

Clustering Multi-Modal Connectomes

Jaewon Chung¹ Benjamin D. Pedigo¹ Carey E. Priebe² Joshua T. Vogelstein¹

¹Department of Biomedical Engineering, Johns Hopkins University

²Department of Applied Mathematics and Statistics, Johns Hopkins University

Summary

- Understanding how brain connectivity is influenced by genetics can improve our understanding of brain function and diseases.
- Existing methods ignore spatial arrangement of the brain or are not statistically justified.
- Our random graph theory and statistical approach enable formulation and testing of different models of connectome heritability.
- We show that brain connectivity is more similar in monozygotic twins compared to those of dizygotic twins and siblings.
- Other frameworks (e.g. joint graph embeddings) can be used for vertex level analysis.

Model: Random Dot Product Graphs (RDPG)

Intuition:

- Each vertex has latent position vector in \mathbb{R}^d
- Probability of edge occurring = dot product of two latent position vectors

Parameters:

- Latent position matrix $X_i \in \mathbb{R}^{n \times d} \forall i \in [m]$

Estimation: Omnibus Embedding

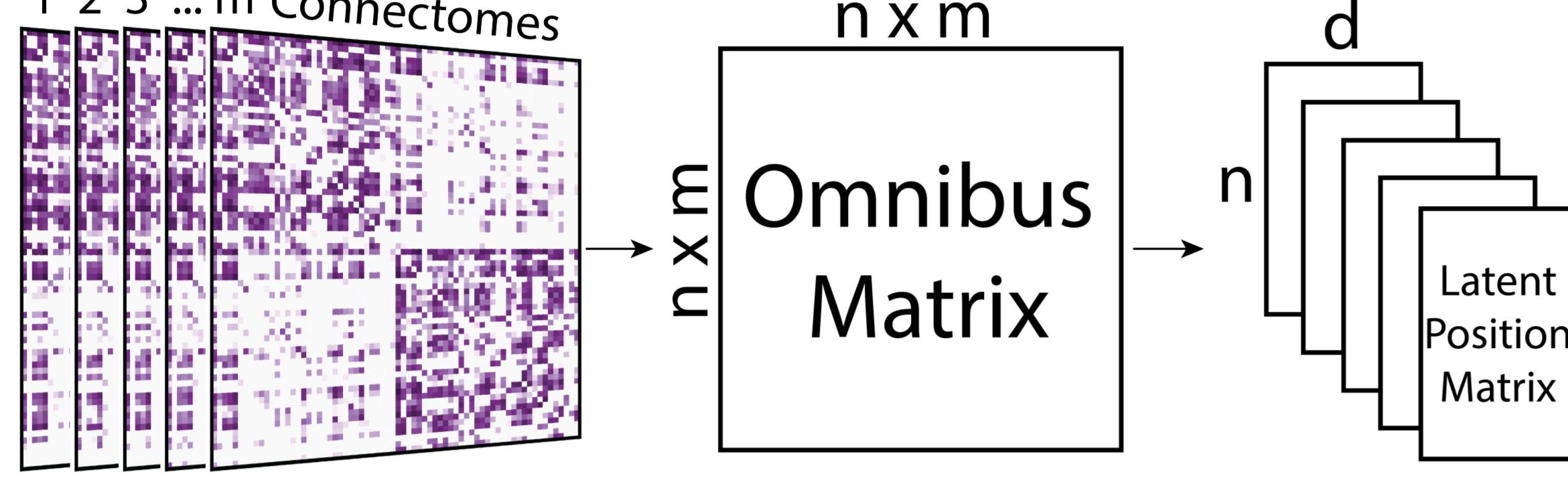


Figure 1. Given multiple connectomes, an omnibus matrix combining all graphs is generated. The omnibus matrix is then decomposed using eigendecomposition to obtain a latent position matrix for each input connectome. Omnibus embedding provides a consistent estimate of latent positions.

Model: Common Subspace Independent Edge Graph (COSIE)

Intuition:

- Each vertex has latent position vector in \mathbb{R}^d .
- All graphs share a common vertex latent position matrix.
- Individual graph is a transformation of latent position matrix.

Parameters:

- Latent position matrix $V \in \mathbb{R}^{n \times d}$
- Individual matrix $R_i \in \mathbb{R}^{d \times d} \forall i \in [m]$

Estimation: Multiple Adjacency Spectral Embedding (MASE)

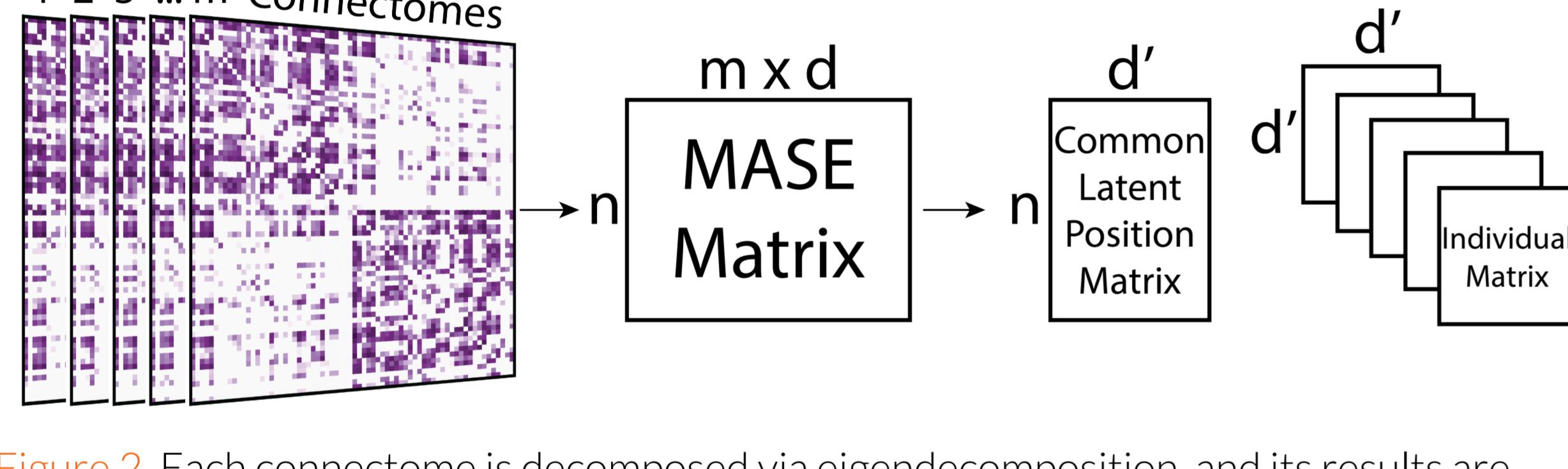


Figure 2. Each connectome is decomposed via eigendecomposition, and its results are concatenated to yield the MASE matrix. The MASE matrix is then decomposed via singular value decomposition to yield a latent position matrix that is shared among all input connectomes and subject matrix that is unique to each connectome. MASE provides a consistent estimate of common latent position matrix, and individual matrices.

Clustering Multi-Modal Connectomes

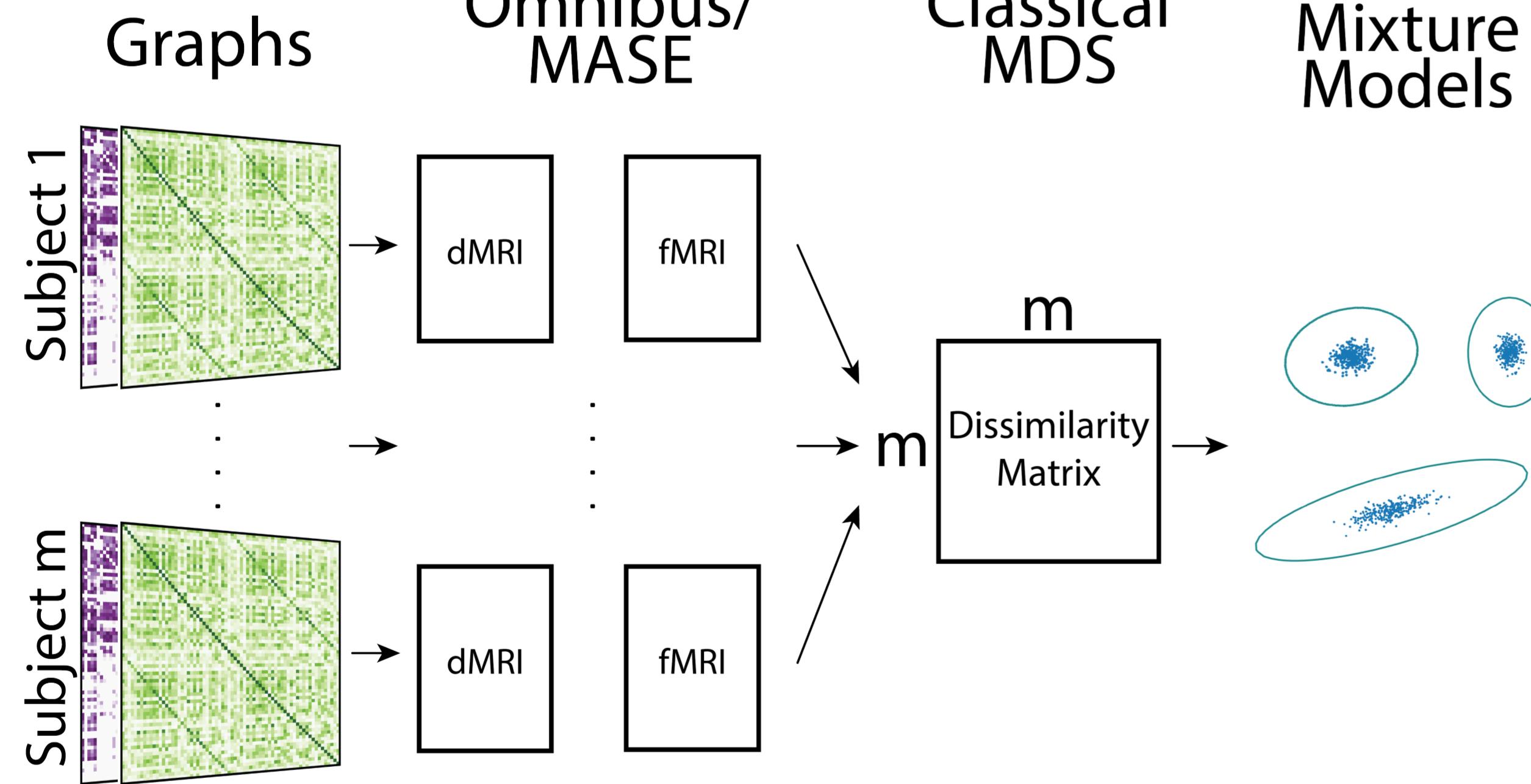


Figure 3. Given populations of graph from different modalities, graphs from each modality is embedded separately. The resulting subject matrices are then concatenated, form a dissimilarity matrix via Euclidean distances, then embedded via classical MDS (cMDS). This results in a feature vector for each subject, which can be clustered via gaussian mixture models.

HNU1 Dataset and Preprocessing

Dataset:

- 30 subjects
- Scanned once every 5 days, 10 total scans per subject
- Preprocessed using ndmg.

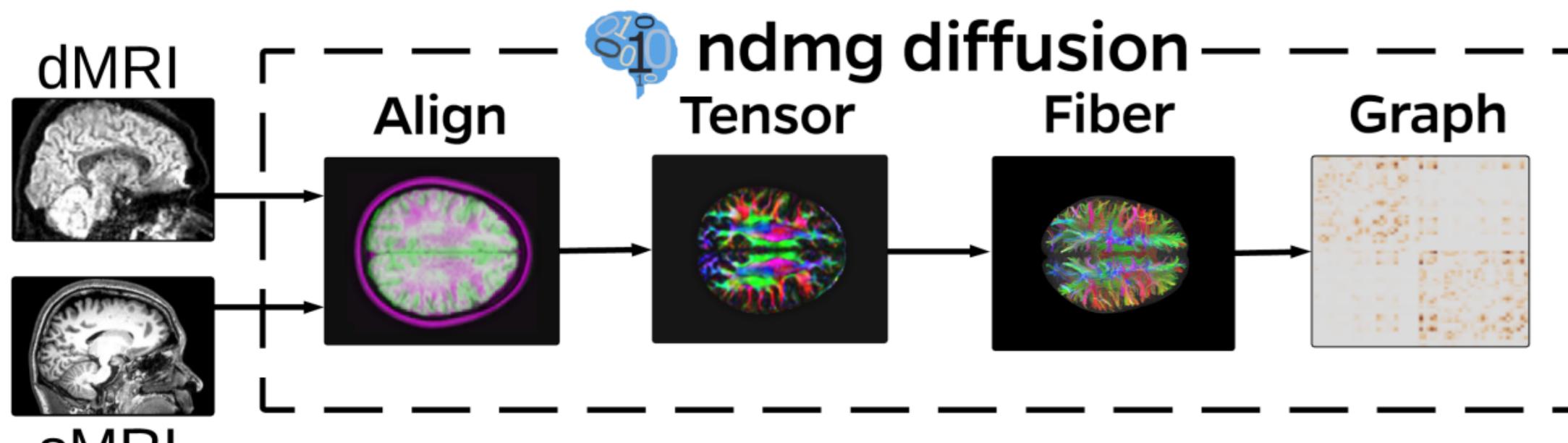
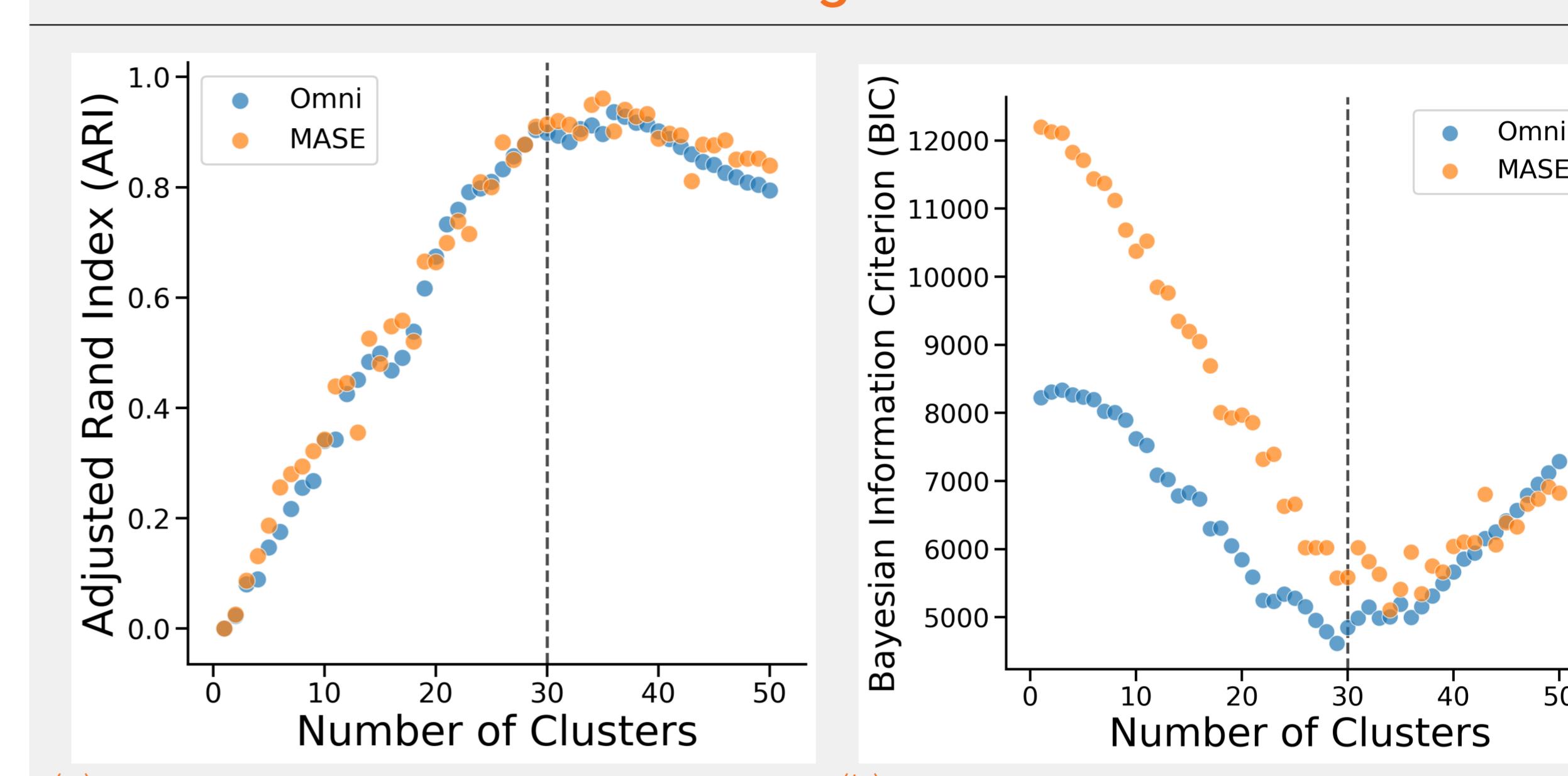


Figure 4. Outline of the ndmg (<https://ndmg.neurodata.io>) pipeline. Image taken from [3].

Summary



(a) Adjusted rand index (ARI) measures the goodness of estimated clusters given ground truth. ARI of 0 means chance clustering, and ARI of 1 means perfect clustering. Dashed vertical line represents 30 clusters. At 30 clusters, ARI is at about 0.9, meaning that the each of the resulting clusters almost perfectly resemble scans from the same subject.
(b) Bayesian information criterion (BIC) measures the goodness of model by penalizing the number of estimated parameters. BIC allows for automated cluster selection by choosing the model with lowest BIC. Dashed vertical line represents 30 clusters. BIC for Omni is minimized at 29 clusters and BIC for MASE is minimized at 34 clusters. Minimized BIC around 30 cluster means that our method forms correct number of clusters, and each cluster is meaningful.

Figure 5. Clustering results from HNU1 dataset using 30 subjects with Omni and MASE.

Visualizations of Intermediate Steps

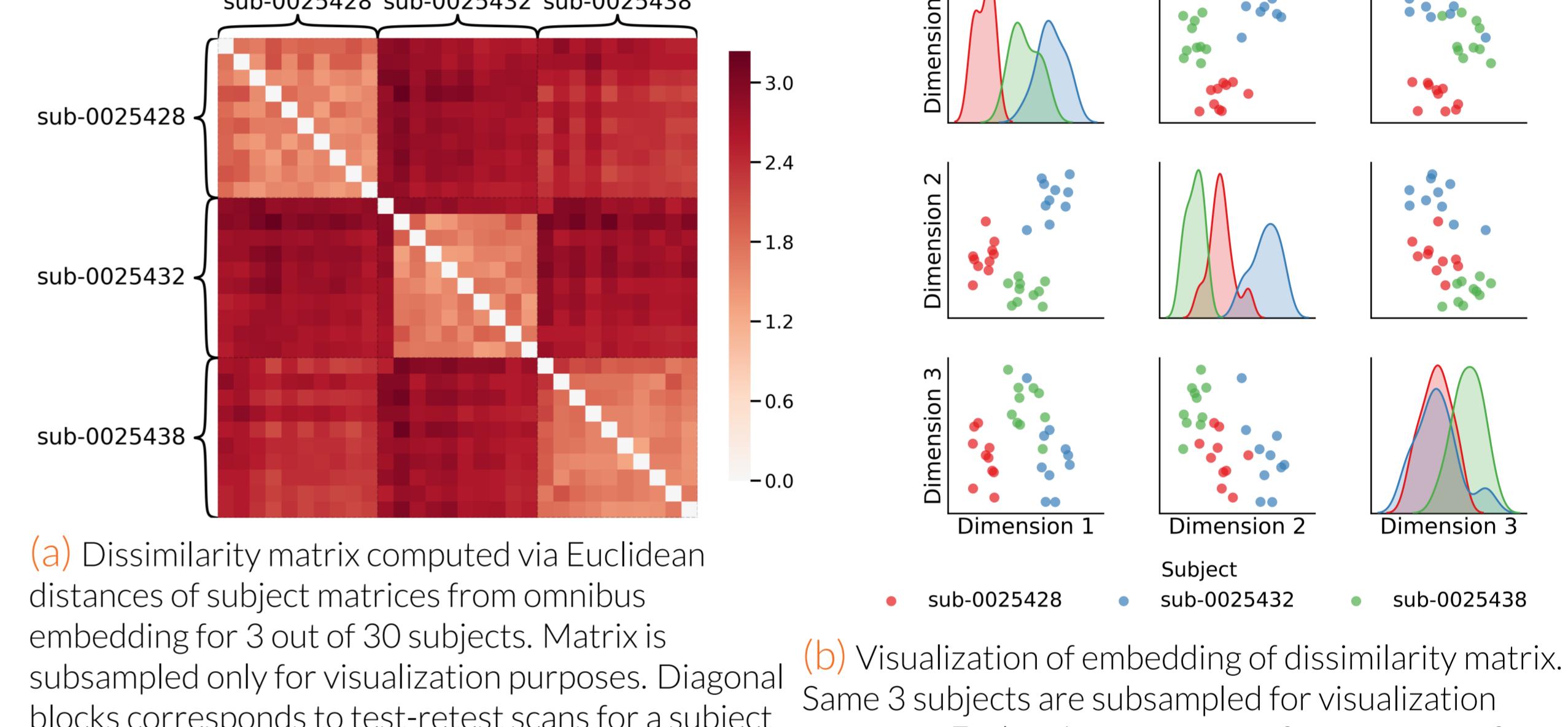


Figure 6. Visualizations of outputs from Classical MDS. The resulting feature vectors from Classical MDS is used for clustering.

Conclusion

- Two new models for representing population of connectomes (RDPG, COSIE).
- Omnibus embedding and multiple adjacency spectral embedding for estimating parameters of RDPG and COSIE, respectively.
- Estimated parameters can be used for any downstream tasks (e.g. hypothesis testing, clustering, classification).
- Clustering on HNU1 data shows near perfect clustering of test-retest scans into correct subject cluster.
- Can be applied to any population of connectomes.

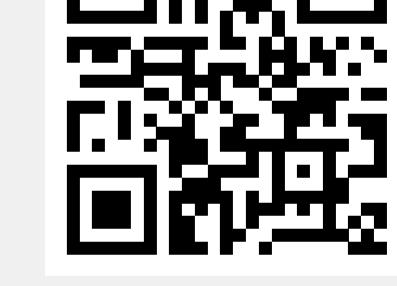
Code and Data

All analysis was performed using an open-source package GrasPy (<https://graspy.neurodata.io>). All dMRI images are open-source and provided by Hangzhou Normal University (http://fcon_1000.projects.nitrc.org/indi/CoRR/html/hnu_1.html).

NeuroData Workshop

NeuroData is hosting a week long workshop for all of our neuroscience and statistical tools.

Date	Tool	Description
8/19	mgc	High dimensional hypothesis testing
8/20	RerF	Decision forest for classification, regression
8/21	GraSpY	Statistical Inference on graphs
8/22	reggie	Non-linear registration
8/23	ndmg	fMRI and dMRI processing



neurodata.io/workshop

Come join us at Baltimore, MD USA!

References

- Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.
- Jaewon Chung, Benjamin D Pedigo, Eric W Bridgeford, Bijan K Varjavand, and Joshua T Vogelstein. Graspy: Graph statistics in python. *arXiv preprint arXiv:1904.05329*, 2019.
- Gregory Kiar, Eric Bridgeford, Will Gray Roncal, Vikram Chandrashekhar, Disa Mhembere, Sephra Ryman, Xi-Nian Zuo, Daniel S Margules, R Cameron Craddock, Carey E Priebe, Rex Jung, Vince Calhoun, Brian Caffo, Randal Burns, Michael P Milham, and Joshua Vogelstein. A high-throughput pipeline identifies robust connectomes but troublesome variability. *bioRxiv*, 2018.
- Keith Levin, Avanti Athreya, Minh Tang, Vince Lyzinski, and Carey E Priebe. A central limit theorem for an omnibus embedding of random dot product graphs. *arXiv preprint arXiv:1705.09355*, 2017.
- Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.
- Xi-Nian Zuo, Jeffrey S Anderson, Pierre Bellec, Rasmus M Birn, Bharat B Biswal, Janusch Blautzik, John CS Breitner, Randy L Buckner, Vince D Calhoun, F Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific data*, 1:140049, 2014.