Learning Joint and Individual Structure in Network Data with Covariates

Modeling multiple heterogeneous networks

Motivation:

- We present the *common subspace independent-edge* (COSIE) model which describes a collection of networks with a shared latent structure on the vertices but potentially different connectivity patterns for each graph [1].
- COSIE encompasses many other popular network representations, including the stochastic blockmodel
- A joint spectral embedding - the *multiple adjacency spectral embedding*- leads to consistent estimation that is computationally efficient
- MASE estimates yield state-of-the-art performance on subsequent inference tasks, including dimensionality reduction, hypothesis testing and community detection

- Models for multiple network data are critical in statistical network theory and across multiple domains, including neuroscience, biology and the social sciences.
- Challenges in modeling graph differences while retaining sufficient model simplicity to render estimation feasible.
- Existing models require strong assumptions that limit their flexibility or scalability

Difussion MRI brain connectomes from the HNU1 data

Common subspace independent edge (COSIE) model

- Consider a sample of m graphs with n labeled nodes.
- Denote the graphs by their adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)} \in \{0,1\}^{n \times n}$
- Each graph $\mathbf{A}^{(i)}$ is modeled as independent-edge with parameter $\mathbf{P}^{(i)} \in \mathbb{R}^{n \times n}$

$$\mathbf{A}^{(i)}_{uv} \sim \text{Ber}(P^{(i)}_{uv}).$$

COSIE model

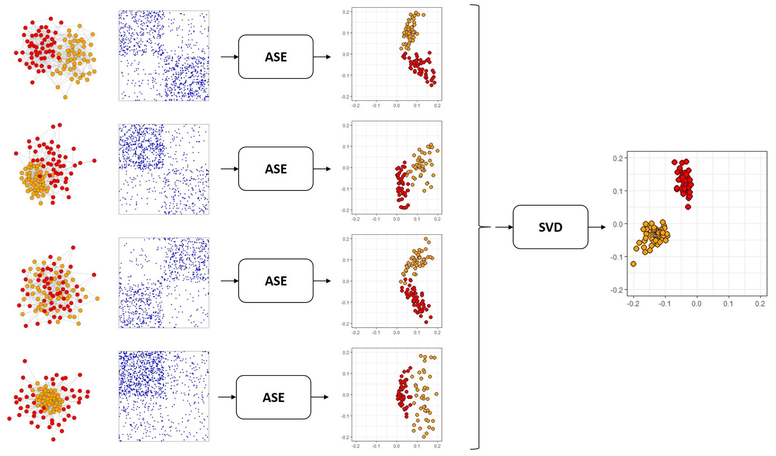
- The sample of graphs is jointly distributed according to the COSIE model if

$$\mathbf{P}^{(i)} := \mathbf{V}\mathbf{R}^{(i)}\mathbf{V}^T.$$

- $\mathbf{V} \in \mathbb{R}^{n \times d}$ is a matrix with orthogonal columns, with its rows representing vertex latent positions
- $\mathbf{R}^{(i)} \in \mathbb{R}^{d \times d}$ is a score matrix, potentially different for each graph
- The parameter d controls the complexity of the model

Multiple adjacency spectral embedding (MASE)

1. For each $i \in [m]$, obtain the d-dimensional unscaled *adjacency spectral embedding* of $\mathbf{A}^{(i)}$, denoted by $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$ and corresponding to the d leading eigenvectors
2. Let $\hat{\mathbf{U}} = \left(\hat{\mathbf{V}}^{(1)} \cdots \hat{\mathbf{V}}^{(m)}\right)$ be the $n \times (md)$ matrix of concatenated ASEs.
3. Let $\hat{\mathbf{V}} \in \mathbb{R}^{n \times d}$ be the matrix containing the d leading left singular values of $\hat{\mathbf{U}}$.
4. For each $i \in [m]$, set $\hat{\mathbf{R}}^{(i)} = \hat{\mathbf{V}}^T \mathbf{A}^{(i)} \hat{\mathbf{V}}$.



Theoretical properties

Consistency of common invariant subspace estimator

- Under some assumptions on the smallest eigenvalue of the score matrices and on the sparsity of the graphs, the estimate of $\mathbf{V}$ obtained by MASE satisfies

$$\mathbb{E}\left[\min_{\mathbf{W} \in \mathcal{O}_d} \left\|\hat{\mathbf{V}} - \mathbf{V}\mathbf{W}\right\|\right] \lesssim \sqrt{\frac{d}{mn}} + \frac{\sqrt{d}}{n}.$$

- When $\mathbf{V}$ has only d different rows, COSIE is equivalent to the stochastic blockmodel, and k-means clustering expected error in community detection is $O\left(\sqrt{\frac{d}{m}} + \frac{1}{\sqrt{n}}\right)$.

Asymptotic normality of the score matrices

- The entries of the estimated score matrices $\hat{\mathbf{R}}^{(i)}$ are asymptotically normally distributed, in particular, as the size of the graphs n increases

$$\frac{1}{\sigma_{ijk}}\left(\hat{\mathbf{R}}^{(i)} - \mathbf{W}\mathbf{R}^{(i)}\mathbf{W}^T + \mathbf{H}\right)_{jk} \xrightarrow{d} \mathcal{N}(0,1)$$

where $\sigma_{ijk} = O(1)$, $\mathbb{E}[\|\mathbf{H}\|_F] = O(d/\sqrt{m})$ and $\|\mathbf{R}^{(i)}\|_F \to \infty$.

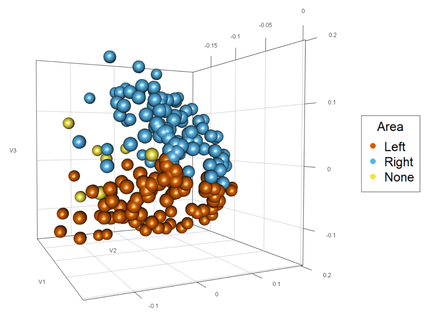
Brain network analysis

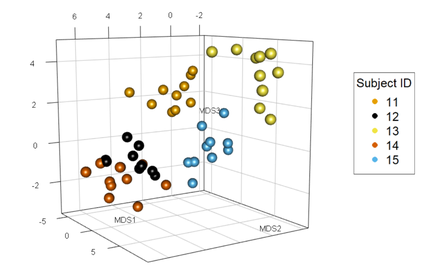
HNU1 data

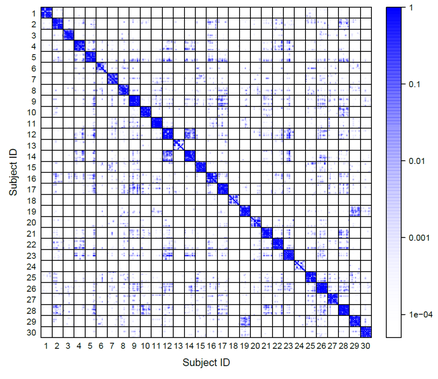
- 300 graphs constructed from diffusion magnetic resonance imaging (dMRI), with $n = 200$ nodes
- 30 different healthy subjects, and 10 graphs per subject
- **Goal:** identify differences and similarities between graphs

Dimensionality reduction

- The vertex embedding $\hat{\mathbf{V}}$ obtained by MASE (top figure) reflects the anatomical location of the vertices in the brain.
- A multidimensional scaling of the distance between the score matrices $\{\mathbf{R}^{(i)}\}_{i=1}^{300}$ (bottom figure) puts graphs from the same subject closer to each other







Matrix of p-values for the equal distribution test of every pair of graphs

Graph hypothesis testing

- For each pair of graphs i and j, we test the hypothesis that their distribution is the same, i.e., $H_0 : \mathbf{R}^{(i)} = \mathbf{R}^{(j)}$
- Test statistic $\|\hat{\mathbf{R}}^{(i)} - \hat{\mathbf{R}}^{(j)}\|_F$
- Semiparametric bootstrap: estimate the parameters with MASE to sample new graphs
- The p-values of the test (left figure) are generally high for pairs of graphs from the same subject (diagonal entries) and low for different subjects

Acknowledgements

This research has been supported by the Lifelong Learning Machines (L2M) program of the Defence Advanced Research Projects Agency (DARPA) via contract number HR0011-18-2-0025. This work is also supported in part by the D3M program of DARPA. We would like to thank Keith Levin and Elizaveta Levina for helpful discussions.

References

[1] Jesús Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E Priebe, and Joshua T Vogelstein, "Inference for multiple heterogeneous networks with a common invariant subspace," *arXiv preprint arXiv:1906.10026*, 2019.

Open R source code for MASE is available at `https://github.com/jesusdaniel/mase`, and in the Python GraSPy package at `https://neurodata.io/graspy`.