Simultaneous prediction and community detection for networks with application to neuroimaging

Jesús Arroyo

September 29th, 2020

Joint work with Liza Levina

Outline

1. Introduction

2. Classification and prediction with network covariates

3. Simultaneous prediction and community detection with networks
   - Methodology
   - Algorithms for block-structured regularization
   - Theoretical properties
   - Simulations and brain network data

Outline

Networks

- **Graphs** are a popular structure to represent relational data
  - Vertices/nodes represent the units of a system.
  - Edges/links encode interactions between the units.

Networks

- **Graphs** are a popular structure to represent relational data
  - Vertices/nodes represent the units of a system.
  - Edges/links encode interactions between the units.
- Network data appear in multiple fields
  - social networks
  - technological networks
  - protein interactions
  - brain networks
  - ...

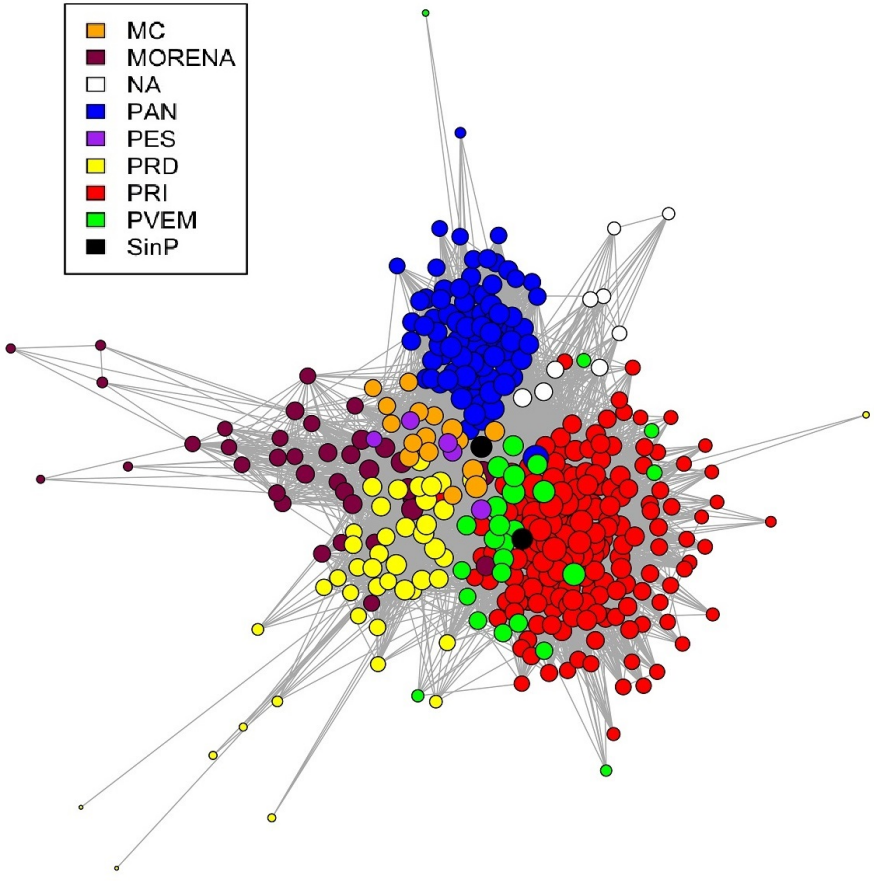
Networks

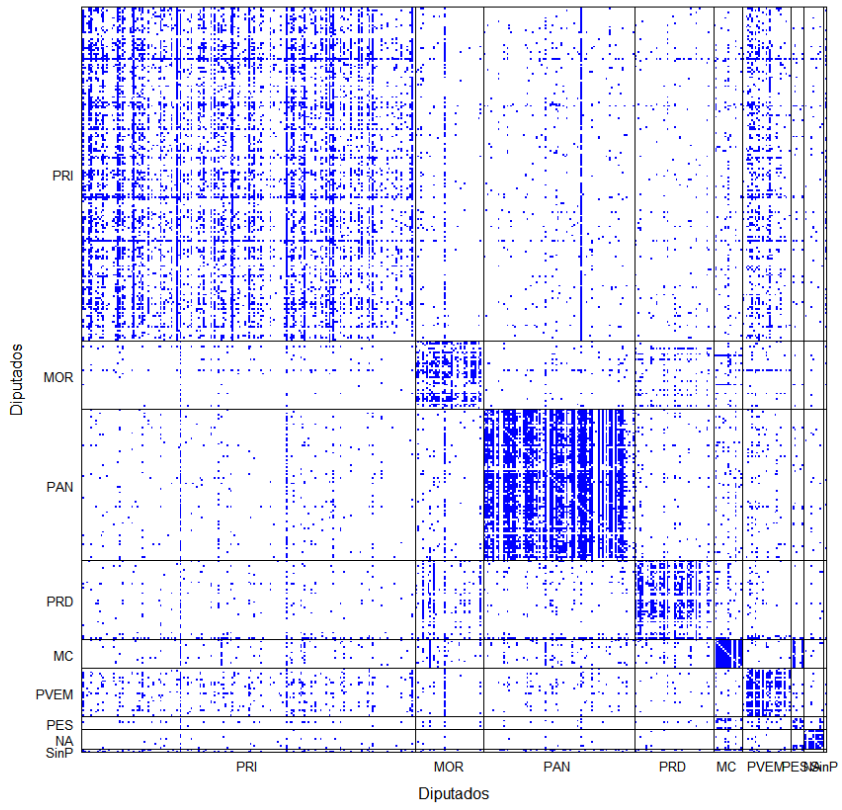
- **Graphs** are a popular structure to represent relational data
  - Vertices/nodes represent the units of a system.
  - Edges/links encode interactions between the units.
- Network data appear in multiple fields
  - social networks
  - technological networks
  - protein interactions
  - brain networks
  - ...
- Inference about the data are conducted at different scales:
  - edges
  - nodes
  - network

Networks

- **Graphs** are a popular structure to represent relational data
  - Vertices/nodes represent the units of a system.
  - Edges/links encode interactions between the units.
- Network data appear in multiple fields
  - social networks
  - technological networks
  - protein interactions
  - brain networks
  - ...
- Inference about the data are conducted at different scales:
  - edges
  - nodes
  - network
  - communities

Mexican Representatives Twitter follow network (2018)

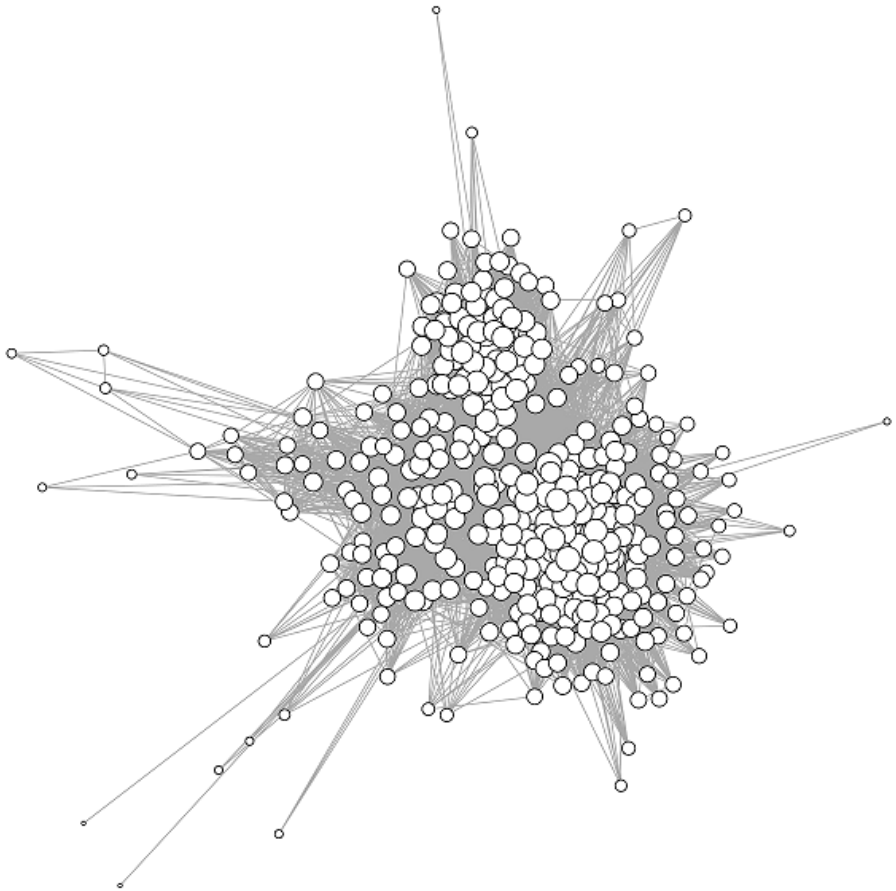




https://github.com/jesusdaniel/spcaCD/

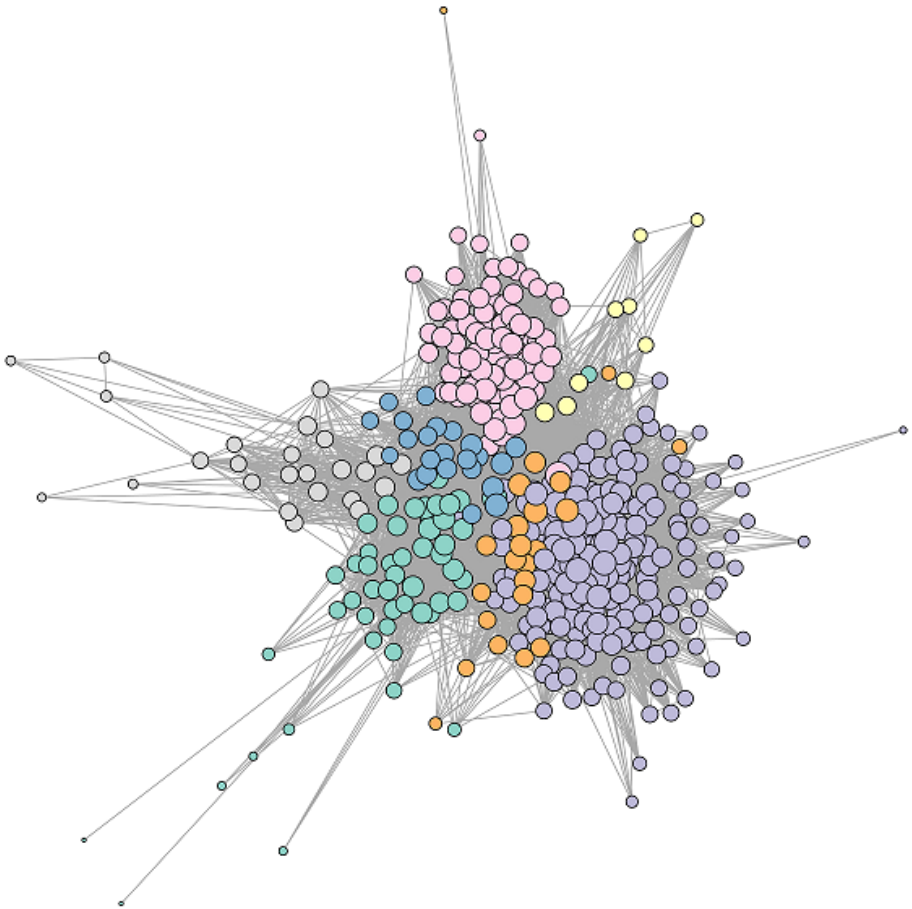
Community detection

Goal: find meaningful groups of nodes (i.e. unsupervised clustering)



Community detection

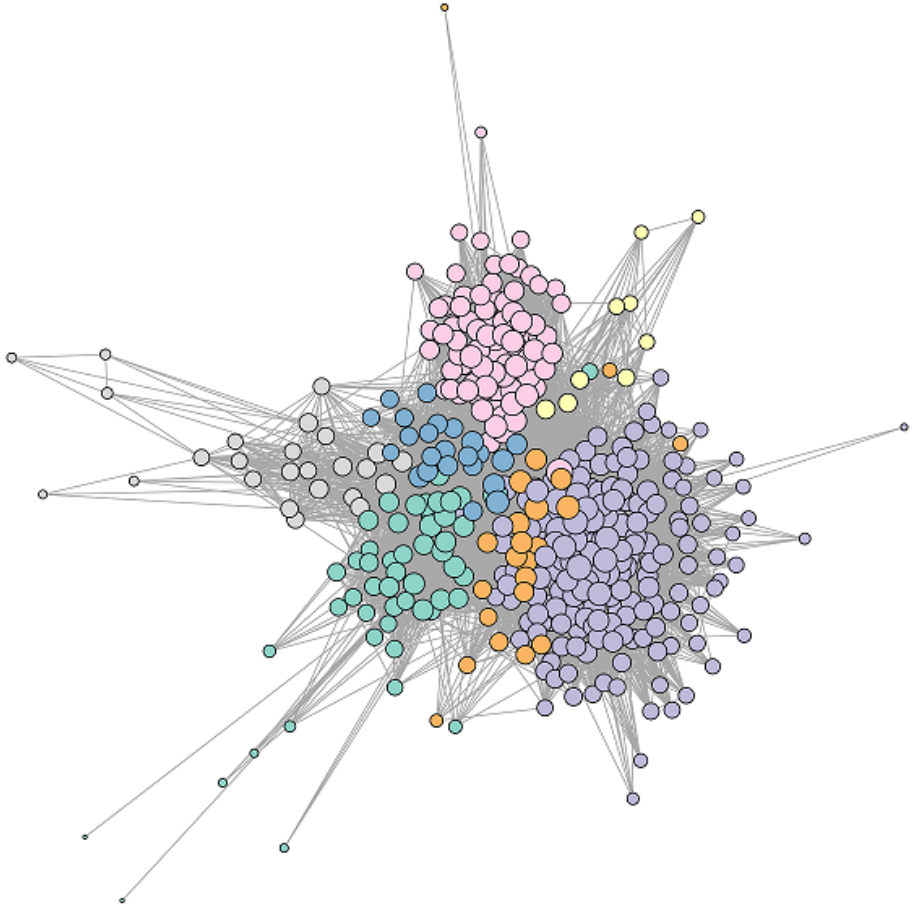
Goal: find meaningful groups of nodes (i.e. unsupervised clustering)

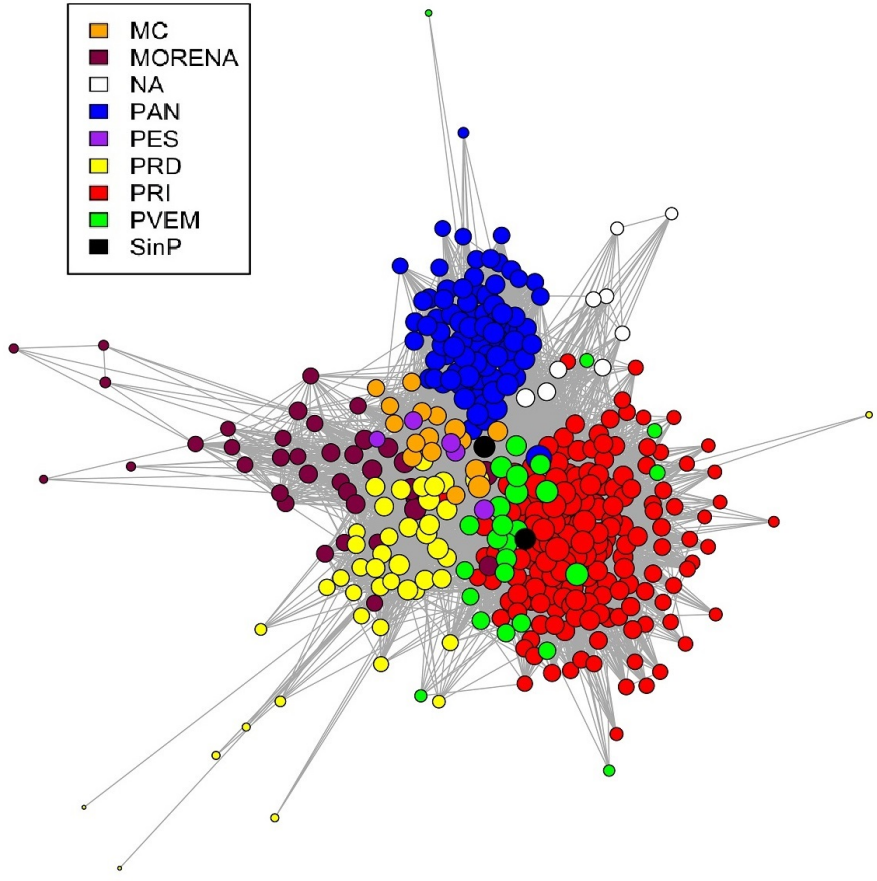


https://arxiv.org/abs/2009.10641

Community detection

Goal: find meaningful groups of nodes (i.e. unsupervised clustering)

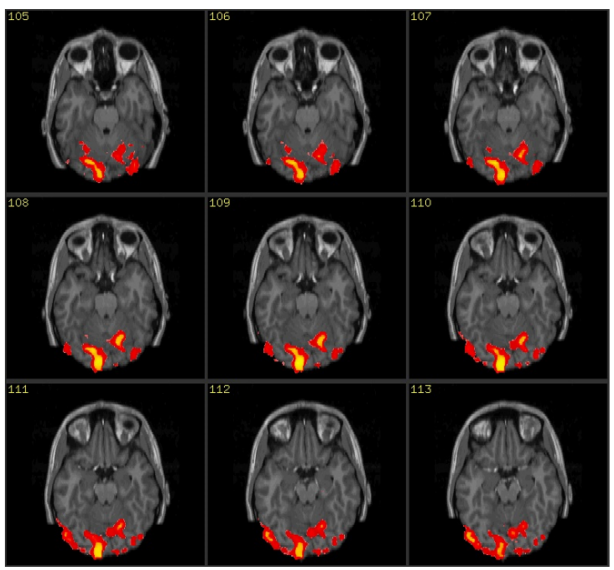


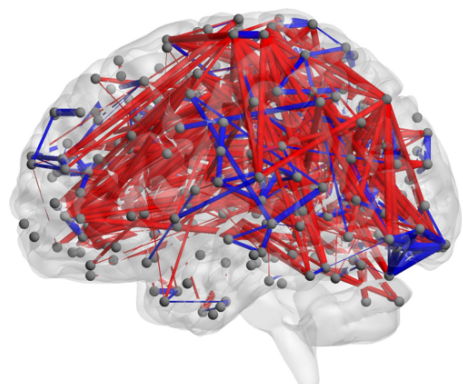


https://arxiv.org/abs/2009.10641

Brain networks

- Subject-specific networks constructed from imaging techniques
- Resting state fMRI data: time series of blood oxygenation level measurements
- Nodes: predefined locations in the brain (regions of interest, ROIs)
- Edges: a functional connectivity measure (Fisher-transformed marginal correlations)
- Pre-processing choices with big downstream effects: noise removal, registration, measures of connectivity





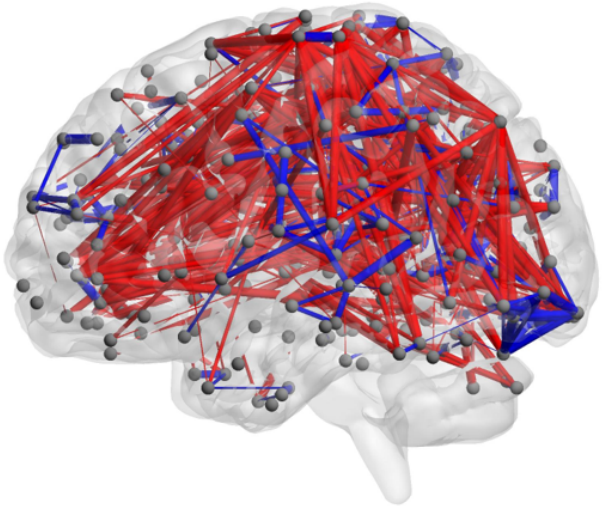
Community structure in the brain

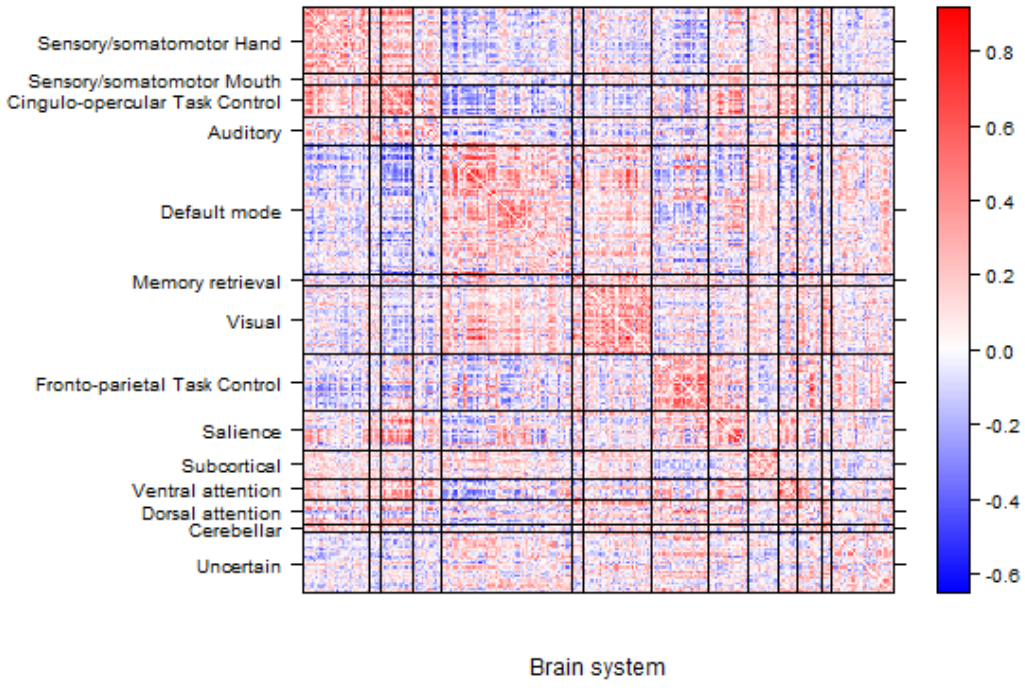
- Community detection algorithms are used to discover meaningful groups of nodes
- Communities correspond to functional brain systems: nodes activate together during tasks and have similar functionality

	Brain system	Nodes (total)
1	Sensory/somatomotor Hand	30
2	Sensory/somatomotor Mouth	5
3	Cingulo-opercular Task Control	14
4	Auditory	13
5	Default mode	58
6	Memory retrieval	5
7	Visual	31
8	Fronto-parietal Task Control	25
9	Salience	18
10	Subcortical	13
11	Ventral attention	9
12	Dorsal attention	11
13	Cerebellar	4
-1	Uncertain	28

Correspondence between communities and functional systems (Power et al., 2011)

Community structure in the brain





Brain network from a subject in the COBRE dataset (Aine et al., 2017)

Community detection: a short summary

- Community detection: unsupervised clustering of the nodes into gropus with similar connectivity patterns
- A well-studied problem in the single-network setting

Community detection: a short summary

- Community detection: unsupervised clustering of the nodes into gropus with similar connectivity patterns
- A well-studied problem in the single-network setting
- Several statistical models proposed to capture different network properties
  - Stochastic block model (Holland et al., 1983) and its extension: degree-corrected SBM, mixed memberships, hierarchical communities, etc.

Community detection: a short summary

- Community detection: unsupervised clustering of the nodes into gropus with similar connectivity patterns
- A well-studied problem in the single-network setting
- Several statistical models proposed to capture different network properties
  - Stochastic block model (Holland et al., 1983) and its extension: degree-corrected SBM, mixed memberships, hierarchical communities, etc.
- Algorithms can be summarized into three types:
  - Likelihood-based approaches
  - Spectral inference
  - Maximization of an objective function (modularity, random-walk objectives)

Community detection: a short summary

- Community detection: unsupervised clustering of the nodes into gropus with similar connectivity patterns
- A well-studied problem in the single-network setting
- Several statistical models proposed to capture different network properties
  - Stochastic block model (Holland et al., 1983) and its extension: degree-corrected SBM, mixed memberships, hierarchical communities, etc.
- Algorithms can be summarized into three types:
  - Likelihood-based approaches
  - Spectral inference
  - Maximization of an objective function (modularity, random-walk objectives)
- In practice, different algorithms might yield different but meaningful communities (Priebe et al., 2019)
- How can we identify meaningful communities for the problem of interest?

Multiple network data

- In many applications, a collection of graphs with matched vertices is observed:
  - Multilayer networks
  - Time-varying networks
  - Multiple samples of networks (e.g. brain networks)

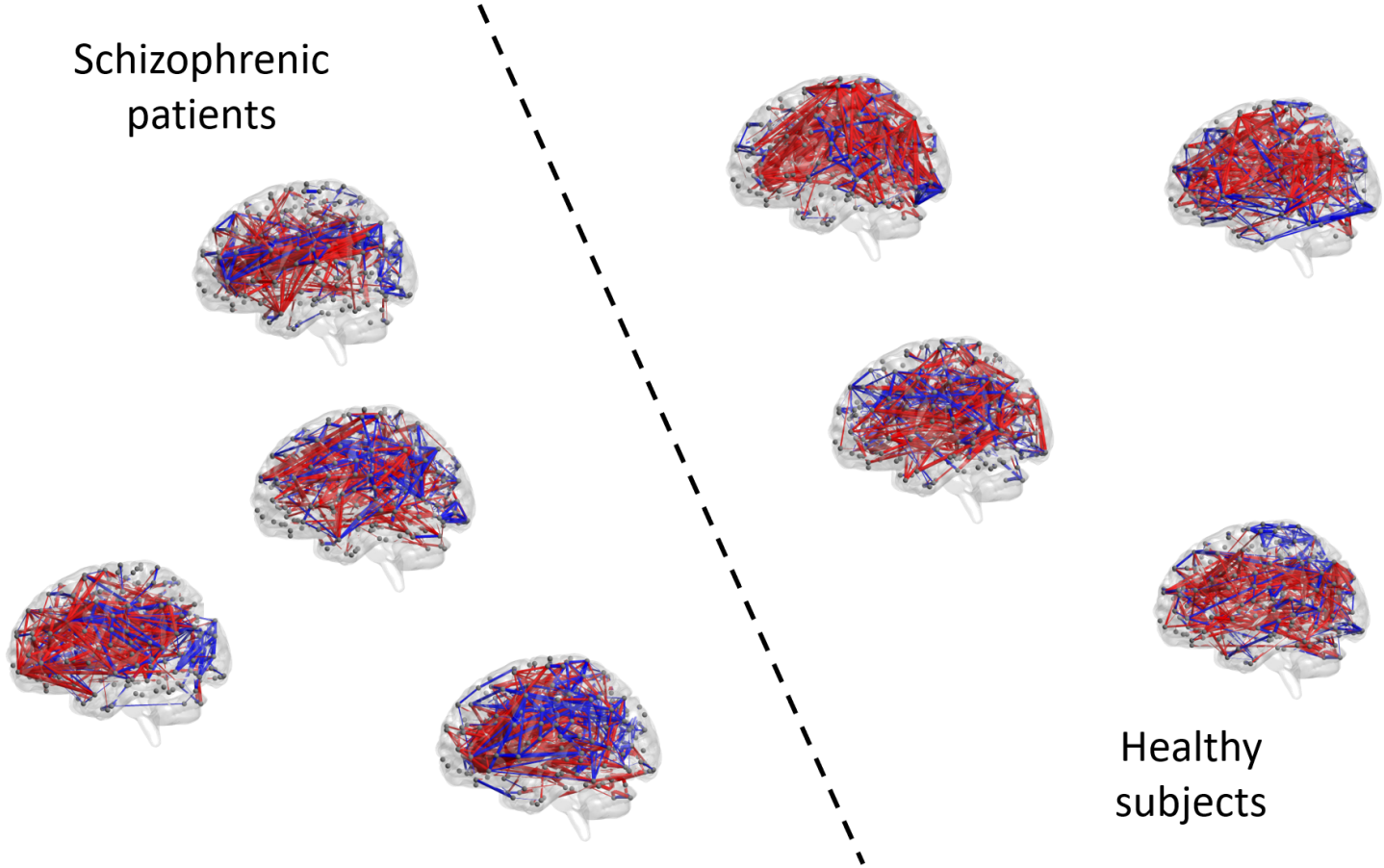
Multiple network data

- In many applications, a collection of graphs with matched vertices is observed:
  - Multilayer networks
  - Time-varying networks
  - Multiple samples of networks (e.g. brain networks)
- This setting is more akin to classical statistics, but observations are network-valued instead of vector-valued data
- How do we incorporate the network structure of the data?

Multiple network data

- In many applications, a collection of graphs with matched vertices is observed:
  - Multilayer networks
  - Time-varying networks
  - Multiple samples of networks (e.g. brain networks)
- This setting is more akin to classical statistics, but observations are network-valued instead of vector-valued data
- How do we incorporate the network structure of the data?
- Often, there are network-specific attributes, and there is interest in understanding the relationship between the connectivity and these attributes

Classification of brain networks



Outline

Setting and notation

Setting: sample of graphs with matched vertices

- n nodes
- m graphs
- $A^{(1)}, \ldots, A^{(m)}$ adjacency matrices (size $n \times n$)
  - Weighted entries ($\mathbb{R}^{n \times n}$)
  - Symmetric
  - Zero-diagonal
- Additional network-specific labels or responses $Y_1, \ldots, Y_m$
  - $Y_i \in \{0, 1\}$ or $Y_i \in \mathbb{R}$

Problem:

- Predict a response Y using the network information A
- Understand the relationship between A and Y

Prediction with network covariates

Two popular approaches to analyzing multiple networks (Bullmore and Sporns, 2009):

1. **Global network summaries** used as features
   - Degree, clustering coefficient, average path length, etc.
   - Captures global network structure but misses local information.

Prediction with network covariates

Two popular approaches to analyzing multiple networks (Bullmore and Sporns, 2009):

1. **Global network summaries** used as features
   - Degree, clustering coefficient, average path length, etc.
   - Captures global network structure but misses local information.
2. **Massive univariate approach**: Use edge weights as features by vectorizing adjacency matrices
   - Standard statistical tools can be applied
   - Captures local information but ignores network structure.

Prediction with network covariates

Two popular approaches to analyzing multiple networks (Bullmore and Sporns, 2009):

1. **Global network summaries** used as features
   - Degree, clustering coefficient, average path length, etc.
   - Captures global network structure but misses local information.
2. **Massive univariate approach**: Use edge weights as features by vectorizing adjacency matrices
   - Standard statistical tools can be applied
   - Captures local information but ignores network structure.

Our goal: develop statistical tools for jointly analyzing samples of networks

- Use local information but incorporate network structure.
- Parsimonious and interpretable solutions.
- Efficient computation and theoretical guarantees.

Methodology

- Linear prediction model for the response (e.g. GLM):

$$\mathbb{E}[Y_i|A^{(i)}] = f(\langle A^{(i)}, B\rangle) = f(\mathsf{Tr}(B^T A^{(i)}))$$

  - Connectivity matrix $\rightarrow$ a linear combination of edge weights
  - $B \in \mathbb{R}^{n\times n}$ symmetric matrix of coefficients.
  - f link function

Methodology

- Linear prediction model for the response (e.g. GLM):

$$\mathbb{E}[Y_i|A^{(i)}] = f(\langle A^{(i)}, B\rangle) = f(\mathsf{Tr}(B^T A^{(i)}))$$

  - Connectivity matrix $\rightarrow$ a linear combination of edge weights
  - $B \in \mathbb{R}^{n\times n}$ symmetric matrix of coefficients.
  - f link function

- In general, B is estimated by minimizing some loss function

$$\ell(B) = \frac{1}{m}\sum_{i=1}^{m} \ell(Y_i, \langle A^{(i)}, B\rangle)$$

  - least squares, logistic function for binary classification, etc.

Methodology

- Linear prediction model for the response (e.g. GLM):

$$\mathbb{E}[Y_i|A^{(i)}] = f(\langle A^{(i)}, B\rangle) = f(\mathsf{Tr}(B^T A^{(i)}))$$

  - Connectivity matrix $\rightarrow$ a linear combination of edge weights
  - $B \in \mathbb{R}^{n\times n}$ symmetric matrix of coefficients.
  - f link function

- In general, B is estimated by minimizing some loss function

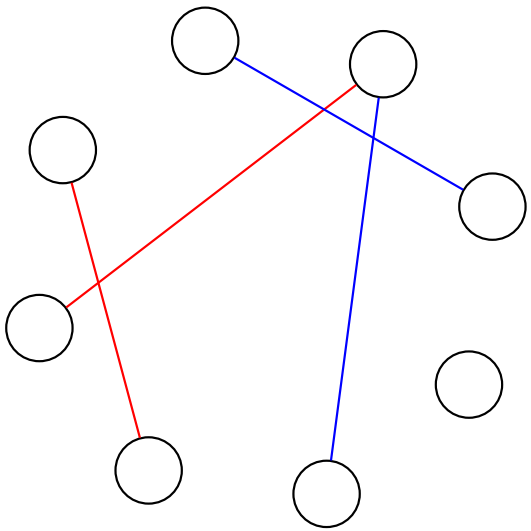
$$\ell(B) = \frac{1}{m}\sum_{i=1}^{m} \ell(Y_i, \langle A^{(i)}, B\rangle)$$

  - least squares, logistic function for binary classification, etc.

- High-dimensional problem: usually $m \ll n^2$.
- Enforce structure and sparsity in the solution through constraints and penalties

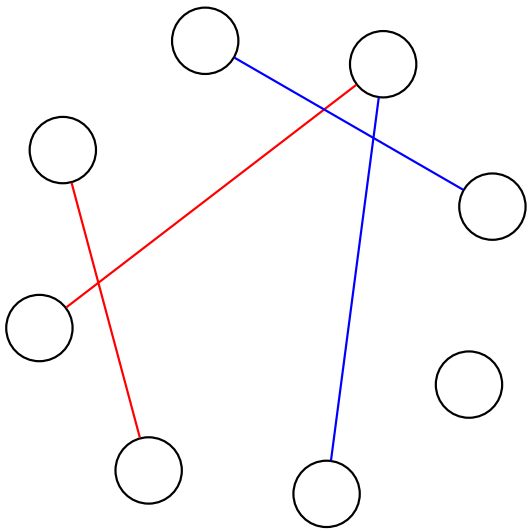
$$\hat{B} = \underset{B\in\mathcal{B}}{\operatorname{argmin}} \{\ell(B) + \Omega_\lambda(B)\}$$

Parsimony in network coefficients (initial approach)

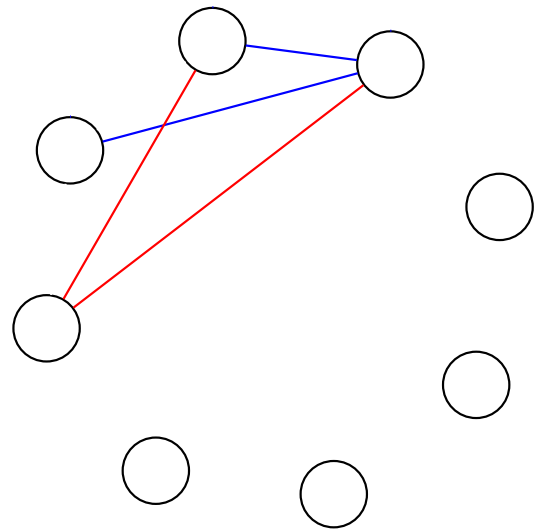


Edge sparsity: small subset of predictive edges

Parsimony in network coefficients (initial approach)

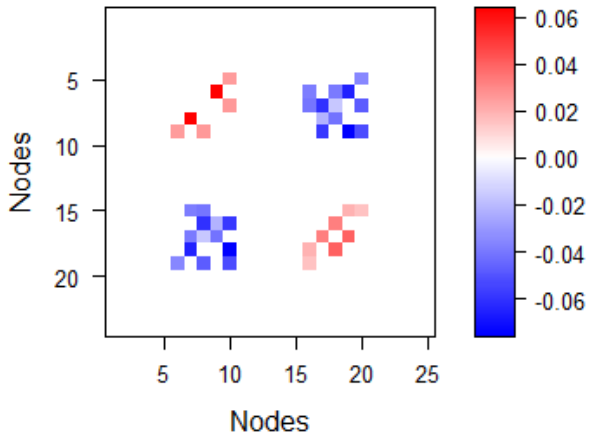


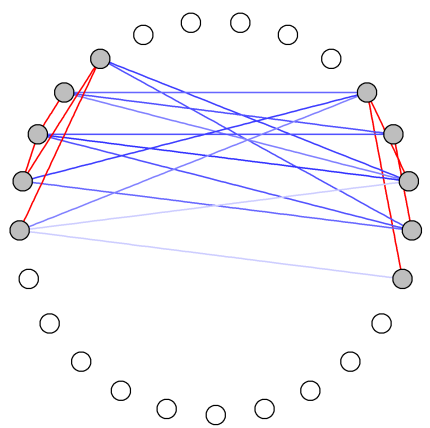
Edge sparsity: small subset of predictive edges



Node sparsity: small subset of nodes with predictive edges

Edge and node selection with overlapping group lasso (Arroyo Relión et al., 2019)

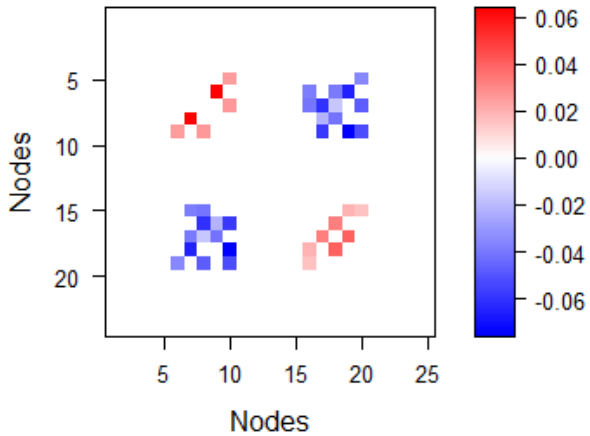




- **Edge selection:** lasso penalty
- Active node: at least one of its edges selected
- **Node selection:** penalize active nodes using group lasso.

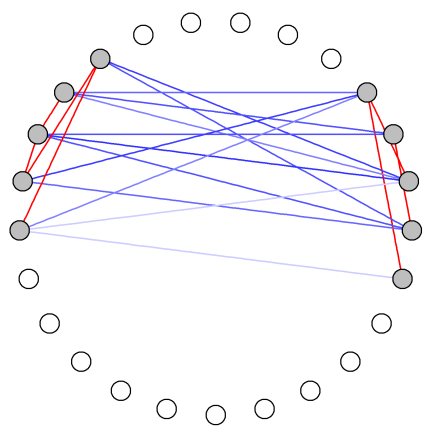
$$\Omega_{\lambda,\rho}(B) = \lambda \left(\sum_{i=1}^{n} \sqrt{B_{i1}^2 + \ldots + B_{in}^2} + \rho \sum_{i,j} |B_{ij}| \right).$$

Edge and node selection with overlapping group lasso (Arroyo Relión et al., 2019)



- **Edge selection:** lasso penalty
- Active node: at least one of its edges selected
- **Node selection:** penalize active nodes using group lasso.

$$\Omega_{\lambda,\rho}(B) = \lambda \left(\sum_{i=1}^{n} \sqrt{B_{i1}^2 + \ldots + B_{in}^2} + \rho \sum_{i,j} |B_{ij}| \right).$$



Proposition

Under regularity conditions, we have

- Consistency: $\|\widehat{B} - B^*\|_F^2 = O_P\left(\frac{G^2 \log N}{n}\right)$
- Node false discovery control: $\hat{\mathcal{G}} \subset \mathcal{G}$ w.h.p.

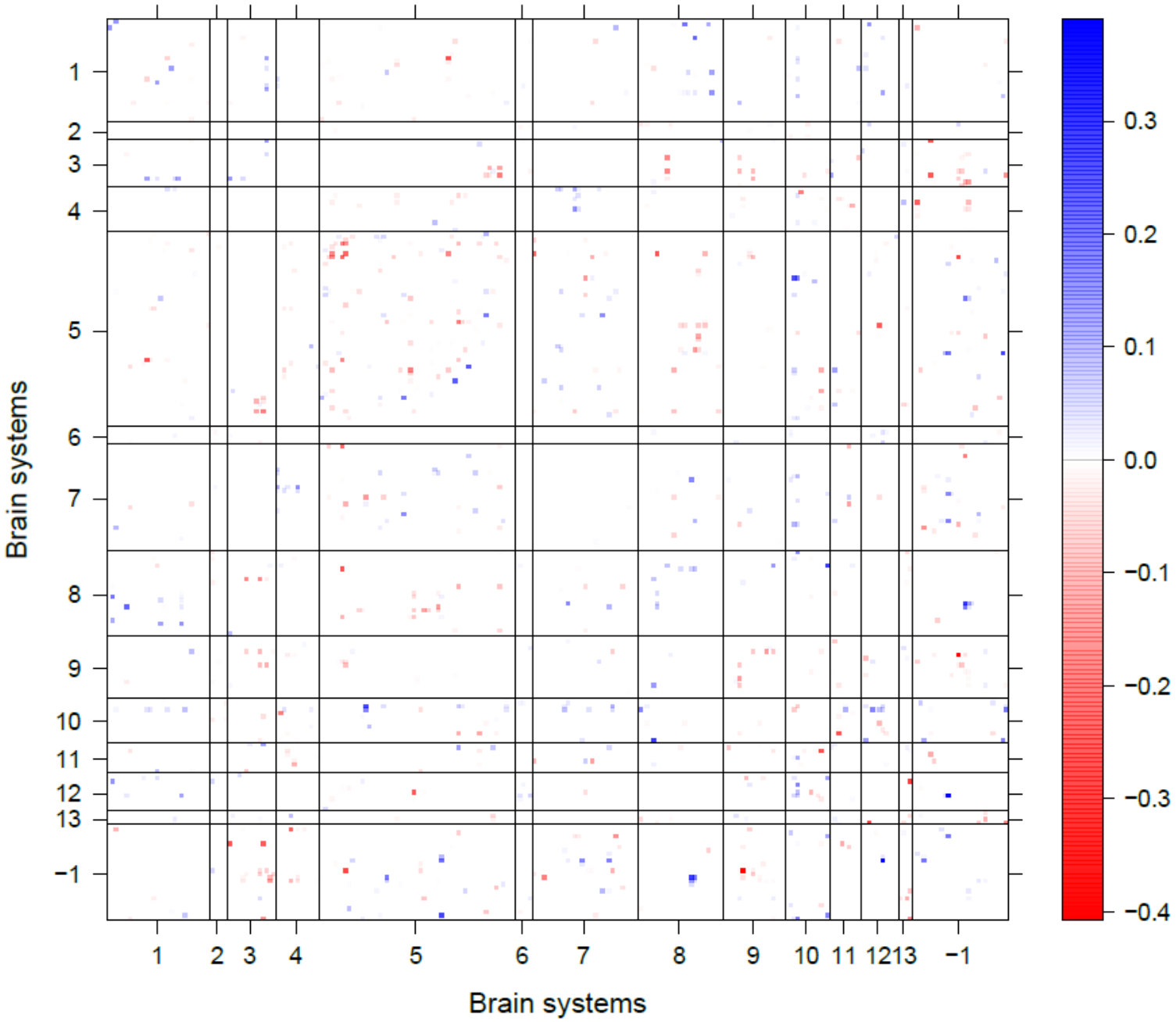
Schizophrenia classification

- COBRE dataset (Aine et al., 2017)
  - 54 schizophrenia patients, 70 healthy subjects ($m = 124$)
  - $m = 263$ regions of interest (nodes) and 34,453 edges
- Good accuracy with a substantially reduced set of edges ($\sim 1,000$).

Method	**% Acc. (s.e.)**
With variable selection	
Our method	92.7 (2.6)
Elastic net	89.5 (1.8)
SVM-L1	87.9 (2.2)
Signal-subgraph	86.1 (3.3)
DLDA	84.6 (3.3)
Lasso	80.1 (5.6)
No variable selection	
SVM	93.5 (2.1)
Ridge penalty	91 (2.6)
Random forest	74.2 (2.6)
Network summaries	64.5 (3.8)

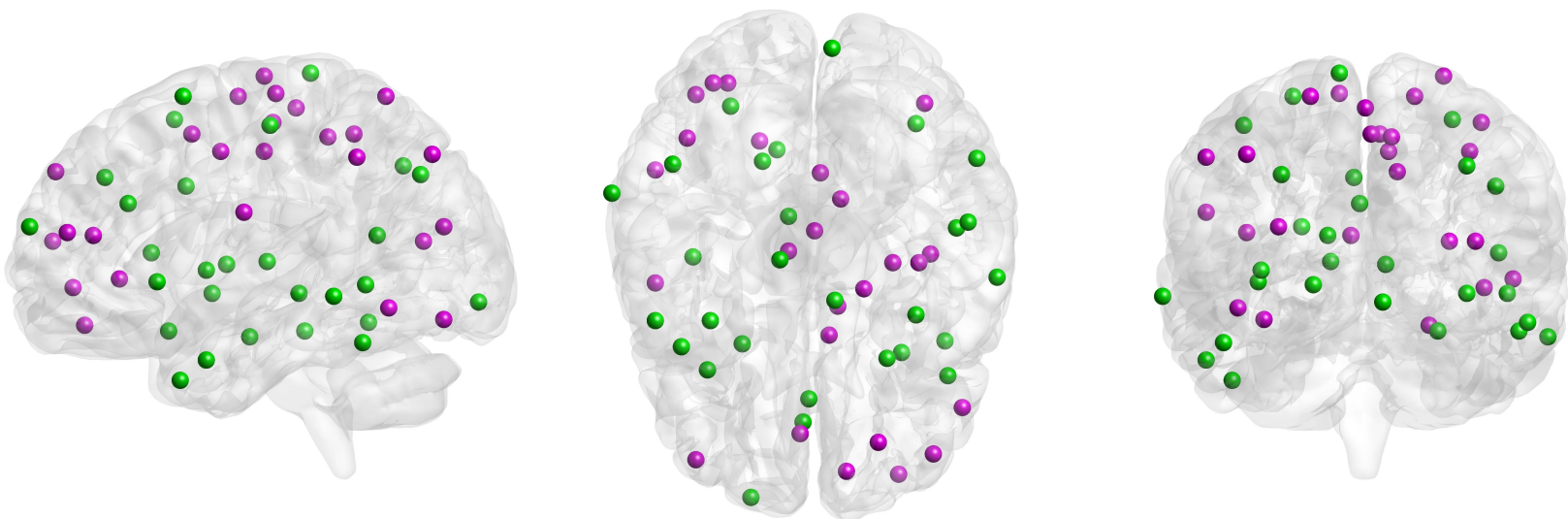
10-fold cross-validation accuracy (standard error)

Matrix of fitted coefficients



Variable selection

- Stability selection (Meinshausen and Buhlman, 2010) to investigate variable selection probabilities.
- Many of the most selected nodes (green) appear in the default-mode network (5) and fronto-parietal task control region (8).
- At least 25 nodes that are consistently not selected for prediction (purple).



Prediction with node selection

- Network-informed penalties improve accuracy and variable selection
- High correlations among edge-variables within a community
- Results are typically interpreted at the community level
- Can we incorporate the community structure in the regularization?

Outline

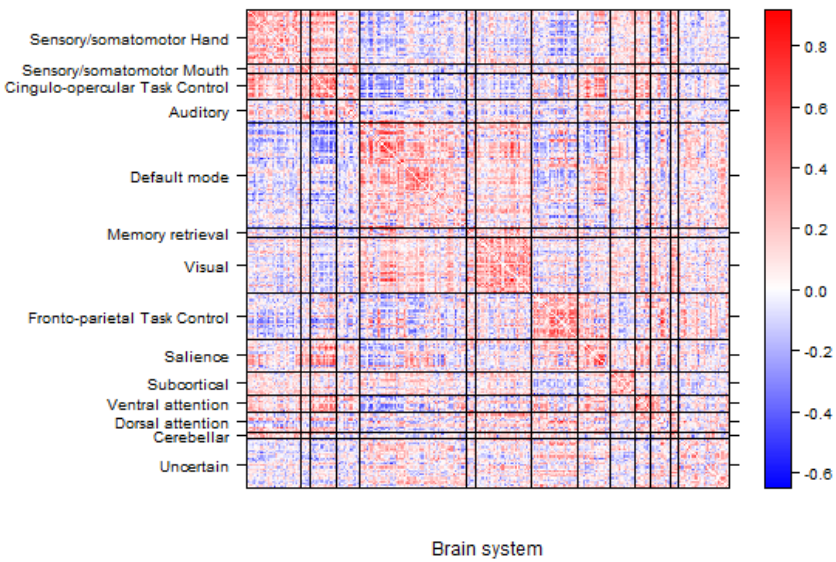
1. Introduction

2. Classification and prediction with network covariates

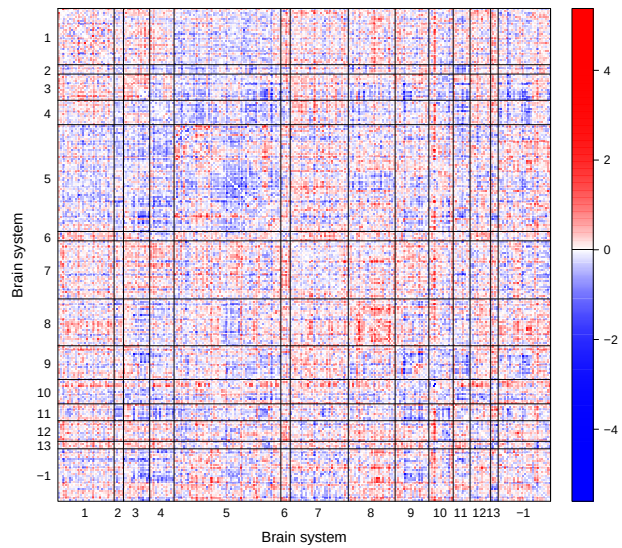
3. Simultaneous prediction and community detection with networks
   - Methodology
   - Algorithms for block-structured regularization
   - Theoretical properties
   - Simulations and brain network data

Community structure in brain networks

- Brain systems are usually derived from healthy subjects.
- They may not fully capture differences between conditions.



A brain network
from a healthy subject



t-statistics for schizophrenic
vs. healthy subjects

Supervised community detection

- Neuroscientific evidence of association between subject-level phenotypes (age, gender, mental illnes) and brain organization (Meunier et al., 2009, Sripada et al. 2014, Kessler et al. 2016)
- Reconfiguration of the brain according to task-induced states (Salehi et al., 2019)
- More generally, clustering is ill-defined, and is usually guided by the problem of interest

Supervised community detection

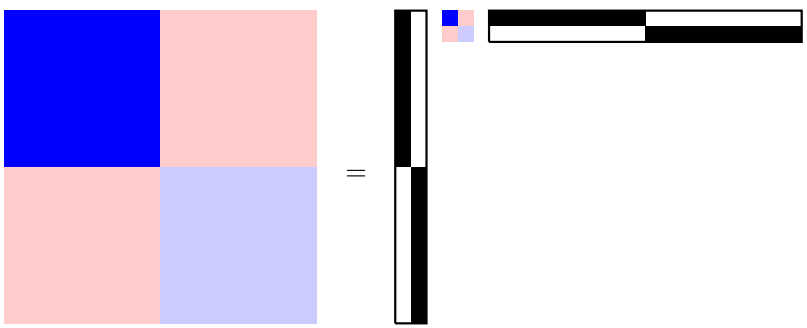
- Neuroscientific evidence of association between subject-level phenotypes (age, gender, mental illnes) and brain organization (Meunier et al., 2009, Sripada et al. 2014, Kessler et al. 2016)
- Reconfiguration of the brain according to task-induced states (Salehi et al., 2019)
- More generally, clustering is ill-defined, and is usually guided by the problem of interest

Our goals:

- Learn community structure and prediction rule simultaneously
- Find network cells with the best predictive power for a specific response of interest

Block structure in coefficients

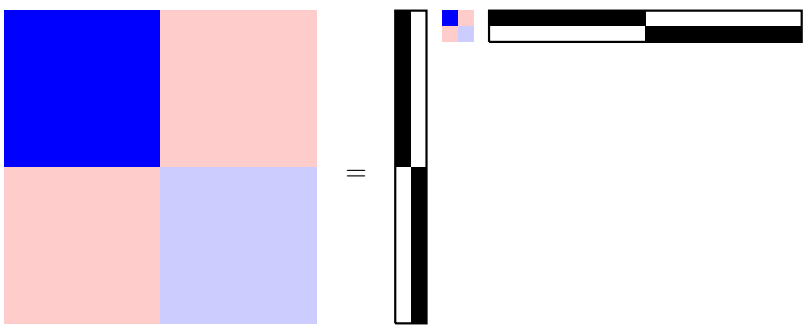
- Enforce community structure in the coefficients B:
  - Partition the nodes into K groups (communities)
  - Constraint $B = ZCZ^\top$ b



- Community membership matrix: $Z \in \{0,1\}^{n \times K}$, $\sum_{k=1}^{K} Z_{jk} = 1$
- Coefficients: $C \in \mathbb{R}^{K \times K}$
- Analogous to the stochastic block model (SBM)

Block structure in coefficients

- Enforce community structure in the coefficients B:
  - Partition the nodes into K groups (communities)
  - Constraint $B = ZCZ^\top$ b



- Community membership matrix: $Z \in \{0,1\}^{n \times K}$, $\sum_{k=1}^{K} Z_{jk} = 1$
- Coefficients: $C \in \mathbb{R}^{K \times K}$
- Analogous to the stochastic block model (SBM)

- Coefficients are clustered into cells with similar predictive behavior.
- Communities simplify interpretation (from $O(n^2)$ to $O(K^2)$ coefficients)

Optimization problem

- Given a value of $K < N$, optimize

$$\begin{aligned} \min_{Z,C} \quad & \ell(ZCZ^\top) + \Omega_\lambda(ZCZ^\top) \\ \text{subject to} \quad & Z \in \{0,1\}^{n \times K}, \qquad Z\mathbf{1}_K = \mathbf{1}_N \\ & C \in \mathbb{R}^{K \times K} \end{aligned}$$

- Given Z, the problem is usually easy to solve.
- Optimizing over Z is a combinatorial problem.

Optimization problem

- Given a value of $K < N$, optimize

$$\begin{aligned} \min_{Z,C} \quad & \ell(ZCZ^\top) + \Omega_\lambda(ZCZ^\top) \\ \text{subject to} \quad & Z \in \{0,1\}^{n \times K}, \qquad Z\mathbf{1}_K = \mathbf{1}_N \\ & C \in \mathbb{R}^{K \times K} \end{aligned}$$

- Given Z, the problem is usually easy to solve.
- Optimizing over Z is a combinatorial problem.

Strategy:

1. Spectral clustering: provides a good initialization value
2. Apply an iterative optimization algorithm (ADMM)

Spectral clustering

- Assume that edge weights are centered and standardized across the sample

$$\sum_{i=1}^{m} A_{uv}^{(i)} = 0, \qquad \frac{1}{m}\sum_{i=1}^{m} (A_{uv}^{(i)})^2 = 1.$$

- We focus on the least-squares loss function

$$\ell(B) = \frac{1}{2m}\sum_{i=1}^{m} \Big(Y_i - \mathsf{Tr}(B^T A^{(i)})\Big)^2$$

Spectral clustering

- Assume that edge weights are centered and standardized across the sample

$$\sum_{i=1}^{m} A_{uv}^{(i)} = 0, \qquad \frac{1}{m}\sum_{i=1}^{m}(A_{uv}^{(i)})^2 = 1.$$

- We focus on the least-squares loss function

$$\ell(B) = \frac{1}{2m}\sum_{i=1}^{m}\left(Y_i - \mathsf{Tr}(B^T A^{(i)})\right)^2$$

- Define $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}} \in \mathbb{R}^{n\times n}$ as the OLS estimate for uncorrelated predictors

$$(\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}})_{u,v} = \widehat{\mathsf{Cov}}(A_{uv}, Y) = \frac{1}{m}\sum_{i=1}^{m} Y_i A_{uv}^{(i)}.$$

Spectral clustering

- Assume that edge weights are centered and standardized across the sample

$$\sum_{i=1}^{m} A_{uv}^{(i)} = 0, \qquad \frac{1}{m}\sum_{i=1}^{m}(A_{uv}^{(i)})^2 = 1.$$

- We focus on the least-squares loss function

$$\ell(B) = \frac{1}{2m}\sum_{i=1}^{m}\left(Y_i - \mathsf{Tr}(B^T A^{(i)})\right)^2$$

- Define $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}} \in \mathbb{R}^{n\times n}$ as the OLS estimate for uncorrelated predictors

$$(\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}})_{u,v} = \widehat{\mathsf{Cov}}(A_{uv}, Y) = \frac{1}{m}\sum_{i=1}^{m} Y_i A_{uv}^{(i)}.$$

- Apply spectral clustering to $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}$
  1. Compute the K leading eigenvectors of $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}$ ($n \times K$ matrix)
  2. Cluster the rows into K groups to estimate node memberships $\widehat{Z}$

Spectral clustering

- Assume that edge weights are centered and standardized across the sample

$$\sum_{i=1}^{m} A_{uv}^{(i)} = 0, \qquad \frac{1}{m}\sum_{i=1}^{m}(A_{uv}^{(i)})^2 = 1.$$

- We focus on the least-squares loss function

$$\ell(B) = \frac{1}{2m}\sum_{i=1}^{m}\left(Y_i - \mathsf{Tr}(B^T A^{(i)})\right)^2$$

- Define $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}} \in \mathbb{R}^{n\times n}$ as the OLS estimate for uncorrelated predictors

$$(\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}})_{u,v} = \widehat{\mathsf{Cov}}(A_{uv}, Y) = \frac{1}{m}\sum_{i=1}^{m} Y_i A_{uv}^{(i)}.$$

- Apply spectral clustering to $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}$
  1. Compute the K leading eigenvectors of $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}$ ($n \times K$ matrix)
  2. Cluster the rows into K groups to estimate node memberships $\widehat{Z}$
- This methods ignores the correlations between edges in the hope that highly correlated edges have similar weights on $\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}$

When does spectral clustering work?

Assumption 1 Centered and standardized covariates

$$\mathbb{E}\left[\sum_{i=1}^{m} A_{uv}^{(i)}\right] = 0, \qquad \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}(A_{uv}^{(i)})^2\right] = 1$$

Assumption 2 Linear model with block-constant coefficients and sub-Gaussian noise

$$Y_i = \mathsf{Tr}(B^T A^{(i)}) + \sigma\epsilon_i,$$

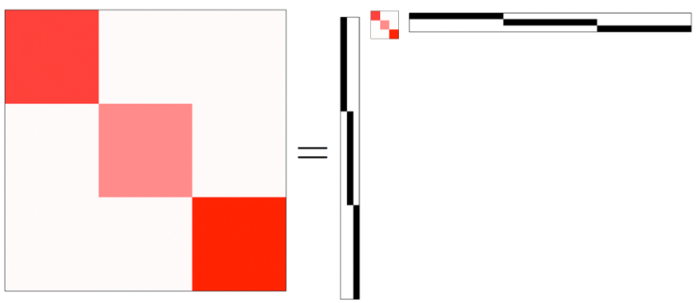
where $B = ZCZ^\top$.

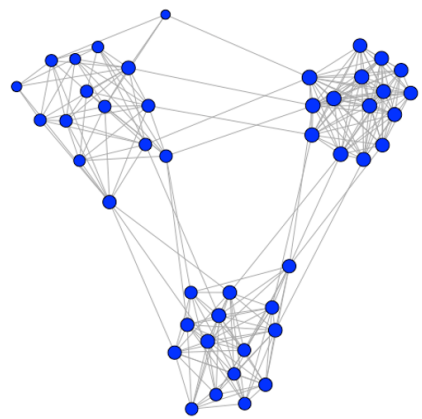
When does spectral clustering work?

Assumption 3 Weighted sub-Gaussian stochastic block model:

- Common membership matrix $Z \in \{0,1\}^{n \times K}$
- Symmetric $K \times K$ matrices $R^{(1)}, \ldots, R^{(m)}$ and $\Psi^{(1)}, \ldots, \Psi^{(m)}$
- $\{A_{uv}^{(m)}\}$ are independent sub-Gaussian r.v. and

$$\mathbb{E}[A^{(i)}] = ZR^{(i)}Z^\top, \qquad \mathsf{Var}(A_{uv}^{(i)}) = (Z\Psi^{(i)}Z^\top)$$



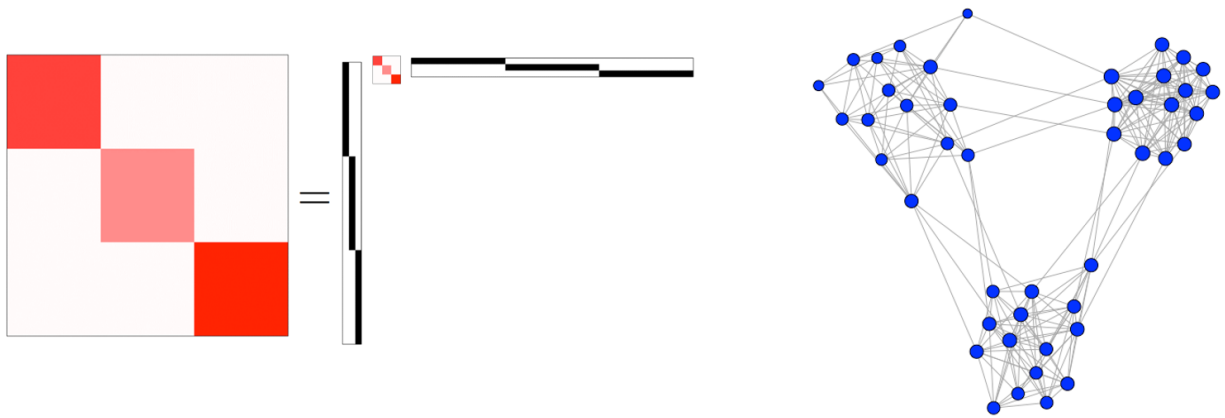


When does spectral clustering work?

Assumption 3 Weighted sub-Gaussian stochastic block model:

- Common membership matrix $Z \in \{0,1\}^{n \times K}$
- Symmetric $K \times K$ matrices $R^{(1)}, \ldots, R^{(m)}$ and $\Psi^{(1)}, \ldots, \Psi^{(m)}$
- $\{A_{uv}^{(m)}\}$ are independent sub-Gaussian r.v. and

$$\mathbb{E}[A^{(i)}] = ZR^{(i)}Z^\top, \qquad \mathsf{Var}(A_{uv}^{(i)}) = (Z\Psi^{(i)}Z^\top)$$



- $\Sigma^{\mathcal{A}}_{[u,v],[s,t]} = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m A_{uv}^{(i)} A_{st}^{(i)}\right]$ is the covariance between edge variables
- Under this model, the covariance between edges only depends on community memberships

$$\Sigma^{\mathcal{A}}_{[u,v],[s,t]} = \frac{1}{m}\sum_{i=1}^m R^{(i)}_{z(u),z(v)} R^{(i)}_{z(s),z(t)}.$$

Consistency of spectral clustering

Theorem

Under regularity conditions, when $n_{min} \asymp n_{max}$*, the estimated membership matrix by spectral clustering satisfies*

$$\mathbb{E}\left[\|\widehat{\mathbf{Z}} - \mathbf{Z}\|_F\right] \lesssim \frac{K^{1/2} n^{3/2} \|B^*\|_F}{m^{1/2} \lambda_K(\mathbb{E}[\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}])} \left(1 + \frac{\sigma}{n\|B^*\|_F}\right)$$

as n *goes to infinity.*

Consistency of spectral clustering

Theorem

Under regularity conditions, when $n_{min} \asymp n_{max}$, the estimated membership matrix by spectral clustering satisfies

$$\mathbb{E}\left[\|\widehat{\mathbf{Z}} - \mathbf{Z}\|_F\right] \lesssim \frac{K^{1/2} n^{3/2} \|B^*\|_F}{m^{1/2} \lambda_K(\mathbb{E}[\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}])} \left(1 + \frac{\sigma}{n\|B^*\|_F}\right)$$

as n goes to infinity.

- When the correlation between edges on different cells ($\Sigma^{\mathcal{A}}_{[u,v],[s,t]}$) is bounded from above, then $\lambda_K(\mathbb{E}[\widehat{\Sigma}^{\mathcal{A},\mathcal{Y}}]) \asymp (n/K)^2 \lambda_{\min}(C^*)$.
- If $\lambda_{\min}(C)$ is bounded away from zero, then

$$\mathbb{E}\left[\|\widehat{\mathbf{Z}} - \mathbf{Z}\|_F\right] \lesssim \frac{K^{5/2} \|B^*\|_F}{m^{1/2} n^{1/2}} \left(1 + \frac{\sigma}{n\|B^*\|_F}\right)$$

Optimization via ADMM

- Introduce new variables and a parameter $\rho > 0$
  - $V \in \mathbb{R}^{n \times n}$
  - $W \in \mathcal{W} = \{ZCZ^T : Z \text{ is a membership matrix}, C \in \mathbb{R}^{K \times K}\}$

Optimization via ADMM

- Introduce new variables and a parameter $\rho > 0$
  - $V \in \mathbb{R}^{n \times n}$
  - $W \in \mathcal{W} = \{ZCZ^T : Z \text{ is a membership matrix}, C \in \mathbb{R}^{K \times K}\}$
- Iterate until convergence

$$B^{(t+1)} = \underset{B}{\operatorname{argmin}} \left\{ \ell(B) + \Omega_\lambda(B) + \frac{\rho}{2} \|B - (W^{(t)} - \frac{1}{\rho} V^{(t)})\|_F^2 \right\} \tag{1}$$

$$W^{(t+1)} = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \left\| W - (B^{(t+1)} + \frac{1}{\rho} V^{(t)}) \right\|_F^2 \tag{2}$$

$$V^{(t+1)} = V^{(t)} + \rho \left(B^{(t+1)} - W^{(t+1)} \right). \tag{3}$$

Optimization via ADMM

- Introduce new variables and a parameter $\rho > 0$
  - $V \in \mathbb{R}^{n \times n}$
  - $W \in \mathcal{W} = \{ZCZ^T : Z \text{ is a membership matrix}, C \in \mathbb{R}^{K \times K}\}$
- Iterate until convergence

$$B^{(t+1)} = \underset{B}{\operatorname{argmin}} \left\{ \ell(B) + \Omega_\lambda(B) + \frac{\rho}{2} \|B - (W^{(t)} - \frac{1}{\rho} V^{(t)})\|_F^2 \right\} \tag{1}$$

$$W^{(t+1)} = \underset{W \in \mathcal{W}}{\operatorname{argmin}} \left\| W - (B^{(t+1)} + \frac{1}{\rho} V^{(t)}) \right\|_F^2 \tag{2}$$

$$V^{(t+1)} = V^{(t)} + \rho \left(B^{(t+1)} - W^{(t+1)} \right). \tag{3}$$

- Problem is non-convex, use spectral clustering for a good initial value.
- Efficient updates
  - $B^{(t+1)}$ is the solution of a convex problem.
  - Solving for $W^{(t+1)}$ requires combinatorial optimization
    - ⋆ We approximate the solution via spectral clustering

Simulations

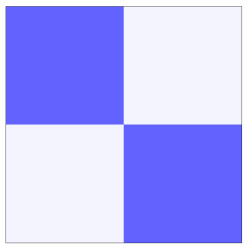
- Parameters that affect complexity of the problem
  - Regression noise
  - Correlation between edge variables
  - Sample size

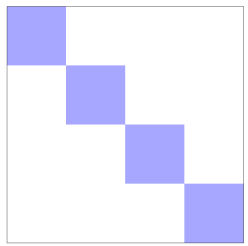
Simulations

- Parameters that affect complexity of the problem
  - Regression noise
  - Correlation between edge variables
  - Sample size
- Generate m weighted networks with $n = 40$ nodes and 4 communities

$$U_i \sim \text{Uniform}(-0.5, 0.5)$$

$$\mathbb{E}[A^{(i)}|U_i] = \quad + tU_i$$

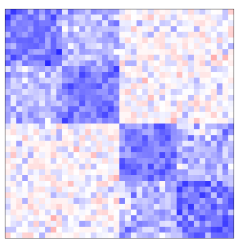


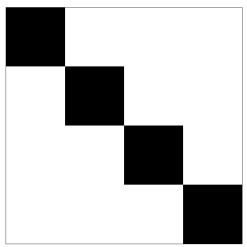


$$A_i = \mathbb{E}[A^{(i)}|U_i] + 0.1 \times Z, \quad Z \sim N(0, 1)$$

- Generate response using the linear regression model

$$Y_i = \left\langle \quad , \quad \right\rangle + \sigma\epsilon_i, \qquad \epsilon_i \sim N(0, 1).$$





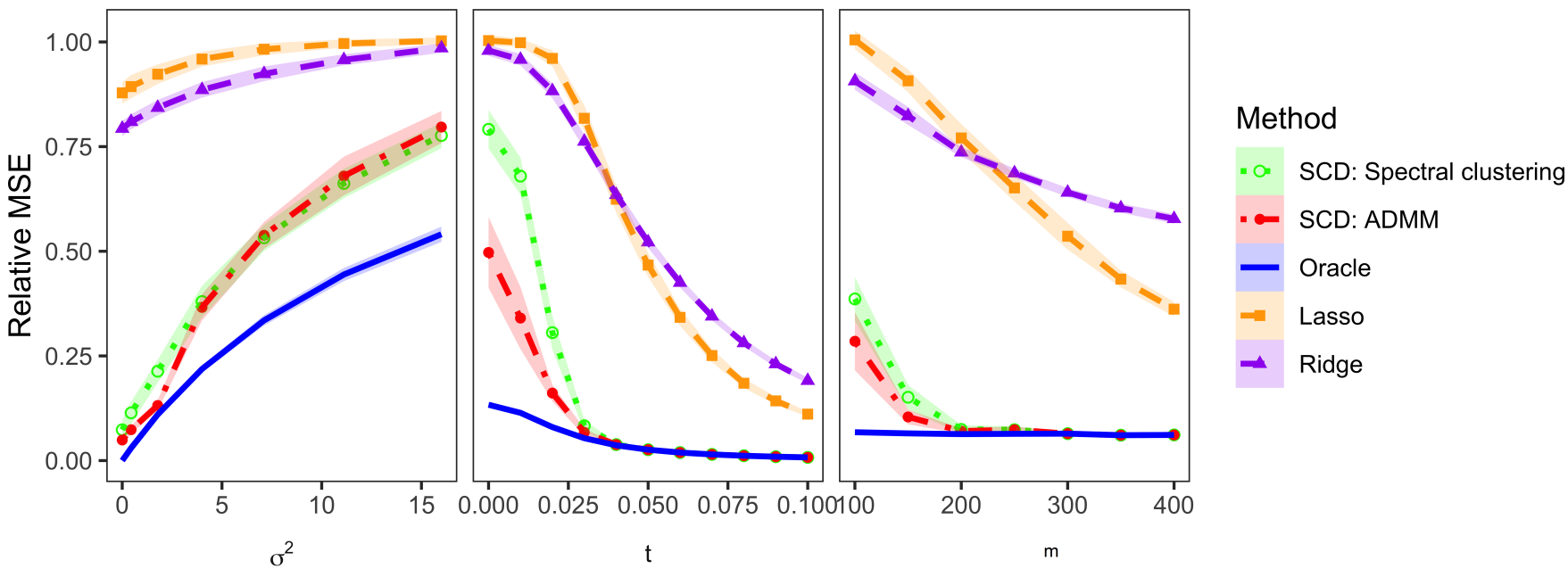
- Set $(\sigma^*, t^*, m^*) = (1, 0.025, 150)$ (low signal) and vary one parameter at a time

Simulations: prediction

- Prediction accuracy measured by relative MSE on a held-out set

$$\text{RMSE} = \frac{\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2}{\hat{\sigma}^2}$$

- Prediction benchmarks:
  - Lasso and ridge regression (no network structure)
  - An oracle that knows true communities

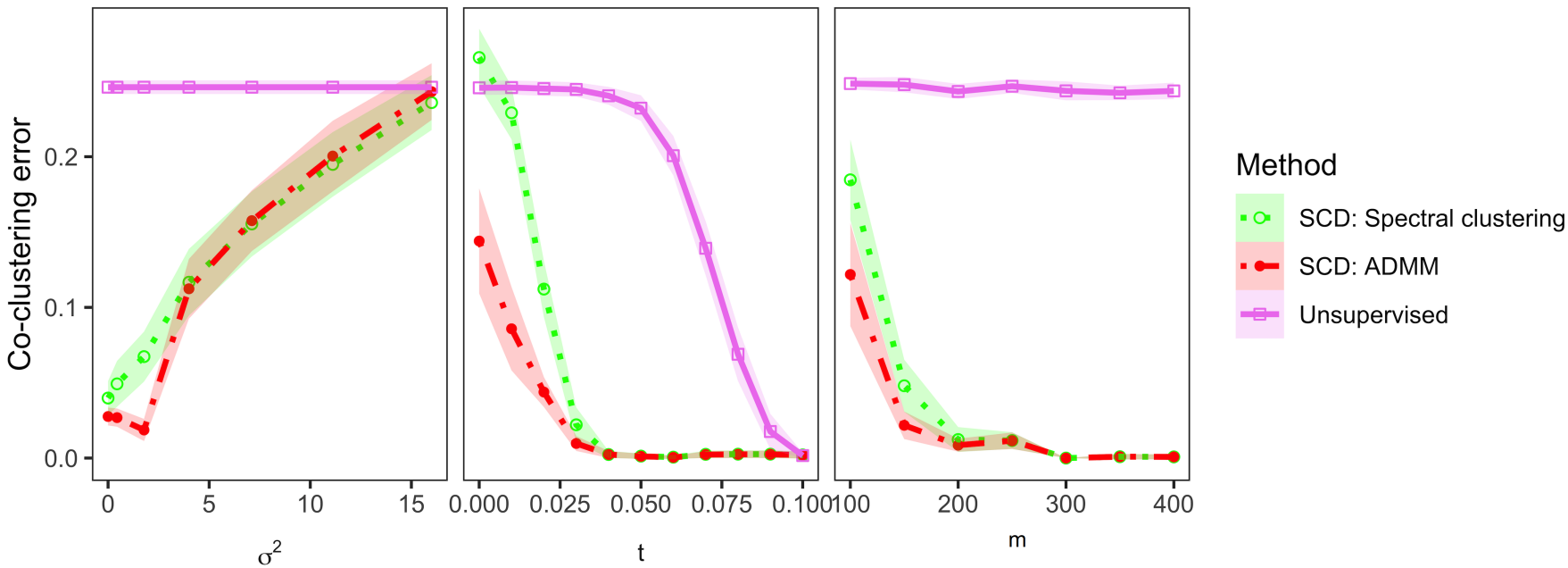


Simulations: community detection

- Community detection accuracy measured by co-clustering error

$$E(Z, \hat{Z}) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} (ZZ^T - \hat{Z}\hat{Z}^T)_{ij}$$

- Community detection benchmark (unsupervised)
  - Spectral clustering for multiple sparse networks (Bhattacharyya and Chatterjee, 2018): spectral clustering on the sum of squared adjacency matrices



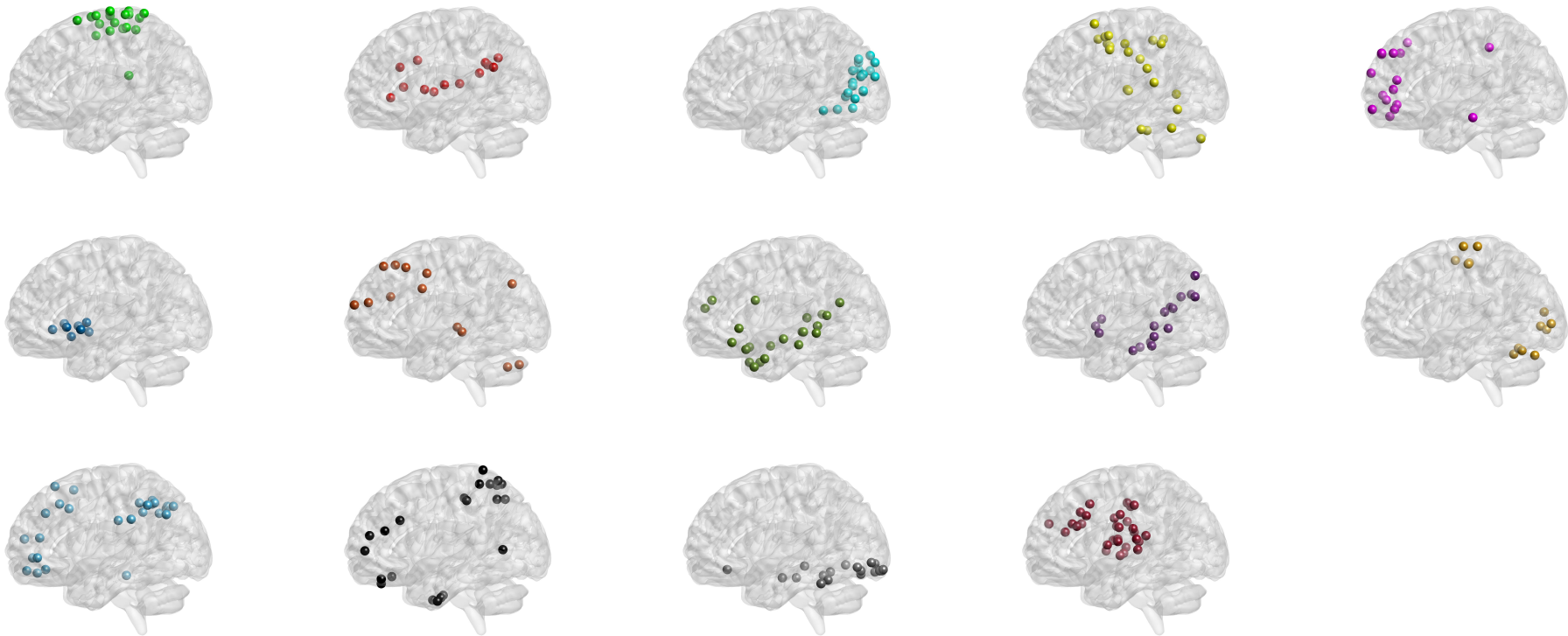
Schizophrenia classification

- Supervised community detection with logistic loss and a lasso penalty
- Accuracy evaluated on COBRE dataset
- Benchmark: Power et al. (2011) communities (14 brain systems)
- Supervised community detection using $K = 14$ communities shows superior performance in accuracy with the same number of parameters.

Power brain systems
CV accuracy: 62%

Supervised community detection
CV accuracy: 79%

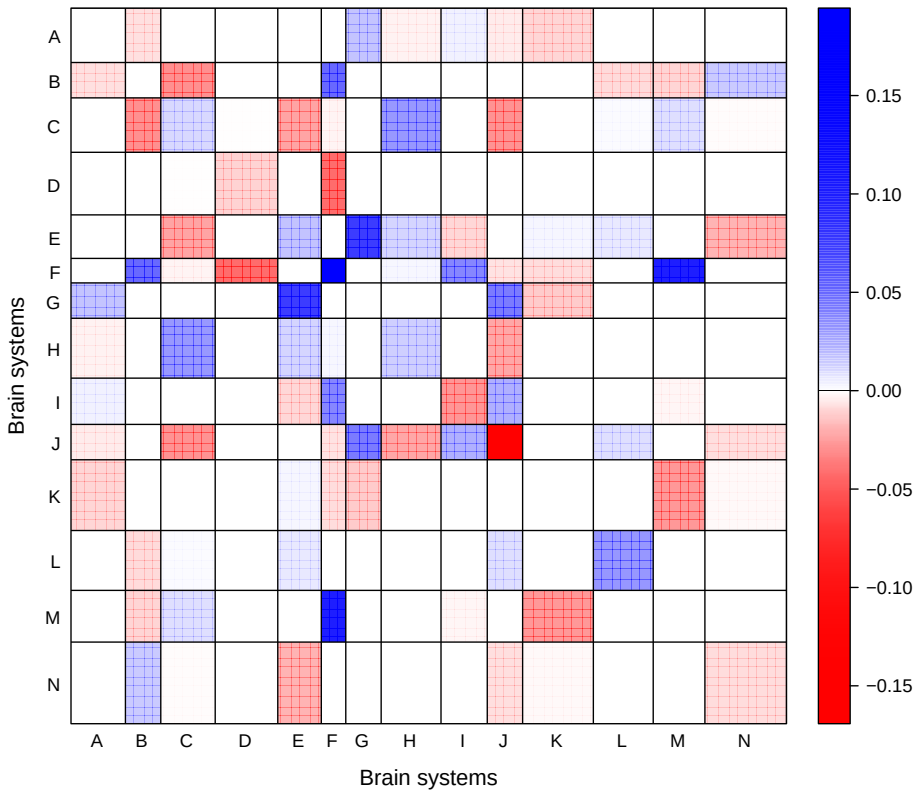
Supervised communities in COBRE data



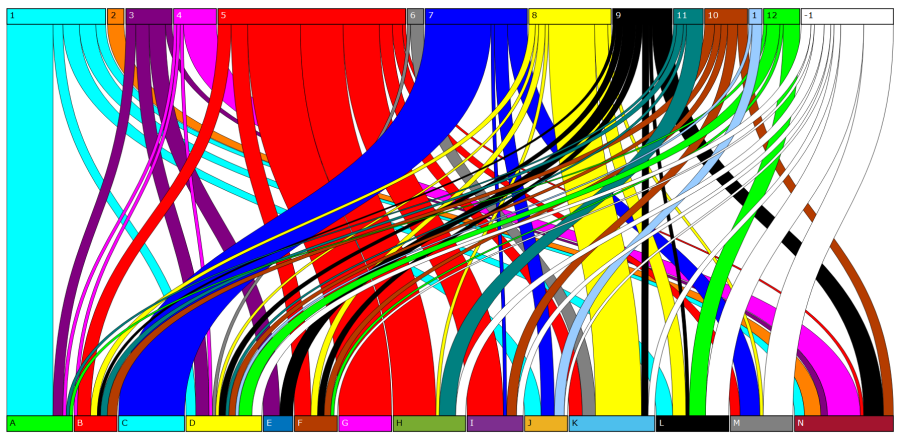
Most communities are spatially contiguous.

Block-structured solution

- SCD+Lasso selects 49 non-zero coefficients (out of 105)
- Default mode network (Power system 5) has been split into three (G, H, I), with different signs



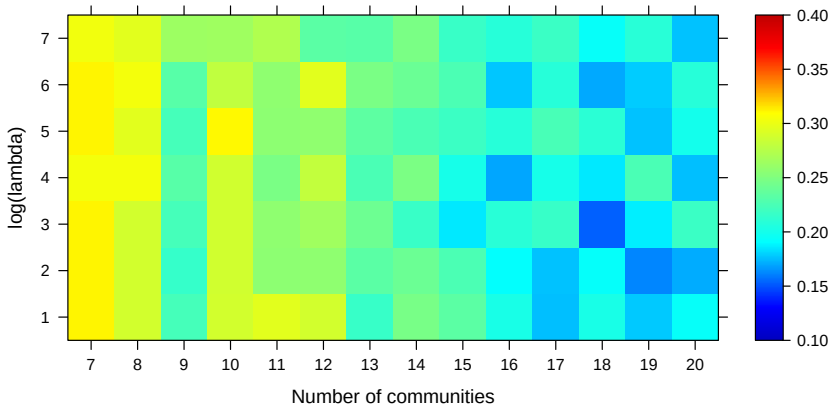
SCD: fitted coefficients



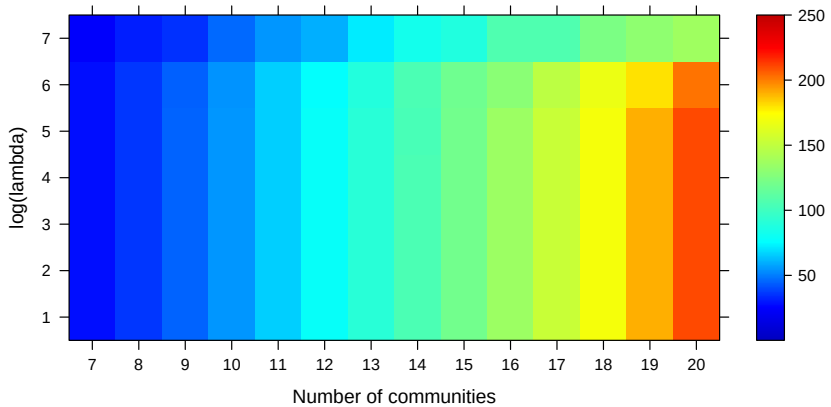
Sankey diagram:
Power (top) vs SCD (bottom)

Cross-validated prediction errors

Cross-validation error



Number of non-zero coefficients



Summary

- A new regularization for prediction problems with network-valued covariates
- Supervised communities simplify interpretation and provide more insights about the specific outcome variable
- Future work:
  - Extensions of the regularization to problems of interest
  - Valid statistical inference
  - Task-dependent brain regions in other relevant scenarios

Summary

- A new regularization for prediction problems with network-valued covariates
- Supervised communities simplify interpretation and provide more insights about the specific outcome variable
- Future work:
  - Extensions of the regularization to problems of interest
  - Valid statistical inference
  - Task-dependent brain regions in other relevant scenarios

Thank you!

Arroyo, J., Levina, E. (2020). *"Simultaneous prediction and community detection for networks with application to neuroimaging"*, arXiv:2002.01645.

References I

- Aine, C. J. et al. (Oct. 2017). "Multimodal Neuroimaging in Schizophrenia: Description and Dissemination". In: *Neuroinformatics* 15.4, pp. 343–364.
- Arroyo Relión, Jesús D. et al. (Jan. 2019). "Network classification with applications to brain connectomics". In: *Annals of Applied Statistics* 13.3, pp. 1648–1677. arXiv: 1701.08140.
- Holland, Paul W., Kathryn Blackmond Laskey, and Samuel Leinhardt (June 1983). "Stochastic blockmodels: First steps". In: *Social Networks* 5.2, pp. 109–137.
- Power, Jonathan D et al. (2011). "Functional network organization of the human brain". In: *Neuron* 72.4, pp. 665–678.
- Priebe, Carey E et al. (2019). "On a two-truths phenomenon in spectral graph clustering". In: *Proceedings of the National Academy of Sciences* 116.13, pp. 5995–6000.