Final Project Report

Zebang Zhang (zz3309)

2025-12-16

Code repository (GitHub): https://github.com/ZebangZHANGstat/P8124-Final-Project.git

1. Goal

The goal of this project is to compare resting-state functional connectivity network structure between individuals with Autism Spectrum Disorder (ASD) and neurotypical controls (NT) using ROI-level fMRI time series (160 ROIs). I address two related questions: (i) whether estimated subject-level connectivity networks are, on average, more different between groups (ASD–NT) than within groups (ASD–ASD and NT–NT), and (ii) whether any apparent ASD–NT differences remain after accounting for alternative explanations such as time series length, age at scan, and sex.

2. Motivation

Graphical models provide a principled framework to represent conditional dependence relationships among brain regions. In a Gaussian graphical model (GGM), an edge corresponds to a nonzero entry in the precision matrix and can be interpreted as conditional dependence between two ROIs given all others. This representation is scientifically meaningful for functional connectivity because it targets *direct* statistical associations rather than marginal correlations that can be induced by shared connections with other regions. In fMRI applications, the number of nodes p can be large relative to the number of time points T, so sparse estimation is essential for stability and interpretability. More broadly, comparing ASD vs NT using estimated graphs can test whether diagnosis status is associated with systematic differences in network organization while also evaluating how sensitive such conclusions are to modeling assumptions, tuning choices, and data-quality factors that affect estimation.

3. Data

I use the ABIDE CMU subset (NIAK preprocessing) summarized in two metadata files and 27 subject-level ROI time series files. The analytic sample includes $n = 27$ subjects, with 14 ASD and 13 NT. For each subject, the data are a multivariate time series $X \in \mathbb{R}^{T \times 160}$, where T varies across subjects due to preprocessing steps (e.g., removal of frames). Group counts and the distribution of T by group are reported in (Table~1).

A known data issue is that each ROI file contains one extra column; all 27 subjects exhibited the same extra column name (`#223`). I removed this extra column and retained the first 160 ROIs for every subject. After column correction, I standardized each subject's time series by z-scoring each ROI (column-wise) within subject, ensuring each ROI has mean 0 and standard deviation 1 for that subject. This improves comparability across ROIs and subjects and stabilizes covariance-based estimation. All intermediate objects (e.g., standardized matrices, fitted graph objects) were cached to disk to make the pipeline fully reproducible for instructors and TAs.

4. Methods

4a. Algorithms and analytical methods

I estimate one undirected graph per subject using two different approaches and compare ASD vs NT using whole-network distances and network summary measures. The key design choice is to use *two complementary estimators* of GGM structure: a precision-matrix-based method (glasso) and a regression-based method (nodewise lasso). These methods target the same underlying object—conditional dependence structure—but differ in optimization and in how sparsity is induced. This provides an explicit "alternative approach'' required by the project and serves as a robustness check for conclusions about ASD vs NT differences.

Method 1: Graphical lasso (glasso) for Gaussian graphical models. For each subject, let S be the sample covariance of the standardized ROI time series. I estimate a sparse precision matrix $\hat{\Theta}$ using graphical lasso. In a GGM, the support pattern of $\hat{\Theta}$ encodes the estimated graph: $\hat{\Theta}_{jk} \neq 0$ implies an edge between nodes j and k, interpreted as conditional dependence given all other ROIs (under the GGM assumptions). For interpretation and distance calculations, I transform $\hat{\Theta}$ into the corresponding partial correlation matrix $\hat{P}$:

$$\hat{P}_{jk} = -\frac{\hat{\Theta}_{jk}}{\sqrt{\hat{\Theta}_{jj}\hat{\Theta}_{kk}}}, \qquad \hat{P}_{jj} = 1.$$

Method 2: Nodewise lasso (neighborhood selection). For each subject and each node j, I regress X_j on the remaining nodes X_{-j} using lasso and define neighbors as predictors with nonzero coefficients. I then symmetrize the directed neighborhood selections using the OR rule to obtain an undirected adjacency matrix $\hat{A}$. This approach estimates the graph through conditional regressions rather than a single global precision-matrix optimization, providing a qualitatively different estimation route for the same scientific target.

Question (i): distance-based between-vs-within comparison. To test whether ASD–NT networks are more dissimilar than within-group networks, I compute subject-to-subject graph distances and compare three pair types (ASD–ASD, ASD–NT, NT–NT). For glasso, the estimated networks are *weighted* (partial correlations). I vectorize the upper triangle of $\hat{P}$ and compute Euclidean distances between subjects. This choice treats each subject as a point in a high-dimensional space of edge weights and measures overall differences in connectivity patterns beyond simple edge presence/absence. For nodewise lasso, the estimated networks are *binary* (edge sets). I vectorize the upper triangle of $\hat{A}$ and compute Jaccard distance between edge sets. Jaccard distance is natural for binary graphs because it directly quantifies the proportion of non-overlapping edges relative to the union of edges. I summarize distances by pair type and test whether the mean between-group distance differs from the pooled mean within-group distance using a label-permutation test (Table~3a; Table~3b; Figure~1; Figure~2).

Question (ii): group comparison of sparsity and adjustment for confounders. In addition to whole-network distances, I compare ASD vs NT on network sparsity summaries (edge counts and edge density) for each method using Wilcoxon rank-sum tests and visualization (Figure~3; Figure~4). I then evaluate associations between network summaries and potential confounders (T, age at scan, sex) using Spearman correlations and simple visualizations (Table~4a; Figure~5; Figure~7). Finally, I re-test ASD vs NT differences after adjustment for T (and optionally age and sex) via (A) linear models on $\log(1 + \text{edges})$ and (B) residualization followed by Wilcoxon tests on residuals (Table~4b; Figure$_{5-\text{Figure}}$6).

4b. Statistical subroutines and objective functions

Glasso objective. For glasso, $\hat{\Theta}$ is obtained by minimizing the penalized negative Gaussian log-likelihood:

$$-\log\det(\Theta) + \operatorname{tr}(S\Theta) + \rho \left\lVert \Theta \right\rVert_{1,\text{off}},$$

where $\rho \geq 0$ is the sparsity penalty and $\|\Theta\|_{1,\text{off}}$ denotes the sum of absolute values of off-diagonal entries (diagonal entries are not penalized). The target is the subject-specific conditional dependence structure under an approximate multivariate normality assumption for the standardized time series. Regularization is necessary because $p = 160$ is large relative to typical T values, making unregularized precision estimation unstable.

Nodewise lasso subroutine. For nodewise lasso, each neighborhood is estimated by minimizing squared error with an ℓ_1 penalty. Because the data are time series with temporal dependence, I use blocked cross-validation folds to reduce information leakage from autocorrelation when selecting the lasso penalty.

Permutation testing rationale for Question (i). Pairwise distances are *not independent* because each subject contributes to many pairs. Standard parametric tests that treat pairwise distances as independent observations are therefore not well-justified. Instead, I use a label-permutation test that only relies on label exchangeability under the null hypothesis of no group effect. Specifically, I repeatedly shuffle diagnosis labels across subjects and recompute

$$\Delta = \overline{d}_{\text{between}} - \overline{d}_{\text{within}},$$

where $\overline{d}_{\text{between}}$ is the mean distance over ASD–NT pairs and $\overline{d}_{\text{within}}$ is the mean distance pooled over ASD–ASD and NT–NT pairs. The two-sided permutation p-value is computed from the empirical null distribution of Δ (Table~3b).

Adjustment models and parametric form. To assess whether ASD–NT differences persist after accounting for T and demographics, I fit linear models with covariates and also perform residual-based Wilcoxon tests. I model edge counts using $\log(1 + \text{edges})$ because edge counts can be skewed and can include zeros (particularly for very short T), and $\log(1 + \cdot)$ stabilizes variance while remaining defined at zero. The residual-based Wilcoxon procedure provides a distribution-robust complement to linear-model inference in a small sample, and the adjusted group-comparison results are summarized in (Table~4b).

4c. Tuning parameters and hyperparameters

Glasso tuning via EBIC. For glasso, I choose ρ separately for each subject using EBIC with $\gamma = 0.5$ over a log-spaced grid of 30 candidate values from 0.01 to 1.0. EBIC extends BIC with an additional complexity penalty that favors sparser graphs in high-dimensional settings, and $\gamma = 0.5$ is a conservative choice that further discourages spurious edges. A log-spaced grid is used because changes in ρ can have nonlinear effects on sparsity, and log spacing provides more uniform coverage of the sparse-to-dense regime. The selected value is denoted ρ_{best} and is used to produce $\hat{\Theta}$ and $\hat{P}$ for that subject (Table~2; Figure~7).

Nodewise lasso tuning via blocked CV and $\lambda_{1\text{se}}$. For nodewise lasso, I use 5-fold blocked cross-validation and select the penalty using the $\lambda_{1\text{se}}$ rule. Compared to $\lambda_{\min}$, $\lambda_{1\text{se}}$ typically yields sparser and more stable neighborhood estimates, which is desirable for structure learning in a small sample. I symmetrize neighborhoods using the OR rule to obtain an undirected adjacency matrix.

Permutation settings. For permutation tests, I use $B = 2000$ label permutations with a fixed random seed for reproducibility (Table~3b).

5. Results

Graph estimation summaries. Across subjects, glasso produced comparatively sparse networks, whereas nodewise lasso produced substantially denser networks, reflecting methodological differences in how sparsity is induced and how edges are defined (Table~2). Edge counts and densities by group are visualized in (Figure~3) and (Figure~4).

Question (i): between-group vs within-group network differences. Using glasso-based partial correlation networks, the distance distributions for ASD–ASD, ASD–NT, and NT–NT pairs were broadly similar (Table~3a; Figure~1), and the permutation test for Δ did not provide evidence of a difference

between pooled within-group and between-group mean distances (Table~3b). Using nodewise lasso adjacency networks with Jaccard distance, the corresponding distance distributions were again similar (Table~3a; Figure~2) and the permutation test was likewise non-significant (Table~3b). Overall, whole-network distance analyses did not detect a diagnosis-related shift in network structure beyond within-group variability in this sample.

Question (ii): alternative explanations and adjusted comparisons. Unadjusted group comparisons of sparsity summaries suggested no clear ASD–NT differences for either method, consistent with the overlap in the boxplots (Figure~3; Figure~4). Importantly, several diagnostics indicated that estimation and tuning are sensitive to effective sample size: glasso edge counts were strongly positively associated with time series length T, and EBIC-selected ρ_{best} decreased as T increased (Table~4a; Figure~5; Figure~7). This pattern is consistent with the fact that larger T yields more stable covariance estimates, allowing weaker regularization and producing less sparse graphs. Therefore, T is a plausible confounder for naive group comparisons because it can influence estimated network structure even if the underlying biology is unchanged.

After adjusting for T (and optionally age and sex), ASD–NT effects remained non-significant in residual-based Wilcoxon analyses (Table~4b). The nodewise lasso analyses produced smaller but still borderline p-values under adjustment, illustrating that quantitative conclusions can depend on the modeling class and penalty selection strategy (Table~4b). Overall, variability associated with T and related factors was at least as important as diagnosis status for explaining differences in estimated graph summaries in this dataset.

6. Conclusions

Using two subject-level graphical modeling approaches—EBIC-tuned glasso (precision-matrix-based estimation) and nodewise lasso neighborhood selection (regression-based estimation)—I found no robust evidence that ASD and NT subjects differ in overall estimated connectivity network structure in this CMU ABIDE subset. Whole-network distance analyses showed that ASD–NT distances were not detectably different from pooled within-group distances under either method (Table~3a; Table~3b; Figure~1; Figure~2). Group comparisons of sparsity summaries likewise showed substantial overlap (Figure~3; Figure~4).

A key practical conclusion is that time series length T is strongly associated with glasso-derived sparsity and with the EBIC-selected penalty ρ_{best} (Table~4a; Figure~7). This provides a concrete alternative explanation for apparent group differences in estimated networks: differences in effective sample size and preprocessing-related frame removal can influence regularization strength and sparsity, potentially masking or mimicking biological effects. After adjustment for T (and age/sex where available), the ASD–NT effect remained non-significant (Table~4b), suggesting that diagnosis status is not the dominant driver of the observed variability in these estimated networks within this sample. Finally, comparing glasso and nodewise lasso emphasized that modeling choices (objective functions, penalty selection, and binary vs weighted graph representations) can materially affect estimated sparsity and distance metrics, underscoring the importance of methodological robustness checks in small-sample neuroimaging graph analyses.