize that any significant differences in average brain activity by diagnosis and sex can be captured in low dimensional representations. I first plot the scree plot and calculate the elbow point to find the optimal number of principal components to keep. I then plot biplots of the first two principal components colored by diagnosis and sex.



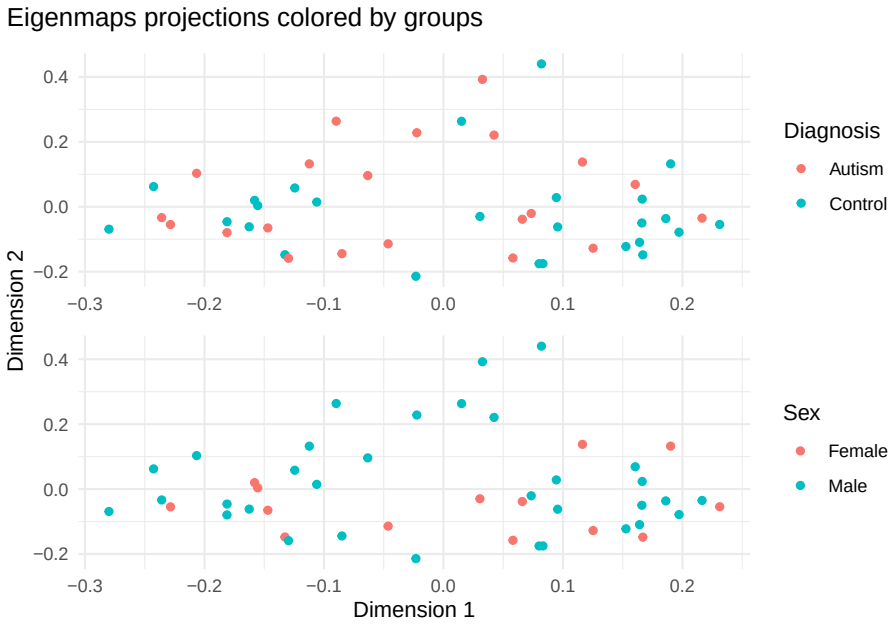


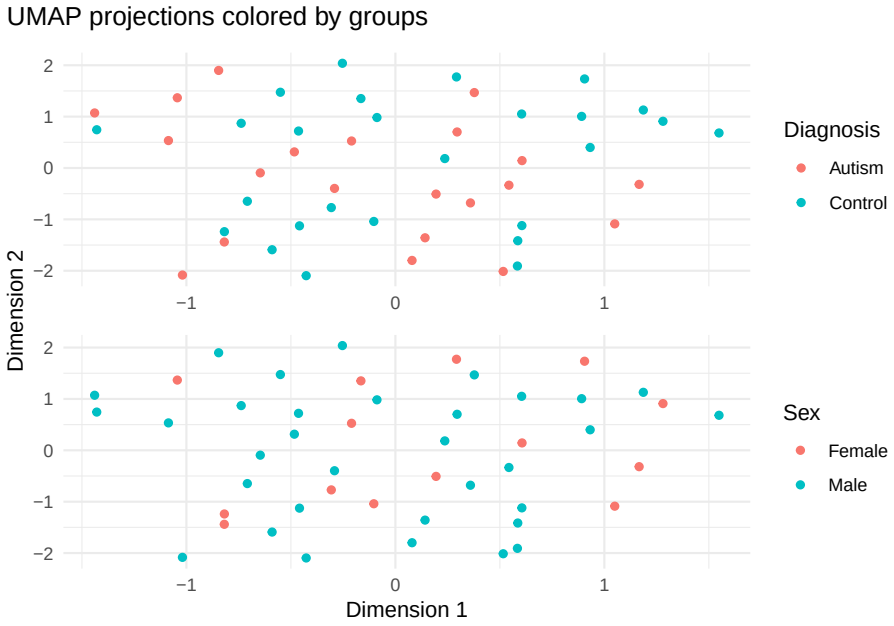


Notice that the elbow point has been calculated to be the last principal component, which is equal to the number of observations in the data. This indicates that PCA is not ideal for this task, likely because the relations between the variables are non-linear. Indeed, the biplots do not show any distinct clusters of projected brain activity by diagnosis or sex, meaning I cannot currently conclude that average brain activity varies by diagnosis and sex. This can also be seen in the results from MDS (which is equivalent to PCA in its classical form) below.



Since non-linear dimensionality reduction methods might fare better than the previous linear methods, I proceed with Laplacian eigenmaps and UMAP. These methods create low dimensional manifolds for visualizing any relations between brain activity and the demographic variables.





Again, the projections do not show any distinct clusters of points by diagnosis or sex. This may be due to, among other reasons, insufficient hyperparameter tuning in the methods or the fact that the patterns are overly complex. While these results may be disappointing, they do not indicate that there is no relation between the demographic variables and brain activity averaged over time.

Brain connectivity

To determine if diagnosis and sex relate to average brain connectivity, I construct graphical models for modeling partial correlations between regions in each demographic group. Since stepwise methods are computationally intractable due to the large number of regions, I use the graphical LASSO method and set $\lambda = 0.01$ for convex optimization.

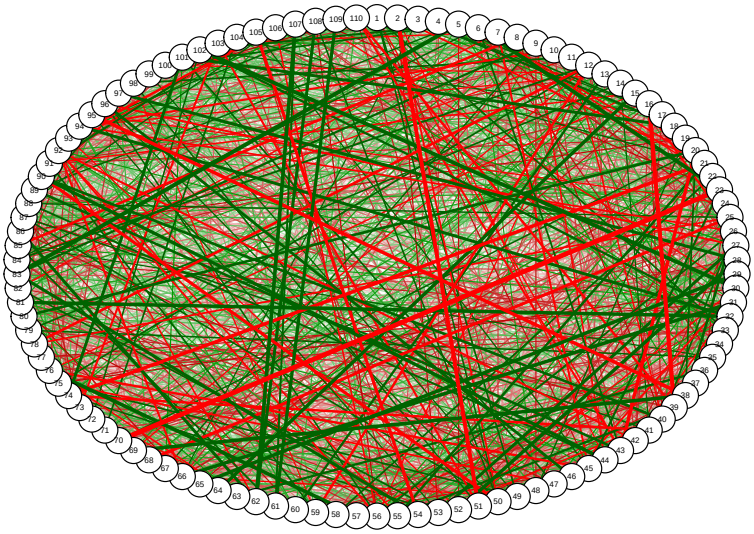


Figure 4: Partial correlation network between regions for the autism group

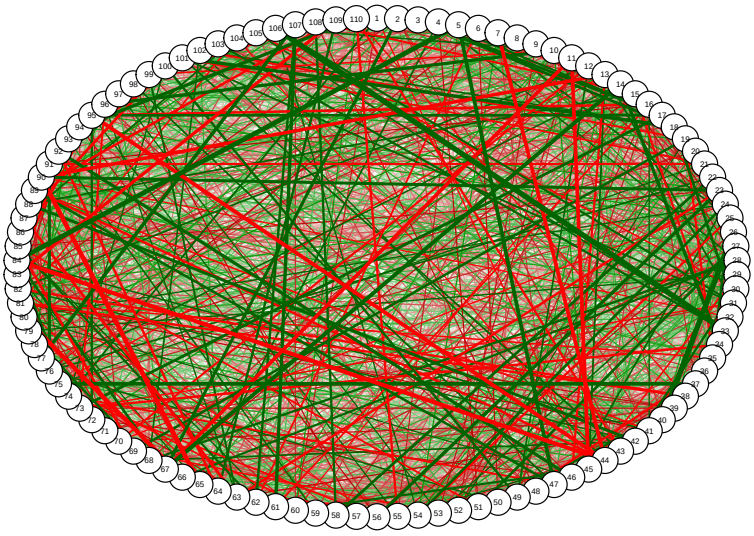


Figure 5: Partial correlation network between regions for the control group

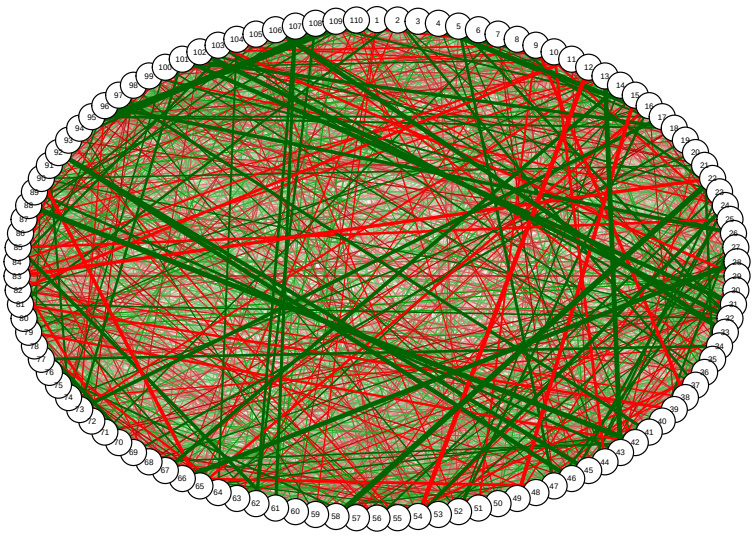


Figure 6: Partial correlation network between regions for males

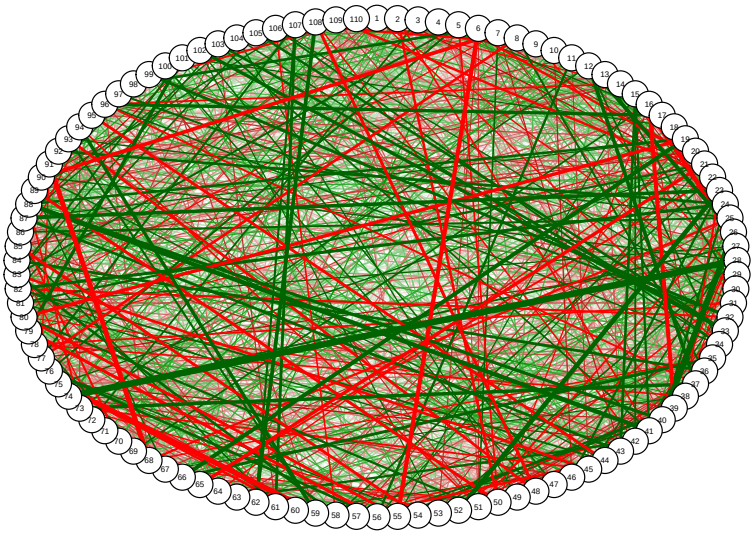
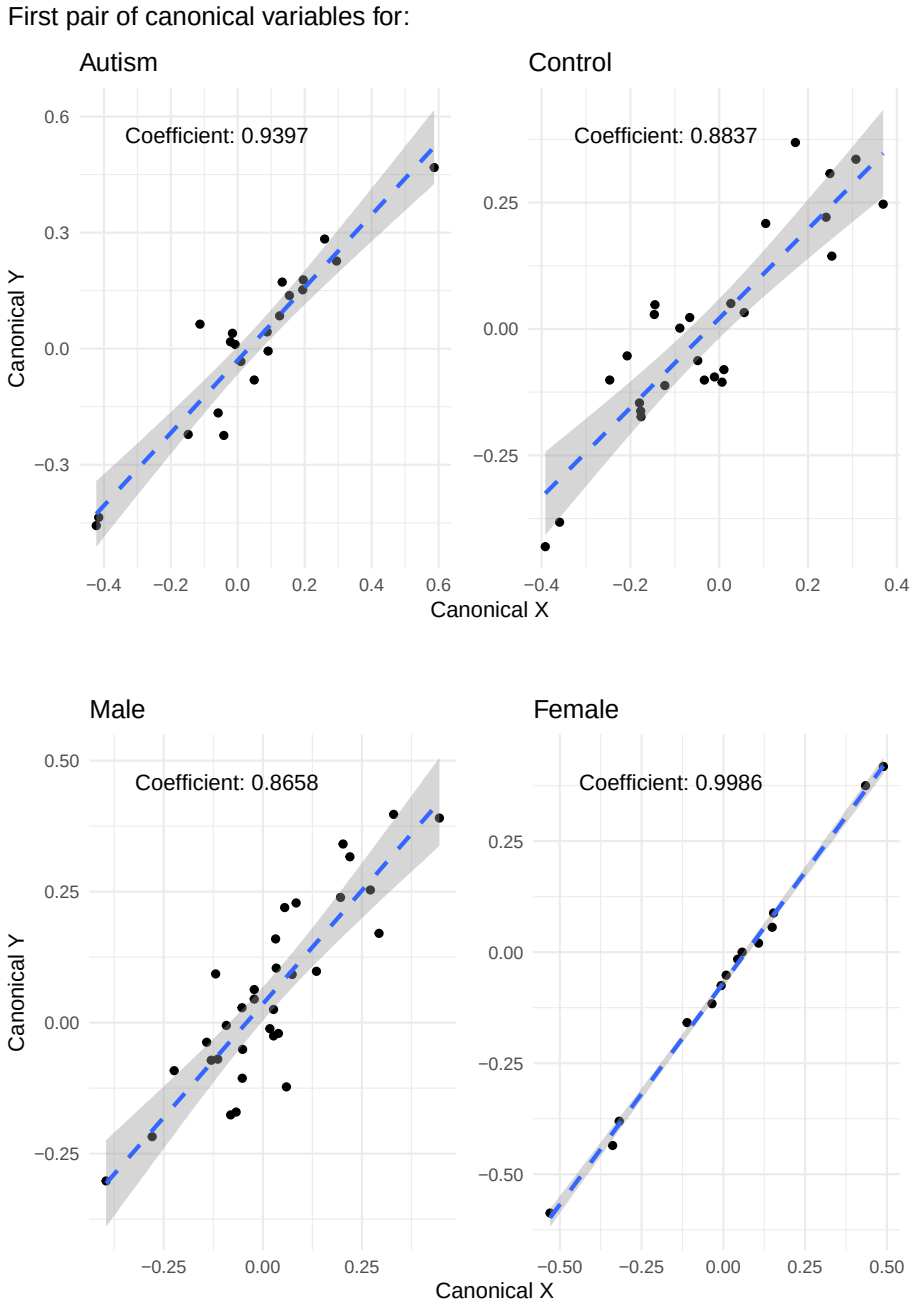


Figure 7: Partial correlation network between regions for females

The above partial correlation networks differ between each demographic group considered here. In particular, the most important correlations for each network are distinct; for instance, the network for males has a prominent connection between region numbers 12 (53, right hippocampus) and 54 (2002, left supramarginal gyrus posterior division), while the network for males has a noticeable connection between region numbers 6 (18, left amygdala) and 55 (2101, right angular gyrus). Thus, it is reasonable to conclude that diagnosis and sex likely have some sort of relation to brain connectivity.

To further verify this, I focus on a subset of the regions, partition it into two halves, and employ CCA to find the maximally correlated linear combinations between each half for each demographic group. Under this construction, each pair of linear combinations represents some sort of partial correlation within the brain. Thus, any differences in each pair between demographic groups might translate to differences in brain connectivity.



Consistent with the findings from graphical models, the maximally correlated linear combinations between each half of the subset differ depending on diagnosis and sex, further supporting the notion that diagnosis and sex likely relate to average brain connectivity.

Latent factors

To find if there are any latent factors affecting brain activity, I next perform factor analysis; however, I use two methods to resolve the issue of high correlation between regions, which affect the numerical convergence of the algorithm. In the first method, I run Horn's parallel analysis to determine the optimal number of factors, but with a fallback to setting the number to 5 if numerical issues are encountered in the algorithm. In the second method, I first remove highly correlated regions (empirically > 0.6462), run Horn's parallel analysis, and proceed with the algorithm. The results of both are shown below as a figure and text output respectively.

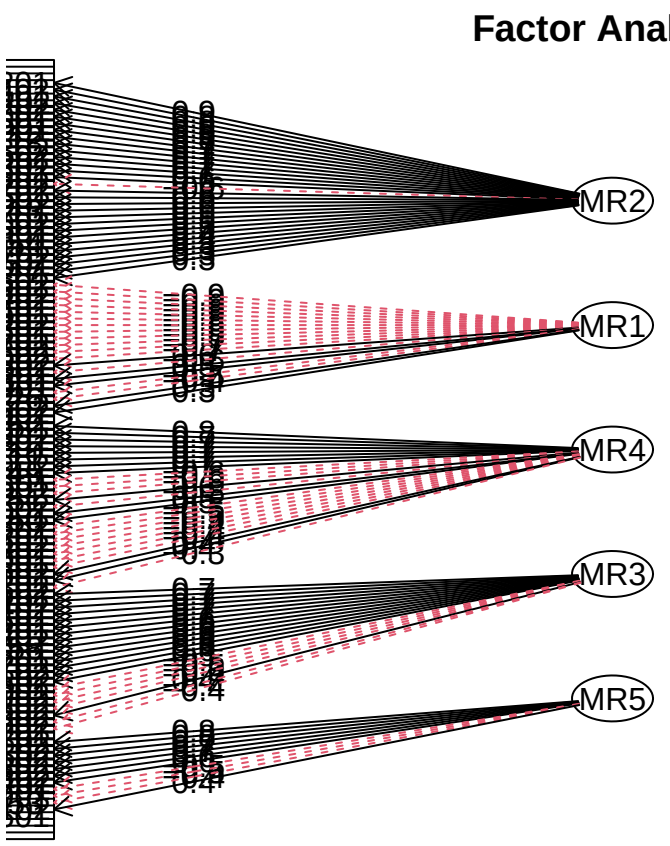


Figure 8: Visualization of the factor structure using the first method

```
>                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
> SS loadings     5.1780  4.4999  3.9537  3.8989  3.8289  3.2836  2.2248
> Proportion Var  0.1126  0.0978  0.0860  0.0848  0.0832  0.0714  0.0484
> Cumulative Var  0.1126  0.2104  0.2963  0.3811  0.4643  0.5357  0.5841
>
> Test of the hypothesis that 7 factors are sufficient.
> The chi square statistic is 1207.77 on 734 degrees of freedom.
> The p-value is 1.01e-25
```

The first method obtains the fallback value of 5 for factor analysis; the relations between the latent factors and the brain regions can be seen in the figure. However, this value and resulting factor structure are not based on the data. By contrast, according to the text output, the second method finds 7 latent factors with a cumulative explained variance of 0.5841, though it rejects the hypothesis that they are sufficient using a chi-squared test. As such, neither method performs particularly well, and I cannot determine the number of latent factors that underlie brain activity.

Conclusion

Using dimensionality reduction, I cannot conclude if diagnosis and sex relate to average brain activity levels, despite finding potential indicators for associations between them in the exploratory data analysis section. Furthermore, I cannot determine the factor structure that underlie brain activity. Nevertheless, I find that average brain connectivity patterns do differ by diagnosis and sex using graphical models and CCA. These results are not entirely surprising: dimensionality reduction is likely not the ideal method for finding relations between variables in this task, while the functionality of the brain might be too complex to be able to be summarized into distinct latent factors. With regards to brain connectivity, previous literature have identified abnormal connectivity patterns as key biomarkers of autism (Benkarim et al. 2020; Maximo et al., 2014), which is consistent with my findings. Moving forward, further analysis is required to elucidate the interplay between demographics, autism, and brain connectivity.

Citations

Belmonte, M. K., Allen, G., Beckel-Mitchener, A., Boulanger, L. M., Carper, R. A., & Webb, S. J. (2004). Autism and abnormal development of brain connectivity. The Journal of Neuroscience, 24(42), 9228-9231. https://doi.org/10.1523/jneurosci.3340-04.2004

Benkarim, O., Paquola, C., Park, B., Hong, S., Royer, J., De Wael, R. V., Lariviere, S., Valk, S., Bzdok, D., Mottron, L., & Bernhardt, B. (2020). Functional idiosyncrasy has a shared topography with group-level connectivity alterations in autism. https://doi.org/10.1101/2020.12.18.423291

Geschwind, D. H. (2009). Advances in autism. Annual Review of Medicine, 60(1), 367-380. https://doi.org/10.1146/annurev.med.60.053107.121225

Maximo, J. O., Cadena, E. J., & Kana, R. K. (2014). The implications of brain connectivity in the neuropsychology of autism. Neuropsychology Review, 24(1), 16-31. https://doi.org/10.1007/s11065-014-9250-0

Appendix

```
# Setting global knitr parameters (e.g., hiding code until the end)
knitr::opts_chunk$set(comment = ">", echo = F)
# Loading required packages
library(factoextra)
library(ggplot2)
library(glasso)
library(paran)
library(patchwork)
library(psych)
library(qgraph)
library(Rdimtools)
library(tidyr)
```

```r
library(umap)
theme_set(theme_minimal())
# Loading the data and renaming variables for clarity
load("./ABIDE_YALE.RData")
fmri <- YALE_fmri
demo <- YALE_demo_var
colnames(demo) <- c("group", "age", "sex")
demo$group <- as.factor(ifelse(demo$group == 1, "Autism", "Control"))
demo$sex <- as.factor(ifelse(demo$sex == 1, "Male", "Female"))
# Checking for missing values
cat(paste("Number of missing values in the fMRI data: ",
          sum(is.na(fmri)), ".", sep = ""))
cat(paste("Number of missing values in the demographic data: ",
          sum(is.na(demo)), ".", sep = ""))
# Summary table
summary(demo)
# Averaging brain activity over time and renaming variables for clarity
avg_fmri <- do.call(rbind, lapply(fmri, function(x) apply(x, 2, mean)))
colnames(avg_fmri) <- gsub("^#", "", colnames(avg_fmri))
# Normalizing the averages
avg_fmri <- scale(avg_fmri)
# Combining the avg_fmri and demo matrices into long format
combined <- as.data.frame(cbind(avg_fmri[, seq(1, 110, 5)], demo))
combined <- combined |> pivot_longer(cols = -c(group, age, sex),
                                     names_to = "region", values_to = "value")
# Boxplots and scatterplots by region
combined |> ggplot(aes(group, value, color = region)) +
  geom_boxplot(lwd = .5) + facet_grid(~region) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        strip.text.x = element_text(angle = 90), legend.position = "none") +
  labs(title = "Brain activity by diagnosis and region", x = "Diagnosis",
       y = "Activity")
combined |> ggplot(aes(sex, value, color = region)) +
  geom_boxplot(lwd = .5) + facet_grid(~region) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        strip.text.x = element_text(angle = 90), legend.position = "none") +
  labs(title = "Brain activity by sex and region", x = "Sex", y = "Activity")
combined |> ggplot(aes(age, value, color = region)) + geom_point(cex = .7) +
  facet_wrap(~region) + theme(legend.position = "none") +
  labs(title = "Brain activity by age and region", x = "Age", y = "Activity")
# Heatmaps
heatmap(cor(avg_fmri), Colv = NA, Rowv = NA)
heatmap(cor(avg_fmri[which(demo$group == "Autism"),]), Colv = NA, Rowv = NA)
heatmap(cor(avg_fmri[which(demo$group == "Control"),]), Colv = NA, Rowv = NA)
# PCA
pca_result <- prcomp(avg_fmri)
var_exp <- pca_result$sdev^2 / sum(pca_result$sdev^2)
scree_plot <- data.frame(pc = 1:47, var_exp = var_exp)
elbow <- which.min(var_exp[2:47] / var_exp[1:46]) + 1
scree_plot |> ggplot(aes(pc, var_exp)) + geom_point() + geom_line() +
  geom_vline(xintercept = elbow, linetype = "dashed") +
  labs(title = "Scree plot with the elbow point shown",
       x = "Principal component", y = "Proportion of variance explained")
```

```r
fviz_pca_biplot(pca_result, axes = c(1, 2), geom = "point",
                col.ind = demo$group, palette = "jco", repel = T,
                title = "Biplot of PC1 vs PC2 colored by diagnosis")
fviz_pca_biplot(pca_result, axes = c(1, 2), geom = "point",
                col.ind = demo$sex, palette = "jco", repel = T,
                title = "Biplot of PC1 vs PC2 colored by sex")
# MDS
mds_result <- cmdscale(dist(avg_fmri), k = 2)
mds_result |> data.frame() |>
  ggplot(aes(x = X1, y = X2, color = demo$group)) + geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2", color = "Diagnosis") +
mds_result |> data.frame() |>
  ggplot(aes(x = X1, y = X2, color = demo$sex)) + geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2", color = "Sex") +
plot_layout(nrow = 2, axis_titles = "collect") +
plot_annotation(title = "Classical MDS projections colored by groups")
# Laplacian eigenmaps
lapeig_result <- do.lapeig(avg_fmri, ndim = 2, type = c("proportion", 0.1))
lapeig_result$Y |> as.data.frame() |>
  ggplot(aes(V1, V2, color = demo$group)) + geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2", color = "Diagnosis") +
lapeig_result$Y |> as.data.frame() |>
  ggplot(aes(V1, V2, color = demo$sex)) + geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2", color = "Sex") +
plot_layout(nrow = 2, axis_titles = "collect") +
plot_annotation(title = "Eigenmaps projections colored by groups")
# UMAP
umap_result <- umap(avg_fmri)
umap_result$layout |> as.data.frame() |>
  ggplot(aes(V1, V2, color = demo$group)) + geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2", color = "Diagnosis") +
umap_result$layout |> as.data.frame() |>
  ggplot(aes(V1, V2, color = demo$sex)) + geom_point() +
  labs(x = "Dimension 1", y = "Dimension 2", color = "Sex") +
plot_layout(nrow = 2, axis_titles = "collect") +
plot_annotation(title = "UMAP projections colored by groups")
# Graphical models
qgraph(glasso(cor(avg_fmri[which(demo$group == "Autism"),]), 0.01)$w,
       graph = "pcor")
qgraph(glasso(cor(avg_fmri[which(demo$group == "Control"),]), 0.01)$w,
       graph = "pcor")
# Graphical models
qgraph(glasso(cor(avg_fmri[which(demo$sex == "Male"),]), 0.01)$w,
       graph = "pcor")
qgraph(glasso(cor(avg_fmri[which(demo$sex == "Female"),]), 0.01)$w,
       graph = "pcor")
# CCA
cca_helper <- function(arr, value) {
  df <- avg_fmri[which(arr == value), seq(1, 110, 9)]
  ind <- floor(ncol(df) / 2)
  X <- df[, 1:ind]
  Y <- df[, (ind + 1):ncol(df)]
  cca_result <- cancor(X, Y)
```

```r
  data.frame(x = (as.matrix(X) %*% cca_result$xcoef)[, 1],
             y = (as.matrix(Y) %*% cca_result$ycoef)[, 1]) |>
    ggplot(aes(x, y)) + geom_point() +
    geom_smooth(formula = y ~ x, method = "lm", linetype = "dashed") +
    labs(title = value, x = "Canonical X", y = "Canonical Y") +
    annotate("text", -Inf, Inf, hjust = -0.25, vjust = 3.5,
             label = paste("Coefficient:", round(cca_result$cor[1], 4)))
}
cca_helper(demo$group, "Autism") + cca_helper(demo$group, "Control") +
plot_layout(ncol = 2, axis_titles = "collect") +
  plot_annotation(title = "First pair of canonical variables for:")
cca_helper(demo$sex, "Male") + cca_helper(demo$sex, "Female") +
plot_layout(ncol = 2, axis_titles = "collect")
# First parallel analysis
# pa_result <- fa.parallel(avg_fmri, fa = "fa")
# num_factors <- ifelse(!is.na(pa_result$ncomp), pa_result$ncomp, 5)
# First factor analysis
fa_result <- fa(avg_fmri, nfactors = 5, rotate = "varimax")
fa.diagram(fa_result)
# Removing highly correlated regions
fa_cor <- cor(avg_fmri)
diag(fa_cor) <- 0
ind <- unique(floor(which(abs(fa_cor) > 0.6462) %% ncol(avg_fmri)))
# Second prallel analysis
# paran(avg_fmri[, -ind], 200)
# Second factor analysis
fa_result <- factanal(covmat = cov(avg_fmri[, -ind]), factors = 7, n.obs = nrow(avg_fmri))
capture.output(print(fa_result, digits = 4, cutoff = .3, sort = T),
               file = "temp.txt")
cat(readLines("temp.txt")[66:73], sep = "\n")
```