

Multi-State Brain Network Discovery

Hang Yin

Worcester Polytechnic Institute
Worcester, MA, USA
hyin@wpi.edu

Yao Su

Worcester Polytechnic Institute
Worcester, MA, USA
ysu6@wpi.edu

Xinyue Liu

Worcester Polytechnic Institute
Worcester, MA, USA
xliu4@wpi.edu

Thomas Hartvigsen

University of Virginia
Charlottesville, VA, USA
hartvigsen@virginia.edu

Yanhua Li

Worcester Polytechnic Institute
Worcester, MA, USA
yli15@wpi.edu

Xiangnan Kong

Worcester Polytechnic Institute
Worcester, MA, USA
xkong@wpi.edu

Abstract—Brain network discovery aims to find nodes and edges from the spatio-temporal signals obtained by neuroimaging data, such as fMRI scans of human brains. Existing methods tend to derive representative or average brain networks, assuming observed signals are generated by only a *single* brain activity state. However, the human brain usually involves *multiple* activity states, which jointly determine the brain activities. The brain regions and their connectivity usually exhibit intricate patterns that are difficult to capture with only a single-state network. Recent studies find that brain parcellation and connectivity change according to the brain activity state. We refer to such brain networks as *multi-state*, and this mixture can help us understand human behavior. Thus, compared to a *single-state* network, a *multi-state* network can prevent us from losing crucial information of cognitive brain network. To achieve this, we propose a new model called MNGL (Multi-state Network Graphical Lasso), which successfully models multi-state brain networks by combining CGL (coherent graphical lasso) with GMM (Gaussian Mixture Model). Using both synthetic and real world ADHD-200 fMRI datasets, we demonstrate that MNGL outperforms recent state-of-the-art alternatives by discovering more explanatory and realistic results.

Index Terms—brain networks, edge detection, graphical lasso, mixture model

I. INTRODUCTION

Motivation. Brain network discovery [1, 2, 3] is one of the most pervasive paradigms in neuroscience and involves two main tasks: *brain parcellation* and *edge detection*. In brain networks, nodes represent brain regions, and edges represent the functional/structural connections between regions. In general, brain networks are modeled by first finding nodes that contain coherently-functioning brain regions (*i.e.*, performing *brain parcellation*), then identifying edges between these nodes according to an observed sequence of brain activity (*i.e.*, *edge detection*). Precise discovery of these networks cultivates a more refined model of the human brain. Such models become instrumental in diagnosing brain disorders [4] and analyze brain functions [5]. Furthermore, recent studies [6, 7, 8] have found that the difference in brain activity states can infer to distinct brain parcellations and connectivity patterns. Thus, an effective brain network discovery methodology must be adaptive, adjusting to the dynamism of the brain’s activity

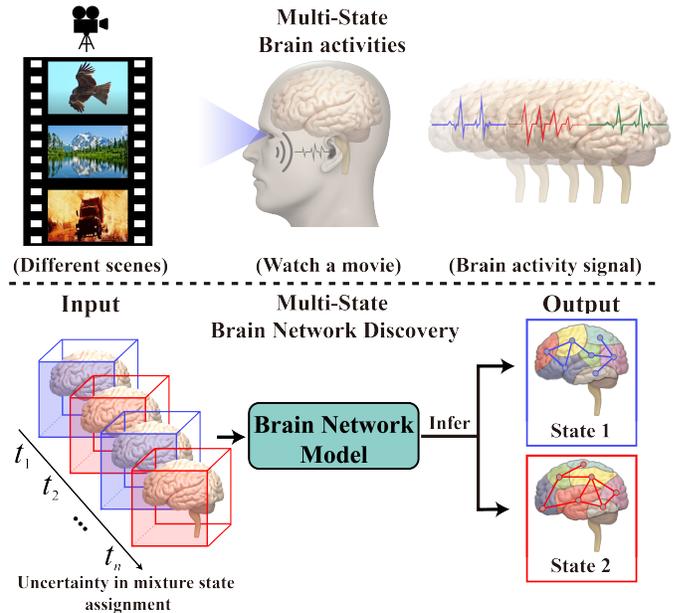
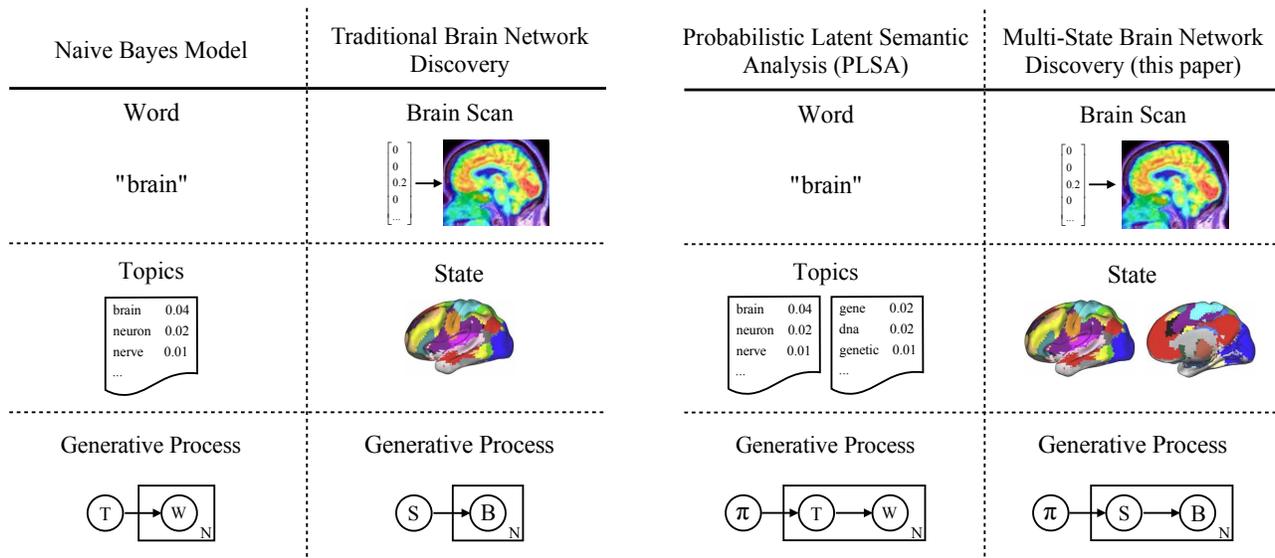


Fig. 1: The problem of multi-state brain network discovery. Brain activities over time may derive from the mixture of multiple brain states (*e.g.*, different brain states appear during different scenes of a movie). Without knowledge of mixture state assignment, our goal is to discover the multiple underlying brain network states, allowing for differing brain parcellation and connectivity.

state. This temporal adaptability (*i.e.*, state-based adjustment) in brain parcellation and connectivity becomes pivotal in understanding human brain networks. Ultimately, characterizing this mixture functional structure of brain parcellation and functional connectivity, as shown in Figure 1, leads to a better understanding of brain function and human behavior.

Knowledge Gap. Formally, the brain network discovery problem is anchored on inferring a set of functionally homogeneous brain regions as the network nodes and the mapping of their connectivity as network edges, all based on a series of brain scans taken over time. While there are some recent solutions [10, 13, 11] can address the problem, they often ignore the flexibility of functional network configurations.



(a) Naive Bayes [9] and Brain Network Discovery Model [10, 11] (b) PLSA[12] and Multi-State Brain Network Discovery Model (ours)

Fig. 2: Two pairs of comparison: (a) naive bayes model and traditional brain network discovery model, (b) PLSA and our model in this paper. In each generative process, the boxes are "plates" representing replicates. The outer plate represents document in naive bayes and PLSA, or observation subject in brain network study, while the inner plate represents the generative process of word (W) in a given document or brain scan (B) in a given subject, each of which word or brain scan is associated with a choice of topic (T) or state (S). π is the topic or state distribution. N denotes the number of words or scans.

They assume that the brain is always in a single activity state, implying that signals extracted from different regions of the brain at different times are members of the *same* network—a notion refuted by recent studies [6]. Additionally, some studies [14] have focused on obtaining parcellation by mapping the brain onto an atlas by image registration [15, 16]. The choice of atlas can influence the derived parcellation and generate distinct activity states. However, in the context of uncertainty in mixture state assignment, identifying the appropriate atlases across time presents a significant challenge. Thus, this paper investigates the problem of multi-state brain network discovery, as shown in Figure 1. The goal is to design methods in brain network discovery to capture *multiple* underlying brain network states, allowing for differing brain parcellation and connectivity.

Challenges. To incorporate the concept of *multiple states* into brain network discovery, our main challenges are:

- *Brain Network discovery*: Edge detection in functional brain network discovery focuses on direct links between the network nodes. However, the raw fMRI data usually do not contain background knowledge of node segmentation. Thus brain network discovery is the first challenge we face. Brain network discovery traditionally aims to use a cohesive model for inferring brain parcellation and edge detection at the same time [13]. However, the brain network may have different brain parcellations in different states and it is impractical to handle networks of each state with uniform node segmentation. Some recent brain network discovery methods [10, 13, 11] can handle learning both nodes and edges, though each has

its clear limitations. [11] aims to infer brain parcellation with spatial continuity constraint for the sake of interpretability, but it fails to distinguish the direct connections and indirect connections among the network nodes. [10] considers the brain network discovery problem as a coherent one, they apply two separate objective functions for two sub-tasks respectively, and update each other alternatively in the same framework. [13] suggests CGL (Coherent Graphical Lasso) deals with coherent brain network discovery, which combines the ideas of orthogonal non-negative matrix factorization with Graphical Lasso. However, this method can not solve the multi-state network in Figure 1 due to the lack of information about state assignment.

- *Mixture of multiple brain networks*: Some recent works [17, 18] have applied mixture models such as JGL (Joint Graphical Lasso) and MGL (Mixture Graphical Lasso) to brain network analysis. However, they all need brain parcellation to be given first. Thus, combining the existing Gaussian mixture model with a brain network discovery method remains unsolved.
- *Dependence between brain activity states and brain network*: Variations in brain activity states influence network estimations. Alterations of state assignments subsequently change the outcomes of corresponding network updates. Given this interdependency, pipeline methodologies that combine baseline techniques like CGL [13] and ON-MtF [11] are inappropriate for multi-state brain network discovery. Because the pipeline frameworks apply the methods for brain parcellation and edge detection sepa-

rately, such approaches lead to estimation inconsistency.

Proposed Method. To tackle the above challenges, we propose a new model, named MNGL, for multi-state brain network discovery, which jointly achieves brain parcellation and edge detection.

We leverage the idea of Probabilistic Latent Semantic Analysis (PLSA) [12], which was originally proposed to adopt a mixture model for natural language. PLSA assumes a given document is a mixture of topics, and the document was generated according to a probabilistic model with latent topics. Inspired by this, first, we view brain scans as mixtures of latent states, where each state S is characterized by a Gaussian distribution with its own covariance matrix Σ_S . Each Σ_S corresponds to a specific brain parcellation and connectivity between nodes. Therefore, in the generation of each brain scan, our model chooses a state S based on the mode distribution π (similar to how PLSA chooses a topic), and then generates a brain scan $B_i \sim \text{Multinomial}(\mathbf{0}, \Sigma_S)$ (as PLSA generates a word based on the topic chosen). Our model MNGL follows the basic idea of PLSA. By contrast, traditional brain network discovery models [11, 13] assume all brain scans are produced by a single state, which is characterized by a unified zero-mean Σ -covariance multivariate Gaussian distribution. Thus, they are analogous to naive bayes model [9]. Figure 2 illustrates these two pairs of comparison. To model this multi-state network, we combine CGL with GMM in a unified objective function to deal with multi-state networks. Compared to other mixture models [17, 18], our model only needs original brain data (a series of brain scans) as input without any prior knowledge related to nodes or their connectivity or assignments of each brain states, and outputs multiple brain networks that include both brain parcellation and connectivity structures.

Contributions. The contributions of this work are:

- We describe the open multi-state brain network discovery problem, which is to find the underlying network structure of hybrid cognitive brain states from a series of brain scans.
- We propose the first solution to this open problem, leveraging recent successes of Gaussian Mixture Models and the Coherent Graphical Lasso.
- We demonstrate that our model outperforms recent state-of-the-art alternatives by discovering more accurate and realistic results on both synthetic and real fMRI datasets.

II. PRELIMINARY

We begin by introducing some basic models to deal with brain parcellation, edge detection, and mixture modelling.

Brain Parcellation. Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$ be the observations of a p -variate Gaussian distribution where p denotes the number of variables and n is the number of observations. Then, we let Σ be the covariance matrix of the n samples. The Non-negative Matrix Factorization (NMF) can then be used to factorize Σ into two non-negative matrices:

$$\Sigma \approx \mathbf{F}\mathbf{G}^\top, \quad (1)$$

where $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_k) \in \mathbb{R}^{p \times k}$ and $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_k) \in \mathbb{R}^{p \times k}$, and k is a pre-specified number of nodes to discover. For network discovery, our target is an absolute covariance matrix Σ . So Σ is a systematic analysis matrix and \mathbf{F} is equal to \mathbf{G} , which we henceforth refer to \mathbf{F} and \mathbf{G} as \mathbf{H} . We can then extend the NMF model to weighted orthogonal non-negative factorization, or ONMtF [19], after which the objective function becomes:

$$\min_{\mathbf{H} \geq 0, \mathbf{H}\mathbf{H}^\top = \mathbf{I}} \|\Sigma - \mathbf{H}\mathbf{S}\mathbf{H}^\top\|^2. \quad (2)$$

By adding non-negativity and orthogonality constraints, the model is equivalent to k -means clustering and the Laplacian-based spectral clustering [20].

Edge Detection. As in [21], directed links among the network nodes can be discovered by minimizing the following objective:

$$\min_{\Theta > 0} (-\log \det \Theta + \text{tr}(\mathbf{S}\Theta) + \lambda \|\Theta\|_1), \quad (3)$$

where $\mathbf{S} = \frac{1}{n}\mathbf{X}\mathbf{X}^\top$ is the empirical covariance matrix, Θ is the precision matrix which is the inverse of the systematic covariance matrix Σ , ℓ_1 regularization is used to force sparsity, and λ is the parameter to control the sparseness of Θ . The edge e_{ij} between \mathbf{x}_i and \mathbf{x}_j exists if and only if $\theta_{ij} \neq 0$, where θ_{ij} is the (i, j) -element of Θ . We prefer Θ rather than matrix S in ONMtF, due to the power of sparse gaussian graphic models on large-scale datasets.

Coherent Graphical Lasso. The Coherent Graphical Lasso (CGL) achieves the two sub-tasks of Brain Network Discovery (node discovery and edge detection) simultaneously [13]. CGL is a special graphical lasso with an orthogonal non-negative matrix factorization, as shown in Equation 4:

$$\begin{aligned} \min_{\mathbf{H}, \Theta^*} & -\log \det \Theta^* + \text{tr}(\mathbf{H}^\top \mathbf{S} \mathbf{H} \Theta^*) + \lambda \|\Theta^*\|_1, \\ \text{s.t.} & \Theta^* \succ 0, \mathbf{H} \geq 0, \mathbf{H}\mathbf{H}^\top = \mathbf{I} \end{aligned} \quad (4)$$

where \mathbf{S} is the empirical covariance matrix, p is the number of features, k is the number of nodes, Θ^* is the inverse of the $k \times k$ absolute inter-node covariance matrix, and \mathbf{H} is a $p \times k$ cluster indicator matrix. However it can not be applied to the problem of the multi-state network discovery directly as it requires prior knowledge of state assignments.

Mixture Model. An attractive and powerful model for multi-state problems is the Gaussian Mixture Model (GMM), where each base distribution in the mixture is a Multivariate Gaussian (MVG) with mean μ_k and covariance matrix Σ_k . The probability of data sample \mathbf{x}_i is then

$$p(\mathbf{x}_i | \theta) = \sum_{k=1}^K \phi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k), \quad (5)$$

where θ is the model parameters, ϕ_k is the prior probability of the k -th distribution chosen to generate a sample and $\sum_{k=1}^K \phi_k = 1$. Next subsection, we introduce a novel model that extends CGL into the framework of a Gaussian Mixture Model, thereby solving the multi-state brain network discovery problem.

TABLE I: Important Notations.

Symbol	Definition
m	The number of gaussian distributions
$\mathbf{X} \in \mathbb{R}^{n \times p}$	n observations of p -variate Gaussian distribution
$\mathbf{H}_j \in \mathbb{R}^{p \times k}$	The clustering indicator matrix and $0 \leq j \leq m$
$\mathbf{Y} \in \mathbb{R}^{n \times k}$	n the projection of \mathbf{X} along \mathbf{H} matrix
$\Sigma_j \in \mathbb{R}^{p \times p}$	The covariance of p -variate Gaussian distribution
$\Sigma_j^* \in \mathbb{R}^{k \times k}$	The projection of Σ along \mathbf{H} matrix
$\Theta_j \in \mathbb{R}^{p \times p}$	The true precision matrix of all variables
$\Theta_j^* \in \mathbb{R}^{k \times k}$	The true precision matrix of all nodes
$\hat{\Theta}_j \in \mathbb{R}^{p \times p}$	The estimate of true precision matrix of all variables Θ
$\hat{\Theta}_j^* \in \mathbb{R}^{k \times k}$	The estimate of true precision matrix of all nodes Θ^*
ϕ_j	The prior probability of the j -th base distribution chosen to generate a sample
γ_{ij}	The posterior probability of the i -th observation generated by the j -th distribution

III. METHODOLOGY

Multi-State Network Graphical Lasso. In this work, we propose the first method for Multi-State Brain Network Discovery, which we refer to as the Multi-State Network Graphical Lasso, or MNGL. Firstly, following the idea of [13], we map the original variable space \mathbf{X} into a new feature space \mathbf{Y} similarly on the covariance matrix Σ :

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{Y} = \mathbf{H}^\top \mathbf{X}, \\ \Sigma &\leftarrow \Sigma^* = \mathbf{H}^\top \Sigma \mathbf{H}, \end{aligned} \quad (6)$$

where \mathbf{Y} denotes the new k -dimensional feature space where each feature represents node, Σ^* represents the inter-node covariance matrix, and \mathbf{H} represents a cluster indicator matrix. Σ^* thus measures the association between each node \mathbf{y}_i [13].

For the rest of this section, we describe our proposed model in terms of \mathbf{y}_i instead of \mathbf{x}_i . k is the index of nodes, j is the index of distributions, i is the index of samples. μ_j^* and Σ_j^* represent the parameters of mean vector and covariance matrix corresponding to the j -th mixture gaussian distribution of \mathbf{Y} , respectively. Then, Θ_j^* represents the inverse matrix of covariance matrix Σ_j^* . More special notations are collected in Table I.

According to the notation above, given the number of base distributions m and the number of node k , we assume the observed sample of target feature space can be mapped into a new feature space (nodes), which also follows a mixture of the k gaussian distributions. The sample size is given as n . Thus, the joint probability of these nodes $\mathbf{Y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top) \in \mathbb{R}^{n \times k}$ is given by

$$p(\mathbf{Y} | \{\Theta_j^*\}, \{\mu_j^*\}, \{\phi_j\}) = \prod_{i=1}^n \sum_{j=1}^m \phi_j \mathcal{N}(\mathbf{y}_i | \mu_j^*, \Sigma_j^*).$$

By assuming $\mu_j^* = \mathbf{0}$ without losing generality, the negative log likelihood (NLL) in terms of $\{\Theta_k^*\}$ is given by,

$$\text{NLL}(\theta) = - \sum_{i=1}^n \log \left(\sum_{j=1}^m \phi_j \mathcal{N}(\mathbf{y}_i | \mathbf{0}, (\Theta_j^*)^{-1}) \right), \quad (7)$$

where $\theta = \{\phi_1, \dots, \phi_m, \Theta_1^*, \dots, \Theta_m^*\}$ is the model parameters.

Latent States. In order to solve the Equation 7, we follow the idea of Jensen inequality and build a latent variable in the sum term of each expression in log. Since there are m separate

latent distributions, each data sample of the corresponding node \mathbf{y}_i could come from one of the K distributions. We therefore construct a latent variable $\mathbf{Q}(z_{ij})$ which we constrain such that $\sum_{j=1}^m \mathbf{Q}(z_{ij}) = 1$. Then, the NLL function can be rewritten as follows:

$$\text{NLL}(\theta) = - \sum_{i=1}^n \log \sum_{j=1}^m \left(\frac{\mathbf{Q}(z_{ij}) p(\mathbf{y}_i | \Theta_j^*, \phi_j)}{\mathbf{Q}(z_{ij})} \right) \quad (8)$$

$$= - \sum_{i=1}^n \log \sum_{j=1}^m \left(\frac{p(\mathbf{y}_i, z_{ij} | \Theta_j^*, \phi_j)}{\mathbf{Q}(z_{ij})} \right). \quad (9)$$

We next prove that this can be treated as the posterior probability of the i -th observation generated by the j -th distribution.

According to the Jensen inequality, the expression in the Equation 8 can be rewritten for the EM algorithm to optimize the function, which can be split into expectation and maximization steps, respectively.

Expectation. First, according to the Jensen inequality, we know that when the optimal function is convex,

$$f(E(x)) \leq E(f(x)). \quad (10)$$

Because NLL is convex, and $\sum_{j=1}^m \left(\frac{p(\mathbf{y}_i, z_{ij} | \Theta_j^*, \phi_j)}{\mathbf{Q}(z_{ij})} \right)$ can be treated as the expectation of $p(\mathbf{y}_i, z_{ij} | \Theta_j^*, \phi_j)$. So we apply Jensen inequality here to find a lower bound:

$$\text{NLL}(\theta) \leq - \sum_{i=1}^n \sum_{j=1}^m \mathbf{Q}(z_{ij}) \log(p(\mathbf{y}_i, z_{ij} | \Theta_j^*, \phi_j)). \quad (11)$$

These terms are only equal when

$$\frac{p(\mathbf{y}_i, z_{ij})}{\mathbf{Q}(z_{ij})} = C, \quad (12)$$

where C is a constant. So, we simply have:

$$\sum_{j=1}^m p(\mathbf{y}_i, z_{ij}) = C \sum_{j=1}^m \mathbf{Q}(z_{ij}) = C, \quad (13)$$

$$\mathbf{Q}(z_{ij}) = \frac{p(\mathbf{y}_i, z_{ij})}{\sum_{j=1}^m p(\mathbf{y}_i, z_{ij})} = r_{ij}. \quad (14)$$

The equation of $\text{NLL}(\theta)$ is correct only when the constraint of $\mathbf{Q}(z_{ij})$ is true. Thus we can conclude that the latent variable is the posterior probability of the i -th observation generated by the j -th distribution. Therefore, we can compute each r_{ij} based on the initialization or update results of Θ_j^* and ϕ_j .

Maximization. Given the $\mathbf{r}_{ij}^{(t)}$ from the Expectation step, we update $\hat{\phi}_j$, $\hat{\mathbf{H}}_j$ and $\hat{\Theta}_j^*$, respectively. First, we update $\hat{\phi}_j$ based as follows:

$$\hat{\phi}_j^{(t)} = \frac{1}{n} \sum_{i=1}^n r_{ij}^{(t)}. \quad (15)$$

The remaining problem is to find the optimal estimations of \mathbf{H}_j and Θ_j^* that maximizes the expectation we obtain in the E step. Through a simple proof, it is equivalent to minimize the following function:

$$\begin{aligned} \min \sum_{i=1}^n \sum_{j=1}^m -r_{ij}^{(t)} (\log |\Theta_j^*| - \mathbf{x}_i^\top \mathbf{H}_j \Theta_j^* \mathbf{H}_j^\top \mathbf{x}_i), \\ \text{s.t. } \Theta_j^* \succ 0, \mathbf{H}_j \geq 0, \mathbf{H}_j \mathbf{H}_j^\top = \mathbf{I}. \end{aligned} \quad (16)$$

Intuitively, the problem above is equivalent to m separate conventional graphical lasso sub-problems weighted by $r_{ij}^{(t)}$ where each sub-problem has the form of

$$\begin{aligned} \min -\log |\Theta_j^*| + \text{tr}(\tilde{\mathbf{X}}_j^\top \mathbf{H}_j \Theta_j^* \mathbf{H}_j^\top \tilde{\mathbf{X}}_j), \\ \text{s.t. } \Theta_j^* \succ 0, \mathbf{H}_j \geq 0, \mathbf{H}_j \mathbf{H}_j^\top = \mathbf{I}, \end{aligned} \quad (17)$$

where $\tilde{\mathbf{X}}_j = (\sqrt{r_{1j}/s_j} \mathbf{x}_1^\top, \dots, \sqrt{r_{nj}/s_j} \mathbf{x}_n^\top)$, $\mathbf{r}_{ij} = (r_{1j}, \dots, r_{nj})^\top$ and $s_j = \sum_{i=1}^n r_{ij}$. Then we bring in the ℓ_1 regularization $\lambda \|\Theta_j^*\|_1$ to obtain the final objective function for the Maximization step.

This problem is not convex *w.r.t.* $\{\Theta_j^*\}$, but we could solve it alternatively for each Θ_j^* by regarding other $\Theta_{j' \neq j}^*$ fixed. Each sub-problem of Θ_j^* is exactly in the form of Equation 17 plus the ℓ_1 regularization terms. Thus the estimation of Θ_j^* could be solved by any existing method for solving Graphical Lasso without significant modifications. To estimate \mathbf{H}_j we follow the algorithm similar to NMF, using Karush–Kuhn–Tucker (KKT) complementary slackness conditions to enforce the non-negativity and orthogonality constraints, then solving the estimation of \mathbf{H}_j by the multiplicative update rule. Thus, we have:

$$(\hat{\mathbf{H}}_j^{(t+1)})_{\text{ls}} = \left(\hat{\mathbf{H}}_j^{(t)} \right)_{\text{ls}} \left(\frac{\tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_j^\top \hat{\mathbf{H}}_j^{(t)} \hat{\Theta}_j^{*-} + \hat{\mathbf{H}}_j^{(t)} \lambda_1^-}{\tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_j^\top \hat{\mathbf{H}}_j^{(t)} \hat{\Theta}_j^{*+} + \hat{\mathbf{H}}_j^{(t)} \lambda_1^+} \right)_{\text{ls}}. \quad (18)$$

Here λ_1 is $k \times k$ Lagrangian multip matrices following the non-negativity constraint and its compact expression follows as below:

$$\lambda_1 = -\hat{\mathbf{H}}_j^\top \tilde{\mathbf{X}}_j \tilde{\mathbf{X}}_j^\top \hat{\mathbf{H}}_j \hat{\Theta}_j^*. \quad (19)$$

To make sure each part is non-negative, We divide the λ_1 and $\hat{\Theta}_j^*$ into two parts, respectively:

$$\begin{aligned} \lambda_1 &= \lambda_1^+ - \lambda_1^-, \\ \lambda_1^+ &= \frac{(|\lambda_1| + \lambda_1)}{2}, \\ \lambda_1^- &= \frac{(|\lambda_1| - \lambda_1)}{2}. \end{aligned} \quad (20)$$

The same is true on the $\hat{\Theta}_j^*$. Thus we can make sure the sign of numerator and denominator are all positive, abiding by the non-negative constraint of \mathbf{H}_j .

Algorithm 1 Algorithm for MNGL

Require: i: \mathbf{X} : The observations of D -variate Gaussian distribution

ii: m : the number of Gaussian distributions

iii: k : the number of nodes (groups)

iv: λ_1 : the Lagrangian multiplier of the ℓ_1 regularization in graphical lasso

v: iter_{\max} : the maximum number of iteration

Output: $\hat{\Theta}_j^*$, $\hat{\mathbf{H}}_j$ and $\hat{\phi}_j$

- 1: Initialization: initialize $\phi^{(0)}_j$, $\hat{\Theta}_j^{*(0)}$, $\hat{\mathbf{H}}_j^{(0)}$ and $r_{ij}^{(0)}$
 - 2: **repeat**
 - 3: E step: Update the latent variable $r_{ij}^{(t)}$ with given $\hat{\phi}_j^{(t-1)}$, $\hat{\Theta}_j^{*(t-1)}$ and $\hat{\mathbf{H}}_j^{(t-1)}$
 - 4: M step: Update $\hat{\phi}_j^{(t)}$, $\hat{\Theta}_j^{*(t)}$ and $\hat{\mathbf{H}}_j^{(t)}$ with $r_{ij}^{(t)}$
 - 5: **until** $\text{iter} = \text{iter}_{\max}$ or convergence
-

In each iteration of the Maximization step, the alternating optimization repeats until all estimated $\hat{\Theta}_j^*$, $\hat{\mathbf{H}}_j$ and $\hat{\phi}_j$ become stable or reaches the maximal number of iterations. The final solutions to Equation 17 and the updated $\{\hat{\phi}_j\}$ are obtained using Equation 15 are used in the upcoming iteration of Expectation step to update the responsibility weights $\{r_{ij}\}$. This looping of Expectation and Maximization repeats until the loss function converges. The MNGL algorithm is also summarized in Algorithm 1.

Initialization. As shown in Algorithm 1, we need to provide starting values for each estimator. The following scheme we found empirically works well in our experiments. For each observation $i = 1, \dots, n$, we distribute the observation randomly a class $j \in \{1, \dots, m\}$. Then we assign a weight $\hat{r}_{ij} = 0.9$ for this observation i and distribution k and $\hat{r}_{ij} = \frac{0.1}{m-1}$ for all other distributions. In the Maximization step, we update $\hat{\Theta}_j^*$ from the initial values $\hat{\Theta}_j^{*(0)}$ computed by CGL based on the whole samples and $\hat{\phi}_j$ from the initial values $\hat{\phi}_k = \frac{1}{m}$. Then for $\hat{\mathbf{H}}_j$, according to the Equation 18, we note that if $(\hat{\mathbf{H}}_j^{(t+1)})_{\text{ls}} = 0$ in one iteration, it will never jump out from this local solution. Thus, our experiments we initialize $\hat{\mathbf{H}}_j^{(0)}$ by performing k -means clustering then setting $\hat{\mathbf{H}}_j^{(0)} \leftarrow \hat{\mathbf{H}}_j^{(0)} + 0.2$.

IV. EMPIRICAL STUDY

We begin by evaluating our method using synthetic data where we have access to the ground truth brain states. To comprehensively evaluate the proposed model, we conduct experiments to answer the following research questions:

- *RQ 1:* How does sample size affect MNGL's performance relative to state-of-the-art alternatives?
- *RQ 2:* How robust is MNGL to the presence of noise compared to other recent models?
- *RQ 3:* How do hyper-parameters in comparative experiments impact each model's performance?
- *RQ 4:* How does the number of nodes affect each compared model?

A. Experiment Setup

1) *Synthetic Data with Ground-Truth*: We evaluate the performance of our model on synthetic data, where the ground-truth is known. The first step of generating these synthetic data is to build a mixture Gaussian distribution of network structure. By following the approach of [13] in generating a single network, we generate m different block-diagonal matrix Θ_j and \mathbf{H}_j firstly. We refer to each diagonal block as the node in real-case. For each Θ_j , we give random sparsity structures for each block Θ_{G_i, G_j} . In this paper, we design each diagonal block Θ_{G_i, G_i} in one Θ_j with different scale. Thus by adjusting scale of diagonal blocks in different matrix Θ_j , we can make different network have different node parcelation. To simulate the connectivity of variables among diagonal and off-diagonal blocks, we control the connectivity of each variables on diagonal block with a high density, then giving a low density to each off-diagonal block. Following the above steps, we generate several different Θ_j and \mathbf{H}_j . Then each Θ_j^* can be derived from $\mathbf{H}_j^T \Theta_j \mathbf{H}_j$.

Given Θ_j , we can thus obtain Σ_j , which is the inverse of Θ_j . Due to the assumption of the independence of each Gaussian distribution, we obtain the covariance matrix Σ of the mixture Gaussian distribution. Then we generate n samples randomly from the mixture Gaussian distribution.

2) *Compared methods*: To demonstrate the effectiveness of our proposed method, we test against several state-of-art methods coherent brain network discovery methods:

- *CGL* [13]: CGL aims to achieve node discovery and directed edge detection at the same time. Meanwhile, it can distinguish direct links from indirect connections due to its solid probabilistic formulation.
- *ONMtF* [11]: ONMtF also aims to complete node discovery and edge detection at the same time. However, it focuses on explaining the spatial continuity of results. We only apply it on the task of nodes discovery, due to its inability of directed edge discovery.
- *k-means + CGL*: This pipeline method is more appropriate than CGL for the problem defined in this paper. We first employ *k-means* to assign each x_i to different nodes, then using CGLasso for each group to obtain the final $\hat{\Theta}_j^*$ and $\hat{\mathbf{H}}_j$.
- *k-means + ONMtF*: This is also a pipeline method that first splits the whole sample of x_i into different nodes by using *k-means*, then using ONMtF on each node to obtain each $\hat{\Theta}_j^*$ and $\hat{\mathbf{H}}_j$.
- *k-means + JGL* [17]: A Joint Graphical Model is proposed in [17], which aims to discover a mixture Gaussian distribution. However, it applies to the level of nodes. We therefore employ *k-means* to map x_i into the node space of y_i first.

3) *Experiment Setting*: We simulate four scenarios by changing one parameter and keeping the others fixed. Each scenario aims to study one of aforementioned research questions (RQ). In these situations, we select sample size n , the

standard error of noise σ , the variables number p of x_i , and the group number k as the controlled parameters.

- *Scenario 1*: We fix $p = 70$ (the number of variables), $\sigma = 0$ (the standard error of noise), $k = 5$ (the number of nodes) and then control sample size n from 200 to 2000.
- *Scenario 2*: We fix $n = 2000$, $p = 70$ and $k = 5$, meanwhile control σ from 2 to 5.
- *Scenario 3*: We fix $n = 2000$, $\sigma = 0$ and $k = 5$, and then control p from 70 to 350.
- *Scenario 4*: We fix $n = 2000$, $\sigma = 0$, and then control k from 3 to 11.

To generalize the results of comparative experiments, we sample 10 times for all experiments and average their results to evaluate the precision and stability of our model.

4) *Evaluation Protocol*: To evaluate the quality of edge detection, we employ Accuracy and F1-score in the comparative experiments. We follow [22] to define the accuracy and F1-score of edge detection:

$$\text{Accuracy} = \frac{n_d}{n_g}, \quad (21)$$

$$\text{F1} = \frac{2n_d^2}{n_a n_d + n_g n_d}, \quad (22)$$

where n_d is the number of true edges detected by the algorithm, n_g is the number true edges and n_a is the total number of edges detected. Higher accuracy score or higher F1 score indicates better quality of edge detection. To evaluate the quality of clustering, we follow [23] to use the purity score and normalized mutual information score (NMI). Higher purity score or higher NMI score indicates better quality of clustering.

B. Comparative Results

To study the effect of sample size on the performance of MNGL, we design comparative experiments based on *Scenario 1*. Figure 3 shows the comparative results. We compare our proposed model with five baseline methods. The first row shows the results of the comparison on edge detection; the second row shows the results for node discovery. In the results of all scenarios, we use the same symbol, which is illustrated in the caption below Figure 3. From the results in the first row of Figure 3, we observe that the sample size n indeed affects some methods, especially ONMtF and its derivations. As the sample size increases, the accuracy of these two methods become much higher. Encouragingly, this factor has no significant effect on our model. Overall, we can clearly see that our method MNGL is more accurate and robust than other methods as the sample size n changes. Meanwhile, n does not have a significant impact on the performance of MNGL, which means that our model performs well even with a small training set.

To study the effect of noise on the performance of MNGL, we use the experimental setup *Scenario 2*. Our results are shown in Figure 4 where the horizontal axis in the figure represents the standard error of noise σ . The larger the σ ,

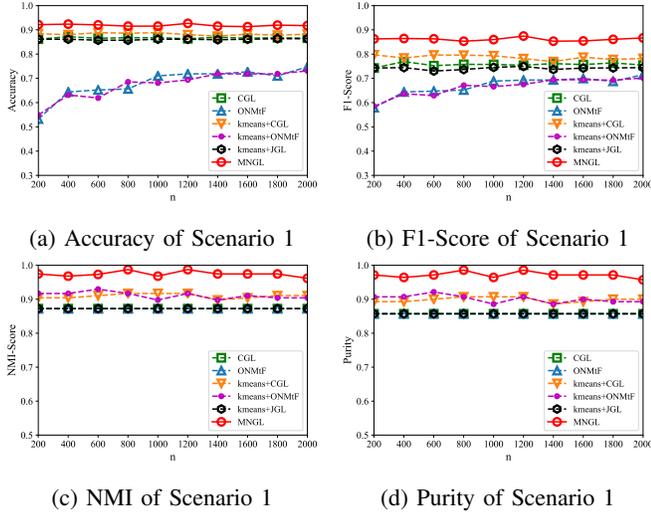


Fig. 3: Comparison of each method on edge detection and node discovery. The first row shows the results of edge detection, and the second shows the results of node discovery. The four sub-figures above consider different sample size n from 200 to 2000. The other parameters are left fixed.

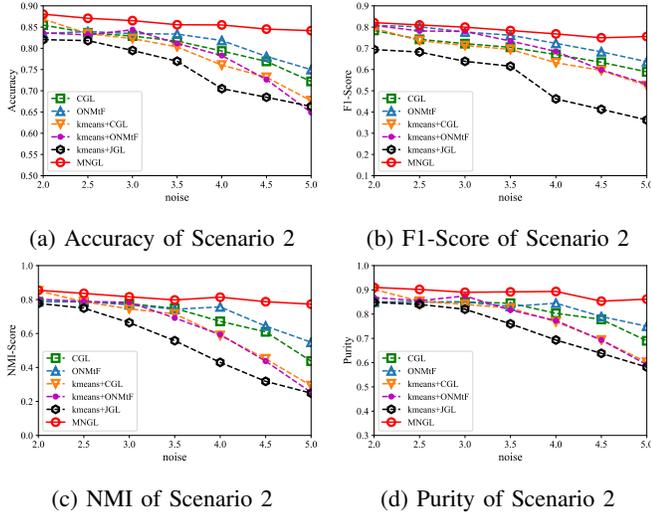


Fig. 4: The four sub-figures above consider different σ (the standard error of noise) from 2 to 5, meanwhile fix the other parameters, which correspond to scenario2;

the stronger the noise. It leads to smaller signal-to-noise ratio, which means it is more difficult to mine the network structure from the available samples. As seen in the four sub-graphs, we find that noise affects all compared methods. In particular, while JGL suffers the most influence, ONMF and derivatives of it are more robust than CGL and its derivatives in this scenario. Meanwhile, our method, MNGL, is better than all other comparison methods in this experiment for both edge detection and node discovery. Furthermore, σ does not significantly decay the performance of MNGL.

To study the effect of the number of variables on the performance of MNGL, we next use *Scenario 3*, the results for which are shown in Figure 5. For the node discovery

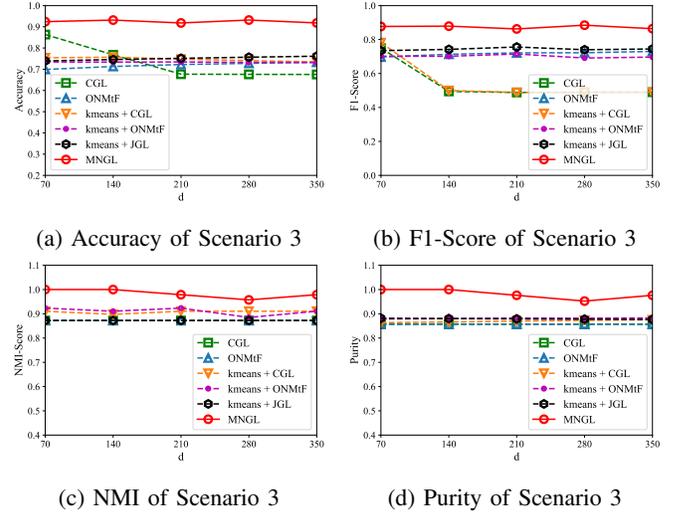


Fig. 5: The four sub-figures above consider different p (the number of variables x_i) from 70 to 350, meanwhile fix the other parameters, which correspond to scenario3;

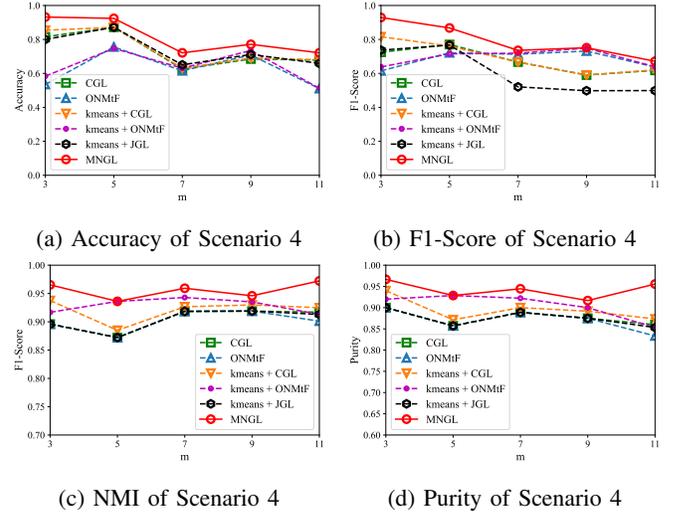


Fig. 6: The four sub-figures above consider different k (the number of nodes y_i) from 3 to 11, meanwhile fix the other parameters, which correspond to scenario4;

task, we see that the dimension of feature space has no impact on the performance of any methods. However, for edge detection, when the dimension is low, the performance of CGL and its derivatives outperforms the other baseline methods. In particular, as the dimension increases, the accuracy of CGL and its derivatives shows a significant downward trend compared to the others. Again, as expected, in this scenario the performance of MNGL is more accurate and robust than the baseline methods.

To study the effect of the number of states on the performance of MNGL, we turn to *Scenario 4* and report our findings in Figure 6. Across all sub-figures, as the number of nodes increases, the robustness of all compared methods shows a downward trend. Specifically, for the edge detection task, the accuracy of CGL and its derivatives perform

slightly better than ONMtF and its derivatives. However, when considering the F1-Score, ONMtF outperforms the CGL methods. For node discovery, ONMtF and its derivatives show more robustness than other baseline methods. Overall, MNGL still significantly outperforms the compared methods in these settings, though there is a small degree of fluctuation.

Combining the four *RQs* raised above and the results of all these comparative experiments, we can draw the following four conclusions: First, ONMtF and its derivatives are not as good as other methods in the case of insufficient samples. Second, CGL and its derivatives are more restrictive in high-dimensional space. Third, Both the accuracy and robustness of all comparative methods will decrease under the impact of noise and the number of nodes. Fourth, compared to the alternative methods, our proposed method MNGL exhibits greater accuracy and robustness in each scenario, indicating that neither sample size n , the dimension of feature space p , noise σ nor the number of nodes k significantly degrades the performance of our method.

C. Real-World Datasets

We also evaluate our proposed method on the fMRI dataset from the ADHD-200 project. Attention Deficit Hyperactivity Disorder, or ADHD, is a chronic and sometimes-devastating condition affecting 5-10% of school-age children. It is also extremely costly to treat – the United States alone has spent more than 36 billion on ADHD [24]. This real-world dataset is distributed by nilearn. Specifically, there are 40 subjects in total. Among them, 20 subjects are labeled as ADHD, and the others are labeled as typically developing children (TDC). The fMRI scan of each subject in the dataset is a series of snapshots of 3D brain images of size $91 \times 109 \times 91$ over ~ 176 time steps. Because the fMRI scanning datasets are contain only voxels, the nodes and connectivity among them are all unknown. [11, 13] put these two tasks in a unified model to find the optimal solution. However, they ignore the assumption of mixture network structure we defined in this paper. Furthermore, this part of the experiment lacks ground-truth as a reference to measure the accuracy and robustness of the model. Therefore we must consider the interpretability and rationality of the results. Specific to our proposed model, we are primarily concerned with whether or not our model can mine different cognitive networks from the fMRI datasets (various node assignments and functional connectivity). Therefore for this subsection, we focus solely on applying our proposed method, MNGL, to this challenging task.

In our experimental setting, we focus on the multi-state brain network discovery among the same subjects, and report the results of both nodes discovery and edge detection. To assign the voxels that can be considered as parts of the brain, we use anatomical automatic labeling (AAL) brain-shaped mask, which is provided by neurology professionals. We follow [25] and use a middle slice of these scan for the ease of presentation. Consequently, each of the brain scans can be represented by about 3281 voxels. So it is more conventional

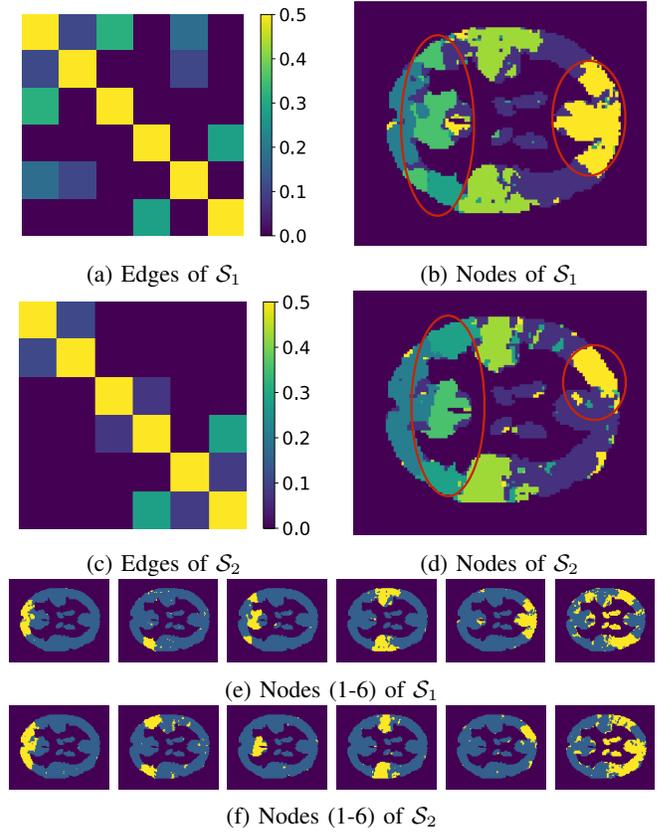


Fig. 7: Discovered results of multi-state brain network in ADHD subjects ($k = 6$).

for the visualization of the results. The datasets is a 3281 (variables) \times 2992 (time steps) datasets and reasonably assume that they are drawn from a mixture Gaussian distribution. However, the number of Gaussian distributions m and the number of nodes k are both unknown and need to be selected in advance. Through repeated experimental observations, we find that $m = 2$ and $k = 6$ can provide the most reasonable results on the data sets.

Figure 7 shows the multi-state network discovered by MNGL on the fMRI datasets. The results of edge detection and node discovery are shown on the first and second line, which corresponds to the functional network of state S_1 and S_2 respectively. Meanwhile, each inferred node is displayed on the third and fourth line individually. First, we can see the difference between the two networks from the discovered edges and nodes. We deliberately mark the differences between the nodes of two networks with red circles. More specifically, in the first line of Figure 7, we find a strong and complete default mode network (DMN) for ADHD subjects, corresponding to group 3 and 5 in the third line. A DMN is a network of interacting brain regions known to have activity highly correlated with each other while being distinct from other networks in the brain, including the Parietal and Occipital Lobes, the Cingulum Region Posterior, and the Frontal Cortex. However, in the second line of Figure 7, this mode is not intact. In particular, the Frontal Cortex is missing

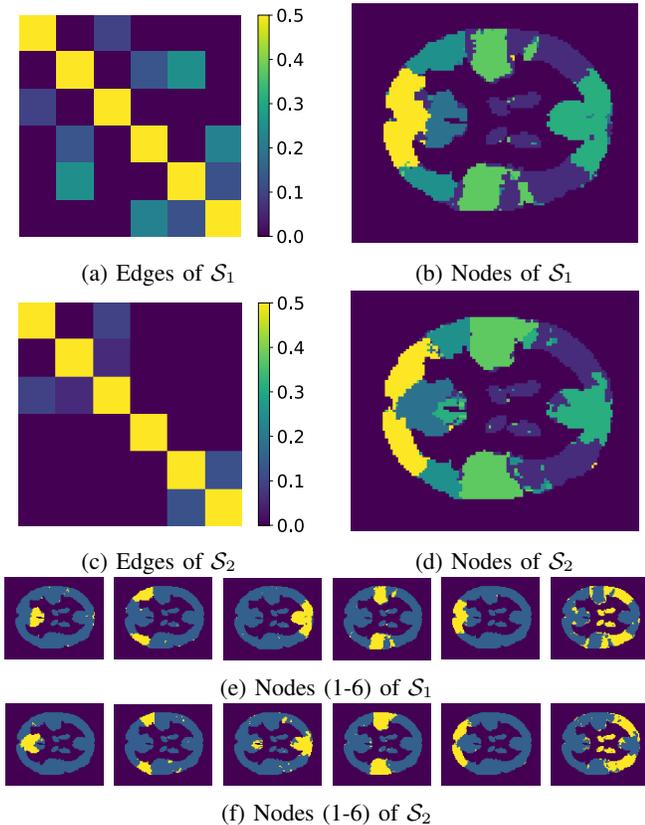


Fig. 8: Discovered results of multi-state brain network in TDC subjects ($k = 6$).

in the network, while the rest of the connections are different from the functional network in the first line. The specific relationship between each node can be found in the sub-figures of discovered edges.

For comparison, we apply MNGL again on the subjects of TDC, which represent the group of typically-developing children. We can observe from Figure 8 that, although there are differences in the node assignments and the connectivity structure among each node, there is no deletion of DMN in each network. At the same time, the network of state \mathcal{S}_1 from TDC and that from ADHD have a certain degree of similarity from the nodes result to the connectivity structure, which allows us to have more reference when analyzing the differences and connections between the two subjects. These results give us reason to believe that the brain scans of these subjects have some similar functional structure correspondences similar to on-task and off-task states.

Despite the lack of ground-truth, we believe that the current results are still consistent with the problem defined in this paper: We find strong evidence that there is a multi-state brain cognitive network in the fMRI datasets, and our proposed model MNGL can effectively mine this mixture network structure.

V. RELATED WORKS

Existing works can be divided into two categories. Firstly, for coherent brain network discovery, ONMtF [26] is a useful

pattern recognition method. [11] extend ONMtF by adding a spatial continuity penalty, which can increase the interpretability of the parcellated regions. This method is a coherent model which can output the result of nodes discovery and edge detection simultaneously. However, it has discovered the edges based on the correlation matrix instead of inferring direct links between each node. Instead of using a correlation matrix, [21, 27] focus on sparse inverse covariance estimation for discovering connectivity of brain network based on large-scale datasets. These kinds of methods can distinguish direct links from indirect connections due to their solid probabilistic foundation. [13] propose a model called CGL to achieve the coherent brain network discovery, including edge detection and node discovery. However, this method ignores the problem of multi-state problem we mentioned in this paper.

For the multi-state problem, we consider the Gaussian Mixture Model (GMM) [28]. GMMs model the distribution of data observations as a weighted sum of parameterized Gaussian distributions. However, a prominent issue related to GMM is estimating the parameters given observations [29]. Through many extensions, the EM algorithm has proven to be a powerful algorithm for the maximum-likelihood estimation of GMMs [30]. Additionally, [31, 32] consider the issue of the number of mixture components in the model, which can lead to over-fitting in practice. GMMs have been widely used in many areas, especially for network discovery [33, 34, 35]. Most existing studies for mixture modeling focus on regularizing only the mean parameters with diagonal covariance matrices [36, 37], though some works [38, 39, 40] have started considering regularization of the covariance parameters. However, these works do not touch on the key issue of identifying the varying sparse structures of the precision matrices across the components of a mixture model in brain network discovery. [17] proposes a joint graphical model (JGL) to deal with cluster-specific networks. [18] aims to edge detection task by combing graphical lasso with GMM. However, these models need brain parcellation to be given first. Thus, existing models related to GMM are thus not suitable for the special problem defined in this paper.

VI. CONCLUSION

In this work, we define the open problem of multi-state brain network discovery, which is to infer various brain parcellations and connectivities across different brain states. Previous works on brain network discovery derive an average brain network based on the assumption that only one single activity state of the brain generates the signals. However, according to recent studies in the area of brain network, assuming single-state networks ignores a crucial of cognitive brain networks. To better understand the temporally-changing functional network of the brain, we propose a novel model called MNGL, which can discover *multiple* brain networks, including nodes and their connectivity based on on only unlabeled fMRI scans. Through extensive controlled experiments, we demonstrate that our proposed model shows more effectiveness and robustness than other baseline models. MNGL also shows expected

and meaningful results on the real ADHD-200 fMRI dataset. We thus have reason to believe that our method can be applied in multi-state brain network for a better understanding of brain function and behavioral performance.

VII. ACKNOWLEDGMENTS

Hang Yin and Xiangnan Kong was supported in part by NSF grant IIS-1718310. Yanhua Li was supported in part by NSF grants IIS-1942680 (CAREER), CNS-1952085, and DGE-2021871.

REFERENCES

- [1] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10, 2009.
- [2] A. Zalesky, A. Fornito, and E. T. Bullmore. Network-based statistic: identifying differences in brain networks. *NeuroImage*, 53(4):1197–1207, 2010.
- [3] L. Zhou, L. Wang, and P. Ogunbona. Discriminative sparse inverse covariance matrix: Application in brain functional network classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3097–3104, 2014.
- [4] Yuhui Du, Zening Fu, and Vince D Calhoun. Classification and prediction of brain disorders using functional connectivity: promising but challenging. *Frontiers in neuroscience*, 12:525, 2018.
- [5] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.
- [6] Pan Lin, Yong Yang, Junfeng Gao, Nicola De Pisapia, Sheng Ge, Xiang Wang, Chun S Zuo, James Jonathan Levitt, and Chen Niu. Dynamic default mode network across different brain states. *Scientific Reports*, 7:46088, 2017.
- [7] Alana J Anderson and Sammy Perone. Developmental change in the resting state electroencephalogram: insights into cognition and the brain. *Brain and Cognition*, 126:40–52, 2018.
- [8] Ibai Diez and Jorge Sepulcre. Neurogenetic profiles delineate large-scale connectivity dynamics of the human brain. *Nature Communications*, 9(1):1–10, 2018.
- [9] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.
- [10] Xinyue Liu, Xiangnan Kong, and Philip Yu. Collective discovery of brain networks with unknown groups. In *Proceedings of the 30th International Joint Conference on Neural Networks*, pages 3569–3576. IEEE, 2017.
- [11] Zilong Bai, Peter Walker, Anna Tschiffely, Fei Wang, and Ian Davidson. Unsupervised network discovery for brain imaging data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 55–64. ACM, 2017.
- [12] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296, 1999.
- [13] Hang Yin, Xiangnan Kong, and Xinyue Liu. Coherent graphical lasso for brain network discovery. In *2018 IEEE International Conference on Data Mining*, pages 1392–1397. IEEE, 2018.
- [14] Yao Su, Zhentian Qian, Lei Ma, Lifang He, and Xiangnan Kong. One-shot joint extraction, registration and segmentation of neuroimaging data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2049–2060, 2023.
- [15] Yao Su, Zhentian Qian, Lifang He, and Xiangnan Kong. Ernet: Unsupervised collective extraction and registration in neuroimaging data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1666–1675, 2022.
- [16] Yao Su, Xin Dai, Lifang He, and Xiangnan Kong. Abn: Anti-blur neural networks for multi-stage deformable image registration. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 468–477. IEEE, 2022.
- [17] Chen Gao, Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Estimation of multiple networks in gaussian mixture models. *Electronic Journal of Statistics*, 10:1133, 2016.
- [18] Hang Yin, Xinyue Liu, and Xiangnan Kong. Gaussian mixture graphical lasso with application to edge detection in brain networks. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1430–1435. IEEE, 2020.
- [19] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, pages 126–135. ACM, 2006.
- [20] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 5th SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.
- [21] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [22] Sen Yang, Zhaosong Lu, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2): 916–943, 2015.
- [23] Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, and Tianyi Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology*, pages 565–576. ACM, 2009.
- [24] Lenard A Adler, David M Shaw, Kimberly Kovacs, and Samuel Alperin. Diagnosing adhd in children and adults. *Attention-Deficit Hyperactivity Disorder in adults and children*, pages 16–23, 2015.
- [25] Chia-Tung Kuo, Xiang Wang, Peter Walker, Owen Carmichael, Jieping Ye, and Ian Davidson. Unified and contrasting cuts in multiple graphs: Application to medical imaging segmentation. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 617–626. ACM, 2015.
- [26] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, 101(12):4164–4169, 2004.
- [27] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.
- [28] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.
- [29] Kerry Back and David P Brown. Implied probabilities in gmm estimators. *Econometrica: Journal of the Econometric Society*, pages 971–975, 1993.
- [30] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.
- [31] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
- [32] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [33] Mark EJ Newman and Elizabeth A Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569, 2007.
- [34] Yunzhang Zhu, Xiaotong Shen, and Wei Pan. Structural pursuit over multiple undirected graphs. *Journal of the American Statistical Association*, 109(508):1683–1696, 2014.
- [35] Geoffrey McLachlan and David Peel. Finite mixture models, wiley series in probability and statistics, 2000.
- [36] Sijian Wang and Ji Zhu. Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics*, 64(2):440–448, 2008.
- [37] Benhuai Xie, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, 2:168, 2008.
- [38] Hui Zhou, Wei Pan, and Xiaotong Shen. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473, 2009.
- [39] Steven M Hill and Sach Mukherjee. Network-based clustering with mixtures of l1-penalized gaussian graphical models: an empirical investigation. *arXiv preprint arXiv:1301.2194*, 2013.
- [40] Meng Yun Wu, Dao Qing Dai, Xiao Fei Zhang, and Yuan Zhu. Cancer subtype discovery and biomarker identification via a new robust network clustering algorithm. *PLoS One*, 8(6), 2013.