The Annals of Applied Statistics
2020, Vol. 14, No. 2, 993–1029
https://doi.org/10.1214/20-AOAS1339

A RANDOM EFFECTS STOCHASTIC BLOCK MODEL FOR JOINT COMMUNITY DETECTION IN MULTIPLE NETWORKS WITH APPLICATIONS TO NEUROIMAGING

BY SUBHADEEP PAUL[1] AND YUGUO CHEN[2]

[1]*Department of Statistics, The Ohio State University, paul.963@osu.edu*
[2]*Department of Statistics, University of Illinois at Urbana-Champaign, yuguo@illinois.edu*

To analyze data from multisubject experiments in neuroimaging studies, we develop a modeling framework for joint community detection in a group of related networks that can be considered as a sample from a population of networks. The proposed random effects stochastic block model facilitates the study of group differences and subject-specific variations in the community structure. The model proposes a putative mean community structure, which is representative of the group or the population under consideration but is not the community structure of any individual component network. Instead, the community memberships of nodes vary in each component network with a transition matrix, thus modeling the variation in community structure across a group of subjects. To estimate the quantities of interest, we propose two methods: a variational EM algorithm and a model-free "two-step" method called Co-OSNTF which is based on nonnegative matrix factorization. We also develop a resampling-based hypothesis test for differences between community structure in two populations both at the whole network level and node level. The methodology is applied to the COBRE dataset, a publicly available fMRI dataset from multisubject experiments involving schizophrenia patients. Our methods reveal an overall putative community structure representative of the group as well as subject-specific variations within each of the two groups, healthy controls and schizophrenia patients. The model has good predictive ability for predicting community structure in subjects from the same population but outside the training sample. Using our network level hypothesis tests, we are able to ascertain statistically significant difference in community structure between the two groups, while our node level tests help determine the nodes that are driving the difference.

1. Introduction: Networks and neuroimaging data analysis. Network analysis has received a plethora of multidisciplinary interest in the last few decades due to its various scientific and industrial applications in a variety of fields including genetics, neuroscience, ecology, economics and social sciences. A rapidly growing application area of network science is in the analysis of neuroimaging data, where it is used to analyze anatomical and functional connectivity among the brain regions (see Bullmore and Sporns (2009), Hutchison et al. (2013), Rubinov and Sporns (2010), Sporns (2014), Fornito, Zalesky and Breakspear (2013, 2015), Stam (2014), for reviews). In network neuroscience a typical approach is to construct functional brain networks based on measures of interregional associations obtained from various sources of measurements, including the functional magnetic resonance imaging (fMRI) with blood oxygen level dependent (BOLD) signals (Bassett et al. (2011), Simpson, Bowman and Laurienti (2013)). Modern neuroimaging experiments typically involve multiple subjects and/or multiple trials across the subjects. Various properties of the functional network (e.g., modularity, connectivity, degree, rich club organization, clustering coefficient,

Received April 2019; revised March 2020.
Key words and phrases. Community detection, neuroimaging, nonnegative matrix factorization, population of networks, random effects stochastic block model.

average path length, small-world property) are then investigated and contrasted among the subjects or groups of subjects (Bullmore and Sporns (2009), Van Den Heuvel and Pol (2010)). The intersubject and intergroup variations in many such network metrics have been related to cognitive ability and diseases in the literature (Bassett et al. (2011), Braun et al. (2015), Stevens et al. (2012), Hutchison et al. (2013), Jones et al. (2012), Wang et al. (2013), Braun et al. (2015), Yu et al. (2012), Lynall et al. (2010), Alexander-Bloch et al. (2012)).

A community or module of vertices in a network is defined as a group of vertices which are more connected among themselves than they are to the rest of the network. Many real world networks, including the functional brain networks, are known to exhibit community structure, whereby the vertices in the network can be roughly divided into such communities or modules. Several authors have uncovered intrinsic module structures in the functional organization of brain regions through network based analysis of spontaneous neuronal activity in resting state fMRI (He et al. (2009), Meunier, Lambiotte and Bullmore (2010), Power et al. (2011), Yu et al. (2011), Moussa et al. (2012)). The identified modules are consistent with several functionally connected subsystems generating spontaneous activities, for example, motor functions, auditory, visual, attention and default mode.

However, due to physiological differences among the subjects, responses to changing environmental conditions and variations in imaging instruments, the measured networks and, consequently, the network community structures vary from subject to subject within a group of subjects or even from trial to trial within a subject (Stevens et al. (2012), Simpson et al. (2013), Moussa et al. (2012), Weber et al. (2013)). Often in neuroimaging studies, researchers are interested in studying brain functional connectivity patterns in two groups of subjects; one is a group of patients diagnosed with a certain condition, and the other group is healthy controls. In such studies jointly analyzing these multisubject networks and comparing between the two groups of networks is of interest (Zalesky, Fornito and Bullmore (2010), Ginestet, Fournel and Simmons (2014), Chen et al. (2015), Stam (2014), Fornito, Zalesky and Breakspear (2015)). In a typical experimental setup the subjects in the two groups serve as random samples from these two populations. Hence, to facilitate comparison of populations beyond those in terms of single-value network summary measures (e.g., modularity), it is important to build statistical models for a random sample of networks from a population of networks.

Unfortunately, most of the literature on networks deal with a single instance of a network. This is primarily due to the fact that network data collection usually involves observing a network at one time point or tracking its evolution over time. However, modern application of networks in neuroscience brings a unique challenge and opportunity in terms of multiple instances of interactions among the same set of nodes through measurements on multiple subjects. The central question then is how to quantify the uncertainty in community structure due to subject specific variations within a group of subjects, so that two groups (populations) of subjects can be statistically compared.

Simultaneously, the problem of statistically testing for differences in community structure has received considerable attention recently in the neuroimaging literature (Alexander-Bloch et al. (2012), GadElkarim et al. (2012), Fujita et al. (2014), Glerean et al. (2016), Kujala et al. (2016)). These papers argue that differences between populations might not be well captured using a single network property measure like modularity, but it might be more meaningful to look at some measure of how different the module structures in the populations are. A permutation test using the average of Normalized Mutual Information (NMI) between pairs of network community assignments was proposed in Alexander-Bloch et al. (2012). GadElkarim et al. (2012) used a nodewise community consistency measure between two community partitions, called "Scaled Inclusivity (SI)" and defined in Steen et al. (2011), to assess how consistent a node's module is in a subject network with a mean module assignment for the whole group (obtained from community detection in the mean connectivity matrix). GadElkarim

et al. (2012) then proposed statistical hypothesis tests based on the SI vectors for the two groups. Glerean et al. (2016) first obtained a consensus partition across the group using a "clustering of clusters" technique similar in function to module allegiance method in Braun et al. (2015) and then determined SI of nodes across subjects with this consensus. The more general problem of hypothesis testing involving networks in functional neuroimaging has also been addressed in the literature (Ginestet, Fournel and Simmons (2014), Ginestet et al. (2017), Narayan and Allen (2016), Chen et al. (2015)).

We approach these problems by developing methodology in the spirit of random effects models, which will help us separate systematic variations between two populations of networks from variations due to subject specific "noise." In particular, we propose a random effects stochastic block model (RESBM) parameterized by a putative mean community assignment matrix, a transition probability matrix and appropriate block parameters. We develop two estimation strategies to estimate the parameters and other quantities of interest, a variational EM algorithm and a model-free two-step approach based on nonnegative matrix factorization (NMF). Finally, we develop resampling based two-sample hypothesis tests to compare two populations of networks in terms of their network level and node level community structures.

The rest of the article is organized as follows. In Section 2 we describe the application to schizophrenia data. In Section 3 we describe the random effects stochastic block model, the estimation strategies and the two-sample hypothesis tests. Section 4 studies the estimation and inference methods in simulated networks under several scenarios and several metrics. In Section 5 we present our results on the schizophrenia data. Section 6 provides a brief discussion.

2. Application to a resting state fMRI study on schizophrenia: COBRE dataset. We apply the methods developed in this article to a publicly available dataset from resting state fMRI experiments performed on subjects diagnosed with schizophrenia along with healthy controls. Network analysis is a key tool employed in analysis of functional connectivity in fMRI based experiments on schizophrenia (Lynall et al. (2010), Liu et al. (2008), Bassett et al. (2012), Yu et al. (2012), van den Heuvel et al. (2010, 2013), Alexander-Bloch et al. (2010, 2012)). Analysis of modular organization (community structure) of brain networks is of particular importance in understanding schizophrenia, since it has been hypothesized that schizophrenia is associated with neurodevelopment and evolution of brain which, in turn, is influenced by modular organization (see Yu et al. (2012) and references therein).

The dataset we analyze is the COBRE dataset publicly available to download from International Neuroimaging Data-sharing Initiative (INDI, 1000 Functional Connectomes project, http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html), that consists of anatomical magnetic resonance and resting state functional magnetic resonance scans from 72 patients diagnosed with schizophrenia and 75 healthy controls with ages ranging from 18 to 65 years. A detailed description of experimental conditions and equipment is available in the aforementioned webpage.

We used an automated preprocessing pipeline (Statistical Parametric Mapping's (SPM) default preprocessing pipeline for volume-based analyses (Penny et al. (2011))) implemented in Matlab toolbox CONN (Whitfield-Gabrieli and Nieto-Castanon (2012)) with parameters similar to earlier studies in Lynall et al. (2010) and Bassett et al. (2012). In particular, the steps included, deleting first three volumes, correcting for head motion by functional realignment and unwarping, functional slice timing correction, functional and structural image coregistration, structural segmentation and normalization, functional normalization to the standard Montreal Neurological Institute (MNI) space, functional outlier detection and spatial smoothing with a Gaussian kernel with 6 mm full width at half maximum (FWHM). Temporal filtering was performed with a high-pass filter with cutoff at 0.008 Hz. The regions of interest

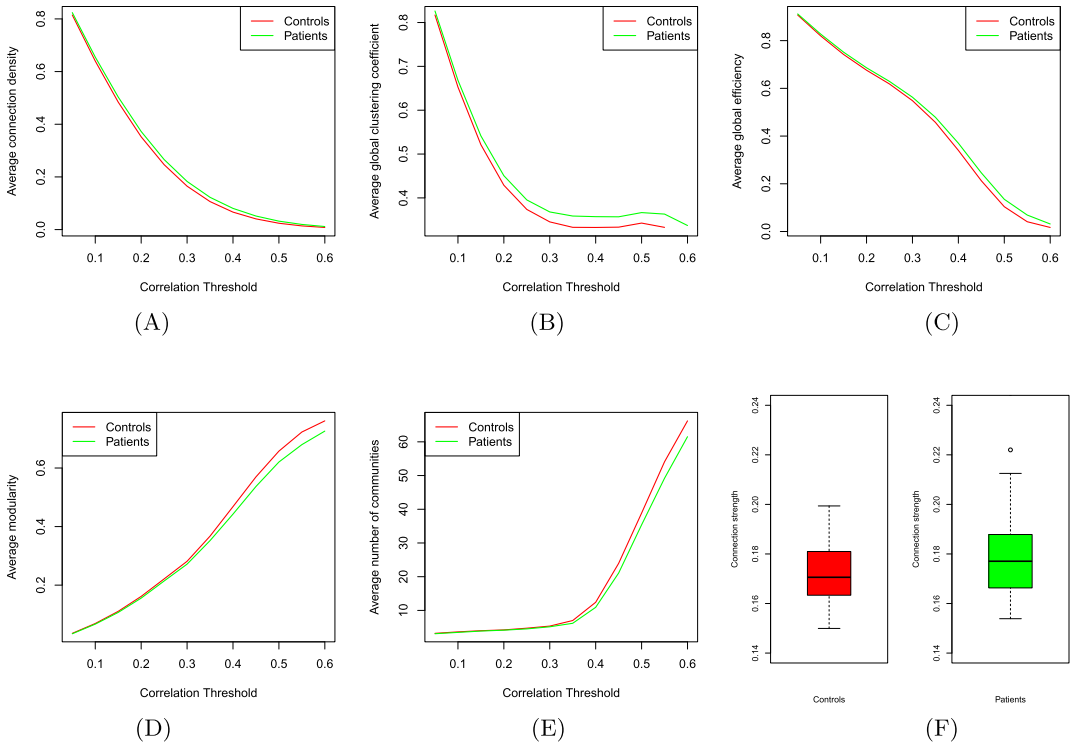


FIG. 1. *Exploratory analysis* 1: *Average values of global network summary measures with varying thresholds* (*A*) *network density*, (*B*) *global clustering coefficient*, (*C*) *global efficiency*, (*D*) *modularity and* (*E*) *number of communities*. *A boxplot of average strength of absolute correlation* (*without thresholding*) *is displayed in* (*F*).

(ROIs) were determined from a whole brain image percellation into anatomically defined regions described in the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al. (2002)). We exclude the cerebellum and vermis and concentrate on the remaining 90 cortical and subcortical ROIs, similar to previous studies (He et al. (2009), Lynall et al. (2010), Bassett et al. (2012)). Mean ROI time series was obtained for each of the ROIs by averaging the time series in the voxels within the ROI.

2.1. *Exploratory analysis*. To compute functional connectivity among the ROIs, we first decompose the mean time series in each ROI to a 4 scale maximal overlap discrete wavelet transform (Percival and Walden (2006)) and, then, take the second scale which roughly corresponds to the frequency range 0.060–0.125 Hz. A ROI to ROI correlation matrix is subsequently constructed from the pairwise correlations among this scale 2 wavelet transformed time series. The second scale of the wavelet transformation is chosen to strike a balance between minimizing the impact of physiological noise that might confound the higher frequencies and not having enough samples to compute correlation matrix in lower frequency ranges (Lynall et al. (2010), Bassett et al. (2012), Alexander-Bloch et al. (2012)). Binary connection matrices were created for each subject through thresholding at 12 levels between 0.05 and 0.60 with increments of 0.05, resulting in graphs of different network connection densities.

2.1.1. *Varying the thresholds*. Figure 1(A)–(E) compare the average values of a number of network global summary measures, namely, network density, global clustering coefficient, global efficiency, modularity and number of communities in the two populations, the controls and the patients over a range of threshold values. The modularity values and number of

TABLE 1

*Average modularity and average number of communities detected by modularity maximization (Spin-glass and Louvain algorithms) in the two groups of subjects for different thresholds. The columns of p-value indicate p-values obtained from Welch two sample t-test for difference in means. A * and ** indicate statistically significant at 5% and 1% FDR multiple comparison corrected levels*

Threshold	Modularity			Communities		
	Controls	Patients	p-value	Controls	Patients	p-value
0.05	0.0354	0.0339	0.0014**	3.20	3.08	0.0668
0.10	0.0689	0.0667	0.0013**	3.63	3.47	0.0875
0.15	0.1105	0.1067	0.0019**	3.98	3.87	0.2032
0.20	0.1615	0.1556	0.0028**	4.21	4.14	0.4581
0.25	0.2212	0.2141	0.0197*	4.72	4.52	0.0710
0.30	0.2814	0.2717	0.0483*	5.35	5.14	0.1681
0.35	0.3673	0.3520	0.0385*	7.00	6.18	0.0007*
0.40	0.4681	0.4425	0.0262*	12.40	10.91	0.0291
0.45	0.5697	0.5368	0.0329*	23.95	21.08	0.0414
0.50	0.6576	0.6206	0.0282*	38.86	35.38	0.0903
0.55	0.7229	0.6798	0.0116*	54.09	49.21	0.0291
0.60	0.7609	0.7261	0.0373*	66.21	61.52	0.0250

communities were detected from modularity maximization using spin-glass (Reichardt and Bornholdt (2006)) and Louvain algorithms (Blondel et al. (2008)) applied to the individual subject networks. Generally, as we observe, the patient group on average had a higher network density, higher global clustering coefficient, higher global efficiency and lower modularity as compared to the control group across thresholds. Figure 1(F) presents a boxplot of the distribution of strength of absolute correlation across the subjects in control and patient groups, respectively. This measure is sometimes reported in the literature as connection strength. The higher efficiency and lower modularity imply the networks in patients are more integrated, as opposed to controls where the networks are more modular and, hence, segregated (Rubinov and Sporns (2010)). While this observation is in agreement with some previous studies on schizophrenia, it is also in contrast to some findings (Yu et al. (2012)). We also see higher clustering coefficient in the patients, but this observation could also be because of higher network density. Since we thresholded all correlation matrices at the same value, the resulting binary networks may have different densities. The optimal number of communities (as detected by modularity optimization methods) appears to be the same for both groups until the threshold of 0.35, after which the number of communities increases exponentially for both groups and there is some difference in the number between the two groups (Figure 1(E)). Our observation on reduced modularity yet similar number of communities is in agreement with most previously reported studies in schizophrenia (Yu et al. (2012)).

For each of these thresholds, Table 1 presents the average of modularity values, the number of communities detected and p-values from Welch two sample t-test for difference in means. We note that the modularity in patients is consistently lower than in controls across all thresholds. The p-values indicate that modularity in patients is significantly lower at a 5% False Discovery Rate (FDR, Benjamini and Hochberg (1995)) corrected significance level for all thresholds, while it is lower at 1% statistical significance level at the first four thresholds. However, the number of communities detected is not significantly different between the two groups at the 5% FDR corrected significance level for any threshold value except 0.35.

In Figure 2 we assess the differences between the two populations in terms of distribution of modularity and number of communities through boxplots and histogram plots, respectively, at thresholds 0.1, 0.2 and 0.3. We note that at the threshold of 0.2, which we will

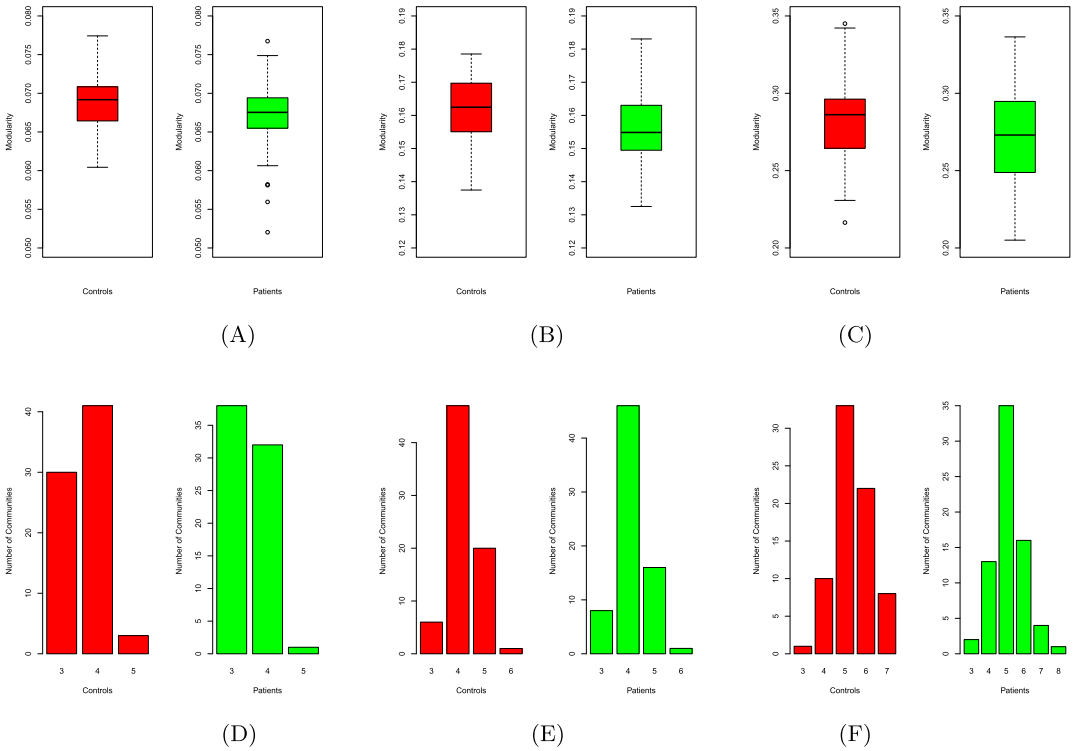


FIG. 2. *Exploratory analysis* 2: (*A*)–(*C*) *boxplots of distribution of maximum modularity across subjects for three different thresholds*—0.1, 0.2 *and* 0.3. (*D*)–(*F*) *histogram plot of distribution of optimal number of communities across subjects for three different thresholds*—0.1, 0.2 *and* 0.3.

primary focus on later in the main analysis, the boxplot of modularity for controls lies substantially above those of patients, giving an indication that the community structure might be quite different at this threshold. On the other hand, at this threshold the distribution of the number of communities appear almost identical with the most common number of communities being four in both cases.

2.1.2. *Removing the effects of covariates*. As a robustness check in our main analysis, we will also perform the analysis on a "residualized" correlation matrix obtained by removing the effect of some of the known covariates from the estimated correlations. For this purpose we run a linear regression of the correlations pulled together across subjects with a number of subject level covariates: age (continuous), gender (categorical with two levels, male and female) and handedness (categorical with three levels, left, right and both). The overall effect of the covariates were statistically significant (test for regression model: F statistic 68.38, p-value < 0.0001); however, the very low adjusted R squared (0.00022) indicates the effect of the covariates is rather low. Individually, all the covariates were found to be statistically significant. To asses the effect of the covariates in our analysis, we take the residuals from this regression and add the intercept term with it to create new residualized correlations. We repeat the above exploratory analysis with these residualized correlations. We do not find any significant differences in the global summary measures and, hence, omit those results. However, in Section 5.4 we run our methods on this residualized correlation matrix and compare the results with those obtained from unresidualized correlation matrix.

3. Models and methods. Suppose we observe a sample of M graphs or networks $\mathcal{G} = \{G^{(1)}, \ldots, G^{(M)}\}$ with a common set of nodes V, of size n and from a population of

networks. We assume the component graphs to be unweighted and undirected. Hence, to each component graph in the sample $G^{(m)}$, we can associate an $n \times n$ square and symmetric adjacency matrix $A^{(m)}$, such that $A_{ij}^{(m)}$ is 1 if there is an edge between nodes i and j in $G^{(m)}$ and 0 otherwise. For each component we can also define a normalized Laplacian matrix as $L^{(m)} = D^{(m)-1/2} A^{(m)} D^{(m)-1/2}$, where $D^{(m)}$ is a diagonal matrix whose elements are the degrees of the nodes in the mth component defined as $D_{ii}^{(m)} = \sum_j A_{ij}^{(m)}$. Our primary goal in this paper is to model the community structure of the population of networks from which the sample is generated.

A problem closely related to our problem is that of community detection in multilayer networks that has received considerable attention in the literature in the last decade (Kivelä et al. (2014), Nicosia and Latora (2015), Boccaletti et al. (2014), Paul and Chen (2016a)). A group of related interactions on the same set of nodes can also be represented as a multilayer network, where each network layer or type of edge represents a component network of the group. The multilayer networks observed in the nature are also known to exhibit community structure (Mucha et al. (2010), Bazzi et al. (2016), Kivelä et al. (2014), Nicosia and Latora (2015), Boccaletti et al. (2014), Peixoto (2015)). The multilayer stochastic block model (MLSBM) is a statistical model for such multilayer networks with community structure (Han, Xu and Airoldi (2015), Valles-Catala et al. (2016), Paul and Chen (2016a), Stanley et al. (2016), Peixoto (2015), Barbillon et al. (2017), Paul and Chen (2016b, 2020a)). Recently, Bacco et al. (2017) proposed a more flexible multilayer mixed membership stochastic block model that allows overlapping clusters. Most of the models described in the literature, with the exception of the strata-MLSBM of Stanley et al. (2016), are constrained by the fact that they assume the community structure to be the same across all layers. The estimation task is usually then to estimate this consensus community structure by fusing information from all layers.

However, in many situations, for example, in the multisubject neuroimaging studies, it may be desirable to model the variation in community structure in different subjects along with finding a consensus clustering. The existing models are not flexible enough to model such data. In a partial remedy of the situation, Stanley et al. (2016) introduced the strata-MLSBM where the community assignments vary across stratas but stay the same within the same strata. Within a strata, they further constrain the block model probability matrix to be identical across layers. A Bayesian nonparametric mixture model for jointly estimating community structure and identifying groups of networks with similar community structure in a collection of exchangeable networks was proposed in Reyes and Rodriguez (2016). A model similar in spirit to our proposed model is the Bayesian hierarchical mixed membership stochastic block model for an ensemble of networks proposed in Sweet, Thomas and Junker (2015); however, the MCMC estimation method is more computationally expensive and difficult to apply for larger networks and no hypothesis testing procedure was provided.

In this paper we propose a random effects stochastic block model (RESBM) to model the community structure of a population of networks. The RESBM is a general and flexible modeling framework which contains the MLSBM as a special case. The model assumes the existence of a putative mean community structure which is a group level parameter representative of the group or the population, but is not necessarily the actual community structure in any of the subjects in the sample. The true community assignment matrices for each of the component networks are random variables generated from this putative mean community assignment through the transition probability matrix. We formally define the model in the next section.

3.1. *Random effects stochastic block model.* We define the n node k block RESBM as follows. To each node i of a network, we associate a k dimensional *community assignment*

vector x_i which takes value 1 in exactly one place and 0's everywhere else. The location of 1 in the vector indicates the community to which the node belongs. We call a matrix $X \in [0,1]^{n\times k}$ a *community assignment matrix* of the nodes of a network, if each row of the matrix is a community assignment vector for one of the n nodes in the network. Let $\bar{Z} \in [0,1]^{n\times k}$ be a *putative mean community assignment matrix* for a group or population of M networks. This is a fixed group or population level parameter. For each member network $G^{(m)}$ in the sample, each node i can randomly switch its community label from this putative mean community label independent of other networks and other nodes.

Formally we use the multinomial unit vector and a transition probability matrix to describe the random deviation from the putative mean structure. A unit vector, similar to the community assignment vector, is a vector whose components are all zeros, except for one component that is one. A k-dimensional random unit vector Y follows a multinomial distribution (some authors also call it multinoulli distribution to draw parallel with Bernoulli distribution) with parameters $N = 1$ and $\mathbf{p} = (p_1, \ldots, p_k)$, $\sum_{i=1}^{k} p_i = 1$, if for unit vector $\mathbf{u} = (u_1, \ldots, u_k)$,

$$P(Y = \mathbf{u}) = p_i \quad \text{if } u_i = 1 \text{ and } u_j = 0 \text{ for } j \neq i.$$

For each member network $G^{(m)}$, the community assignment matrix of the network $Z^{(m)}$ is drawn in the following way. For each node i, the vector of community assignments is

$$Z_i^{(m)} \sim \text{Multinomial}(1, \bar{Z}_i T), \quad i = 1, \ldots, n, m = 1, \ldots, M, \tag{3.1}$$

where $Z_i^{(m)}$ and $\bar{Z}_i$ are the ith row of $Z^{(m)}$ and $\bar{Z}$, respectively. Here, T denotes the $k \times k$ nonnegative transition probability matrix among the communities, with its diagonal elements being $\{\eta_1, \ldots, \eta_k\}$ and the off-diagonal elements in each row q summing to $1 - \eta_q$. The vectors $Z_i^{(m)}$ are independent for all i and m. However, clearly the elements of the vector, that is, $Z_{iq}^{(m)}, q = 1, \ldots, k$, are dependent since $Z_{iq}^{(m)} = 1$ mandates that $Z_{iq'}^{(m)} = 0$ for any $q' \neq q$.

We can also write the expectation and variance of the random vectors $Z_i^{(m)}$s as follows:

$$\begin{aligned} E\big[Z_i^{(m)}\big] &= \bar{Z}_i T = T_q, \\ \text{Var}\big[Z_i^{(m)}\big] &= \text{diag}(\bar{Z}_i T) - T^T \bar{Z}_i^T \bar{Z}_i T = \text{diag}(T_q) - T_q^T T_q, \end{aligned} \tag{3.2}$$

where q is the putative mean community label of node i, T_q is the qth row of T and $\text{diag}(X)$ for any vector X represents the diagonal matrix whose diagonal is the vector X. Here and throughout the paper, T in superscript denotes the matrix transpose. Consequently, $Z_i^{(m)}$ is a multinomial random unit vector with T_q as the vector of parameters (probabilities). This implies that, for each member network, a node i that belongs to the mean community q gets assigned to the same community as its mean community assignment with probability η_q and a different community with probability $1 - \eta_q$. While we do not put any restriction on the transition probability matrix, we note that the model is most interesting when η_q is large as compared to the remaining elements in the qth row. The model then can be interpreted in the context of multisubject networks as follows. While most nodes retain their putative group mean community memberships for the individual subject networks, a few randomly selected nodes change their memberships to another community according to probabilities from the transition probability matrix.

Given the community assignment matrices $\{Z^{(1)}, \ldots, Z^{(M)}\}$, the edges in the M member networks are independently generated following a Bernoulli distribution:

$$A_{ij}^{(m)} | Z^{(m)} \sim \text{Bernoulli}\big(P_{ij}^{(m)}\big), \quad i, j = 1, \ldots, n, m = 1, \ldots, M. \tag{3.3}$$

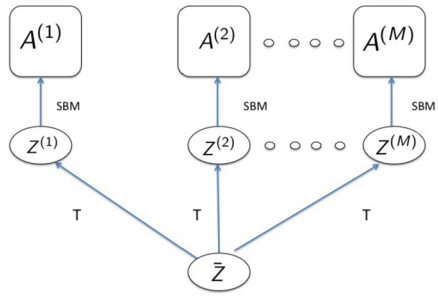


FIG. 3. *Schematic diagram of the RESBM.*

The Bernoulli probabilities can be modeled as a k class stochastic block model (SBM) or a k class degree-corrected stochastic block model with appropriate identifiability constraints. We focus only on the SBM in this paper. We have

$$A_{ij}^{(m)} | (Z_{iq}^{(m)} = 1, Z_{jl}^{(m)} = 1) \sim \text{Bernoulli}(\pi_{ql}^{(m)}), \quad q, l = 1, \ldots, k, m = 1, \ldots, M.$$

The model is schematically represented in Figure 3. However, the model is not identifiable without further constraints. Similar to the discussion in Matias and Miele (2017), in the context of dynamic networks the community labels might get switched between two member networks and still give the same model, leading to incorrect inference. Hence, we need certain constraints on the matrices $\{\pi^{(1)}, \ldots, \pi^{(M)}\}$ such that the communities are identifiable at all member networks. We use the constraint that the diagonal elements of the matrices are identical in each member network, that is, the vector $\{\pi_{11}^{(m)}, \ldots, \pi_{kk}^{(m)}\}$ is the same for all $m = 1, \ldots, M$ (Matias and Miele (2017)).

3.1.1. *Interpretation in fMRI studies.* Often in fMRI neuroimaging studies, there is interest in detecting a group community structure that is representative of the whole population, either a group of healthy control subjects or a group of patients with a certain condition, and then contrast such group community assignments between groups of interest. However, it is also widely recognized that all subjects in a group will not have the same community structure due to individual differences (Alexander-Bloch et al. (2012), Betzel et al. (2019)). The parameter $\bar{Z}$ denotes such an overall group community assignment. This is a "hard" community assignment where we designate a single community structure to be representative of the group. However, the consistency of this structure across subjects is estimated through the transition probability matrix T. Then $\bar{Z}T$ is the expected community assignment for a randomly selected subject from the group. This can be thought of as "mixed" or "overlapping" membership or "soft communities", with each row $(\bar{Z}T)_i$ providing probabilities with which the ROI i belongs to different communities. This is also our prediction for a new subject who is not part of the sample. Hence the parameter T controls how much variation is inherent in the group, that is, how much do we expect a randomly chosen subject to match the putative group mean. In addition, the estimate for $\text{Var}[Z_i^{(m)}]$ in equation (3.2) lets us derive confidence intervals for each ROI in the network. Thus, from this model we can infer an overall community assignment for the group of networks and obtain a quantification of the variability in this group putative community structure through the transition probability matrix.

3.2. *Estimation.* Our estimation goals from this model include estimation of community assignments in each member network $Z^{(m)}$, the putative mean community assignment matrix $\bar{Z}$ and the transition probability matrix T. In what follows we describe two methods to perform these estimation goals.

3.2.1. *A variational expectation-maximization estimator.* The first method we describe is approximate maximum likelihood estimation (MLE) through variational expectation-maximization (EM) algorithm, first introduced in the context of standard SBM in Daudin, Picard and Robin (2008), later extended to multilayer SBMs in Han, Xu and Airoldi (2015), Paul and Chen (2016a) and Barbillon et al. (2017) and to dynamic SBM in Matias and Miele (2017). The asymptotic consistency and limiting distribution of the parameter estimates for the case of standard SBM are investigated in Celisse, Daudin and Pierre (2012) and Bickel et al. (2013), respectively.

Below we derive the update rules for variational EM algorithm approximation to the MLE of the RESBM parameters. For this purpose we view the RESBM from a mixture model perspective and pose the parameters $\bar{Z}_i$s as random variables generated from a multinomial distribution with parameters $\alpha = \{\alpha_1, \ldots, \alpha_k\}$. We denote the unobserved variables $\bar{Z}$ and $Z^{(m)}$s together as X and the model parameters T, π and α together as θ. Further, we denote the observed log-likelihood of the model as $l(A, \theta)$. The complete data log-likelihood, which is the joint likelihood of the observed data and the unobserved model community assignment variables, is given by

$$\begin{aligned}
&\log P(A, X, \theta) \\
&\quad = \sum_{i=1}^{n}\sum_{q=1}^{k} \bar{Z}_{iq} \log \alpha_q + \sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{1\le q,l\le k} \bar{Z}_{iq} Z_{il}^{(m)} \log T_{ql} \\
&\qquad + \sum_{m=1}^{M}\sum_{1\le i<j\le n}\sum_{1\le q,l\le k} Z_{iq}^{(m)} Z_{jl}^{(m)} \{A_{ij}^{(m)} \log(\pi_{ql}^{(m)}) - (1 - A_{ij}^{(m)}) \log(1 - \pi_{ql}^{(m)})\}.
\end{aligned}$$

We define the variational distribution $R(X)$ to have the following form of the product of multinomial densities:

$$R(X) = \prod_{i=1}^{n}\prod_{q=1}^{k} \bar{\tau}_{iq}^{\bar{Z}_{iq}} \times \prod_{i=1}^{n}\prod_{m=1}^{M}\prod_{1\le q,l\le k} (\epsilon_{iql}^{(m)})^{Z_{il}^{(m)}\bar{Z}_{iq}},$$

where $\bar{\tau}$ and ϵ are the variational parameters. We delegate the remaining details of the derivation to the Supplementary Material (Paul and Chen (2020b)). The following fixed point equations are used to update the variational parameters iteratively in the Variational Expectation (VE) step:

$$\hat{\epsilon}_{iql}^{(m)} \propto \exp\left[\log(T_{ql}) + \sum_{j\ne i}\sum_{p=1}^{k} \tau_{jp}^{(m)} \log(b_{ijlp}^{(m)})\right], \tag{3.4}$$

$$\bar{\tau}_{iq} \propto \exp\left[\log \alpha_q + \sum_m \sum_l \epsilon_{iql}^{(m)} \log\left(\frac{T_{ql}}{\epsilon_{iql}^{(m)}}\right) + \sum_m \sum_l \sum_{j\ne i} \epsilon_{iql}^{(m)} \tau_{jl}^{(m)} \log(b_{ijql}^{(m)})\right], \tag{3.5}$$

$$\tau_{il}^{(m)} \propto \sum_{q=1}^{k} \bar{\tau}_{iq} \epsilon_{iql}^{(m)}. \tag{3.6}$$

For the M step we have the following closed form update steps:

$$T_{ql} \propto \sum_m \sum_i \bar{\tau}_{iq} \epsilon_{iql}^{(m)}, \qquad \alpha_q = \frac{1}{n}\sum_i \bar{\tau}_{iq}, \tag{3.7}$$

$$\pi_{ql}^{(m)} = \frac{\sum_{1\le i<j\le n} \tau_{iq}^{(m)} \tau_{jl}^{(m)} A_{ij}^{(m)}}{\sum_{1\le i<j\le n} \tau_{iq}^{(m)} \tau_{jl}^{(m)}}, \quad q \ne l, \tag{3.8}$$

$$\pi_{qq}^{(m)} = \frac{\sum_m \sum_{1\le i<j\le n} \tau_{iq}^{(m)} \tau_{jq}^{(m)} A_{ij}^{(m)}}{\sum_m \sum_{1\le i<j\le n} \tau_{iq}^{(m)} \tau_{jq}^{(m)}}.$$

The proportionalities in the above algorithm are turned into equalities through normalization using the constraints on the parameters.

3.3. *Two-step matrix factorization and maximum likelihood method.* While the variational EM proposed in the previous section attempts to estimate the unknown quantities in RESBM by approximately maximizing the model likelihood, there are two concerns with the approach. It can be computationally expensive, and, being a likelihood based method, it can be susceptible to misspecification of the model. Therefore, next we describe a two-step matrix factorization based approach to nonparametrically estimate some of the model quantities, namely, $\bar{Z}$ and $Z^{(m)}$s. We intuitively argue in the next section why this method is appropriate for those tasks under the RESBM despite not being based on the model.

The two steps are as follows:

1. Solve an optimization problem involving a joint objective function similar to equation (3.9) to obtain the community assignment matrices in the subjects $Z^{(1)}, \ldots, Z^{(M)}$ and the putative mean community assignment matrix $\bar{Z}$ simultaneously. This is a fully nonparametric step not dependent on any model and, hence, is expected to be less influenced by model misspecification.
2. The transition probability matrix T is obtained through conditional maximum likelihood (ML) conditioned on $\bar{Z}$ and $Z^{(m)}$s following equation (3.14).

3.3.1. *Coregularized orthogonal symmetric nonnegative matrix trifactorization algorithm.* We propose an algorithm for the first step of the two-step method based on orthogonal symmetric nonnegative matrix trifactorization (OSNTF) method, described for single networks in Paul and Chen (2016c). The proposed algorithm combines the ideas in linked matrix factorization (Tang, Lu and Dhillon (2009)) and coregularized spectral clustering (Kumar, Rai and Daume (2011)). Similar to the coregularized spectral clustering, the proposed method involves minimizing an objective function with two terms. The first term is the sum of Frobenius norm objective functions for single network OSNTF (Paul and Chen (2016c)), and the second term is a smoothness penalty on the factor matrices obtained at each network to make the subspaces spanned by the factor matrices closer to each other.

The global optimizer of the OSNTF objective function was shown in Paul and Chen (2016c) to consistently recover the community structure from networks generated from stochastic block model and degree-corrected block model. Hence, we expect the first part of the objective function attempts to recover the community structures of the subjects $Z^{(1)}, \ldots, Z^{(M)}$ from the individual subject network adjacency matrices. However, the second part, which is the smoothness penalty, attempts to find $\bar{Z}$, the group mean community structure such that the individual community structures are "shrunk" to this community structure. This penalty allows for both information sharing across subjects and a regularization to a group mean. Intuitively then, this method has similar goals as the RESBM and, therefore, is a reasonable nonparametric approach to estimate the group and subject specific community structures. Finally, we note that this method is also in the same spirit as other nonnegative matrix factorization based multiview learning methods proposed in the literature (Liu et al. (2013), Mankad and Michailidis (2013)). However, the orthogonality constraints on the factor matrices, despite being restrictive, makes the factors sparse which is more appropriate for the task of community detection.

We denote a matrix $U \geq 0$ and call it a nonnegative matrix, if all its elements are nonnegative, and $U > 0$, if all its elements are strictly positive. The method, which we call coregularized orthogonal symmetric nonnegative matrix trifactorization (Co-OSNTF), solves the following optimization problem on the collection of Laplacian matrices:

$$
\begin{aligned}
[\hat{U}^{(1)}, \ldots, \hat{U}^{(m)}, \hat{U}^*] = \underset{\substack{U^{(m)}, S^{(m)} \geq 0, \\ U^{(m)T} U^{(m)} = I, \forall m, \\ U^* \geq 0, U^{*T} U^* = I}}{\arg\min} & \sum_{m=1}^{M} \{ \| L^{(m)} - U^{(m)} S^{(m)} U^{(m)T} \|_F^2 \\
& + \lambda_m \big(k - \| U^{(m)T} U^* \|_F^2\big)\},
\end{aligned} \tag{3.9}
$$

where $U^{(1)}, \ldots, U^{(m)}, U^*$ are $n \times k$ nonnegative matrices with orthonormal columns and $\lambda_m \geq 0$ are user-chosen tuning parameters (scalars). Note that this is a constraint optimization problem on Stiefel manifold with additional nonnegativity constraints. To solve this optimization problem, we employ the method of Lagrange multipliers for the orthogonality constraints and then derive multiplicative update rules. The Lagrangian objective function with the orthogonality constraints incorporated is

$$
\begin{aligned}
\xi \triangleq \operatorname{tr}\Bigg(\sum_{m=1}^{M} \{ & -2 U^{(m)T} L^{(m)} U^{(m)} S^{(m)} + U^{(m)} S^{(m)} U^{(m)T} U^{(m)} S^{(m)} U^{(m)T} \\
& - 2\lambda_m U^{(m)T} U^* U^{*T} U^{(m)} + \Lambda_{U^{(m)}} U^{(m)T} U^{(m)} \} + \Lambda_{U^*} U^{*T} U^* \Bigg),
\end{aligned} \tag{3.10}
$$

where $\Lambda_{U^*} \geq 0$ and $\Lambda_{U^{(m)}} \geq 0$ are $k \times k$ nonnegative symmetric matrices of Lagrange multiplier parameters. Here and throughout the paper, $\operatorname{tr}(\cdot)$ denotes the matrix trace. The goal is now to minimize this new objective function under the constraints that $U^{(m)} \geq 0, U^* \geq 0, S^{(m)} \geq 0$.

To derive an algorithm for the optimization problem, we follow the derivation techniques for multiplicative updates described in Lee and Seung (2001), Ding et al. (2006) and Mirzal (2014). Briefly, the technique is as follows. The KKT conditions for the objective in (3.10) are

$$
\begin{aligned}
U^{(m)} \geq 0, \qquad & U^* \geq 0, \qquad S^{(m)} \geq 0, \\
\nabla\xi|_{U^{(m)}} \geq 0, \qquad & \nabla\xi|_{U^*} \geq 0, \qquad \nabla\xi|_{S^{(m)}} \geq 0, \\
U^{(m)} \odot \nabla\xi|_{U^{(m)}} = 0, \qquad & U^* \odot \nabla\xi|_{U^*} = 0, \qquad S^{(m)} \odot \nabla\xi|_{S^{(m)}} = 0,
\end{aligned}
$$

where $\nabla\xi|_{U^{(m)}}, \nabla\xi|_{U^*}, \nabla\xi|_{S^{(m)}}$ are gradients of the objective function (3.10) with respect to $U^{(m)}, U^*$ and $S^{(m)}$ and $\odot$ represents the Hadamard (elementwise) product. The equations in the last line of the KKT conditions are known as the complimentary slackness conditions which can be used to derive the multiplicative update rules (MUR). If the gradient $\nabla\xi$ with respect to a parameter can be written in the form $\nabla\xi = [\nabla\xi]^+ - [\nabla\xi]^-$, where $[\nabla\xi]^+, [\nabla\xi]^- \geq 0$, then the multiplicative update rule would be

$$
X \leftarrow X \odot \left(\frac{([\nabla\xi]^-)_{ik}}{([\nabla\xi]^+)_{ik}} \right)^{\eta},
$$

where $(\cdot)^\eta$ represents raising each element of the matrix to the power η, and $0 < \eta \leq 1$ is the learning rate. The inner division in the rightmost term is also elementwise. We take $\eta = 1/2$ as learning rate and prove the convergence of the resulting update rules to a stationary point of the objective function.

The details of the derivation are once again delegated to the Supplementary Material (Paul and Chen (2020b)). The algorithm consists of iteratively updating $U^{(m)}$s, $S^{(m)}$s and U^* according to the following update rules:

$$S^{(m)} \leftarrow S^{(m)} \odot \left(\frac{(U^{(m)T} L^{(m)} U^{(m)})_{ik}}{(U^{(m)T} U^{(m)} S^{(m)} U^{(m)T} U^{(m)})_{ik}} \right)^{1/2}, \tag{3.11}$$

$$U^{(m)} \leftarrow U^{(m)} \odot \left(\frac{(L^{(m)} U^{(m)} S^{(m)} + \lambda_m U^* U^{*T} U^{(m)})_{ik}}{(U^{(m)} U^{(m)T} L^{(m)} U^{(m)} S^{(m)} + \lambda_m U^{(m)} U^{(m)T} U^* U^{*T} U^{(m)})_{ik}} \right)^{1/2}, \tag{3.12}$$

$$U^* \leftarrow U^* \odot \left(\frac{(\sum_m \lambda_m U^{(m)} U^{(m)T} U^*)_{ik}}{(\sum_m \lambda_m U^* U^{*T} U^{(m)} U^{(m)T} U^*)_{ik}} \right)^{1/2}. \tag{3.13}$$

Note that λ_ms are user-specified parameters that should be chosen to address the user's preference in the trade-off between optimizing within a member network of the sample and increasing subspace cohesion as well as to reflect the relative importance of different members. The value of λ can also be chosen using cross-validation method to minimize some notion of prediction error. For our numerical experiments and real data analysis in this paper, we choose λ_ms to be 0.01 for all m. This choice of λ_ms works well in our synthetic experiments. In real data analysis in Section 5.7, we display a plot of out-of-sample prediction error as a function of λ, and we find 0.01 to be very close to the optimum. A similar observation was made for joint nonnegative factorization in Liu et al. (2013) for a number of different datasets.

In the Supplementary Material we prove convergence results and comment on some implementation issues. In particular, we prove two results that together imply that the update rules can reach a stationary point of the objective function ξ, provided the solution lies on the positive orthant of the feasible region (all elements of all matrices in the solution contain strictly positive entries), and we start with an initial solution which is in the positive orthant.

An alternative method to estimate the parameters in first step is the (centroid) coregularized spectral clustering method. The method was proposed in Kumar, Rai and Daume (2011); its theoretical properties under the MLSBM were studied in Paul and Chen (2020a). We propose to use this alternative method in conjunction with the conditional MLE described below and call the resulting method Co-Spectral. In our experiments we use the regularization parameters for Co-Spectral as 0.05 for all subject networks.

3.3.2. *The conditional maximum likelihood step*. The first step of the two-step method obtains $\hat{U}^{(1)}, \ldots, \hat{U}^{(m)}, \hat{U}^*$ which are matrices in the Grassmann manifold. In the second step we first obtain community assignments from each of the $\hat{U}^{(m)}$s and $\hat{U}^*$, and then the transition probability matrix is obtained from the estimated matrices through a conditional maximum likelihood (ML) estimator. For Co-OSNTF community assignment is performed by assigning each row to the community corresponding to the largest element in a row, while for coregularized spectral clustering this is accomplished through the k-means clustering of the rows of the matrices. Let $\hat{Z}^{(1)}, \ldots, \hat{Z}^{(M)}$ be the memberwise community assignment matrices and $\hat{Z}^*$ be the mean community assignment matrix obtained in this way. Then, the conditional ML estimate of T is given by

$$\hat{T}_{ql} = \frac{n_{ql}}{\sum_l n_{ql}}, \tag{3.14}$$

where $n_{ql} = \frac{1}{M} \sum_m \sum_i I\{\hat{Z}^*_{iq} = 1, \hat{Z}^{(m)}_{il} = 1\}$, with $I(\cdot)$ being the indicator function. We note that the same estimator for T can be obtained by equating the first moments as well.

3.4. *Two-sample hypothesis testing*: *Whole network-level and node-level tests*. We next develop procedures for performing a two-sample hypothesis test between network community structures of two populations of subjects. Let the two populations to be compared be denoted as A and B (in neuroimaging experiments these correspond to a group of healthy controls and a group of patients) and the sample size from each group be M_1 and M_2, respectively. Further, let the estimated model parameters for the two groups be $\bar{Z}_{(A)}, T_{(A)}, \{Z_{(A)}^{(1)}, \ldots, Z_{(A)}^{(M_1)}\}$ and $\bar{Z}_{(B)}, T_{(B)}, \{Z_{(B)}^{(1)}, \ldots, Z_{(B)}^{(M_2)}\}$. There are two notions of "group mean" of community structure giving rise to two natural ways of testing the difference between the two populations in modular organization. The first notion of "group mean" is $\bar{Z}$, the putative mean community assignment matrix of the group of networks. Accordingly, the notion of group mean at the node level is $\bar{Z}_i$ for node i. However, note that $\bar{Z}$ is not the expectation of the community assignment matrices $Z^{(m)}$s in the group. Instead, the expectation is $\bar{Z}T$. This implies that, for any node, its expected community assignment vector in any network within the group is $\bar{Z}_i T$. Hence, the second notion of "group mean" is represented by $\bar{Z}T$ at the network level and $\bar{Z}_i T$s at the node level.

The first network level test statistic we propose is the distance between $\mathcal{R}(\bar{Z}_{(A)})$ and $\mathcal{R}(\bar{Z}_{(B)})$, the subspaces spanned by the columns of $\bar{Z}_{(A)}$ and $\bar{Z}_{(B)}$, respectively. Note that the columns of both matrices, $\bar{Z}_{(A)}$ and $\bar{Z}_{(B)}$, span subspaces in the Grassmann manifold $\mathcal{G}(k, n)$ (Edelman, Arias and Smith (1999)). A common notion of distance between subspaces in the Grassmann manifold is in terms of norms of the matrix of sines of canonical angles between the subspaces (Stewart and Sun (1990), Edelman, Arias and Smith (1999), Dong et al. (2014)). Formally, we define the "Sine test" statistic as the Grassmann subspace distance between the putative mean community assignment subspaces,

$$S = \left\|\sin\left(\Theta\left(\mathcal{R}(\bar{Z}_{(A)}), \mathcal{R}(\bar{Z}_{(B)})\right)\right)\right\|_F^2 = \frac{1}{2}\left\|\bar{Z}_{(A)} Q_A \bar{Z}_{(A)}^T - \bar{Z}_{(B)} Q_B \bar{Z}_{(B)}^T\right\|_F^2,$$

where $\Theta(\mathcal{R}(\bar{Z}_{(A)}), \mathcal{R}(\bar{Z}_{(B)}))$ is the matrix canonical angles between subspaces spanned by the columns of $\bar{Z}_{(A)}$ and $\bar{Z}_{(B)}$, $Q_A = (\bar{Z}_{(A)}^T \bar{Z}_{(A)})^{-1}$ and $Q_B = (\bar{Z}_{(B)}^T \bar{Z}_{(B)})^{-1}$ are diagonal "scaling" matrices whose elements are the sizes of the communities.

The second test statistic corresponds to the second notion of "group mean," as described earlier. Following similar intuition as the Sine test, the network level "multinomial unit vector (MUV)" statistic is

$$\text{MUV} = \left\|\bar{Z}_{(A)} T_{(A)} T_{(A)}^T \bar{Z}_{(A)}^T - \bar{Z}_{(B)} T_{(B)} T_{(B)}^T \bar{Z}_{(B)}^T\right\|_F^2.$$

Note in this case we have dropped the diagonal scaling matrices that we had for Sine test. Large values of the test statistic will then indicate that the mean module structures in the two groups are markedly different. For each node i, define the node level MUV test statistic

$$\text{MUV}(i) = (\bar{Z}_{(A)i} T_A - \bar{Z}_{(B)i} T_B)^T (\bar{Z}_{(A)i} T_A - \bar{Z}_{(B)i} T_B).$$

Intuitively, $\bar{Z}T$ is a better measure of the group mean since it is the expectation of the subject's community assignment matrices. Consequently, we expect MUV test to be a better test than Sine test, since it takes into account the variation in community structure present in the population. Indeed, we observe the same in our simulations (omitted in the interest of space) and, henceforth, only use the MUV test.

In all the above cases it is difficult to derive asymptotic distribution of the test statistics. Hence, we construct p-value for the test through a permutation test based on resampling from the observed networks. We combine the network samples together and sample without replacement from the combined sample of $M_1 + M_2$ networks to create two samples of sizes M_1 and M_2, respectively. We fit the RESBM to both samples using variational EM and two-step methods and compute the Sine test and MUV test statistics in each case. We repeat

the procedure many times to construct the empirical distribution of the test statistic under the null hypothesis. Comparing the observed value of the test statistics with the constructed empirical distribution yields the p-values. For the node level tests we can perform the same procedures. However, when we make inference using the p-values, we need to account for multiple comparisons through a Family Wise Error Rate (FWER) or False Discovery Rate (FDR) correction (Benjamini and Hochberg (1995)).

4. Performance on simulated networks. In this section we numerically compare the performance of the proposed methods along with some already available or baseline methods in samples of networks simulated from the RESBM. We compare the performance under three metrics: the average performance in community detection across the members of the network sample $Z^{(1)}, \ldots Z^{(m)}$, the performance in estimating the putative mean community structure $\bar{Z}$ and the accuracy in estimating the transition probability matrix T. The normalized mutual information (NMI), an information theoretic measure of similarity between two community structures, is used to assess the accuracy of the estimated community assignments against the true assignments. The NMI between two community assignment vectors measures the degree to which one community assignment vector can be obtained from the knowledge of the other vector. It takes values between 0 and 1, where 0 indicates a random assignment (no overlap of information) and 1 indicates a perfect match, and the higher the value, the better the match. The accuracy of the estimated transition probability matrix is measured in terms of difference in Frobenius norm.

We generate the sample of networks from the RESBM in the following way. We first generate the group putative community labels $\bar{Z}_i$s from a multinomial distribution with k classes and equal probability for each class. To generate the community assignments of the member networks, the community labels of a fraction κ of nodes are then randomly changed to one of the communities other than its original community. Hence, the transition probability matrix T will have $1-\kappa$ as the diagonal elements, while the $k-1$ off-diagonal elements in each row sum to κ. We call the fraction κ "variation factor" since it is an indicator of how much variation there is in the community structure among the members of the sample.

For each of the members, the edges between the nodes are then drawn from a stochastic block model in the following fashion. We first generate the vector of k diagonal elements, which is common for all members (required for the model to be identifiable) as $\lambda \sim U(a, b)$, where $U(a, b)$ denotes the continuous uniform distribution with parameters a and b. Next, in each member the lower half of the $k^2 - k$ off-diagonal elements are generated from $U(a/\rho, b/\rho)$ while the upper half are identical to the lower half. The parameter ρ controls the signal to noise ratio (SNR), and in all our experiments we keep the value close to 2 in all members. The average density of the networks is controlled by another parameter called degree multiplier. Roughly speaking, increasing the degree multiplier by 1 corresponds to an increase of 2% of maximum degree in the degree density per network. As an example, if we have 500 nodes in a network, then a degree multiplier of 3 corresponds to average degree density per network being $500 \times 0.06 = 30$.

4.1. *Methods compared.* In our simulations we compare the performance of three algorithms from the two proposed methods along with a number of baseline and other available methods. In particular, the following methods are compared: (a) *VarEM*: The variational EM algorithm for computing approximate MLE in RESBM; (b) *Co-Spectral*: The two-step coregularized spectral clustering with conditional MLE; (c) *Co-OSNTF*: The two-step coregularized orthogonal nonnegative matrix trifactorization with conditional MLE; (d) *Ind. Spectral*: Spectral clustering in each member network performed independently (Rohe, Chatterjee and Yu (2011)). This method is used only for the comparison on performance in individual networks; (e) *Mean Spectral*: Spectral clustering of the mean adjacency matrix (Han, Xu and

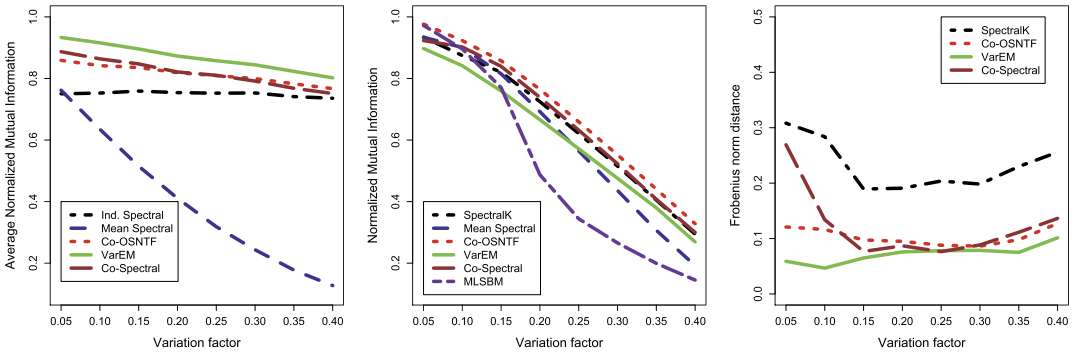


FIG. 4. *Performance of various methods across three metrics*: (*a*) *Average clustering performance across all member networks* (*in NMI*), (*b*) *performance in detecting the mean community structure* (*in NMI*) *and* (*c*) *accuracy in estimating the transition probability matrix* (*in Frobenius norm*) *with increasing variation factor from* 0.05 *to* 0.40. *The number of nodes is* 500, *the number of communities is three*, *the number of member networks is five and the average degree per networks is* 40.

Airoldi (2015), Paul and Chen (2020a)); (f) *SpectralK*: The spectral kernel method for mean community detection (Paul and Chen (2020a)) which is similar to the module allegiance matrix technique in Braun et al. (2015); and (g) *MLSBM*: The variational EM algorithm in MLSBM (Han, Xu and Airoldi (2015), Paul and Chen (2016a), Barbillon et al. (2017)). This method is used only for comparison in terms of putative mean community assignments.

In our comparisons we also include an estimate for the T matrix obtained by using Ind. Spectral in the individual networks and SpectralK for the mean community assignments. Some comments on the initialization and practical implementation of the methods are made in the Supplementary Material (Paul and Chen (2020b)).

4.2. *Increasing variation across members of the sample*. Our first simulation setup fixes n at 500, M at 5, k at 3 and the average degree in each network at 40 (which is about 8% degree density), while it varies the variation factor across the networks from 0.05 to 0.40, in steps of 0.05. Figure 4 shows the results of this simulation across the three aforementioned metrics of comparison.

We note that in terms of community detection in member networks, the performance of all methods, except Ind. Spectral, steadily falls as the variation factor increases (see Figure 4(a)). This is because with increasing variation factor, the networks are increasingly dissimilar and, hence, information sharing across networks does not improve performance as much as it does for low variation factor. The VarEM consistently outperforms all other methods in this metric, while the performance of Co-Spectral and Co-OSNTF trails closely. The Ind. Spectral does not share information across networks and, hence, is agnostic to variation factor. Consequently, although it gives inferior performance initially, its performance almost catches up with Co-Spectral and Co-OSNTF with increasing variation factor. Assuming the same community structure in all networks and using the community assignments obtained from Mean Spectral for all member networks gives considerably worse performance, especially when the variation factor is large, since the networks become very dissimilar.

For the putative mean community assignments all methods, except MLSBM, behave similarly (see Figure 4(b)). The two-step methods (Co-Spectral and Co-OSNTF) and SpectralK slightly outperform VarEM in this case. While the performance of all methods falls with increasing variation factor, the fall is slightly steeper for Mean Spectral as compared to the two-step methods. Finally, the estimate of T is most accurate with VarEM and stays almost

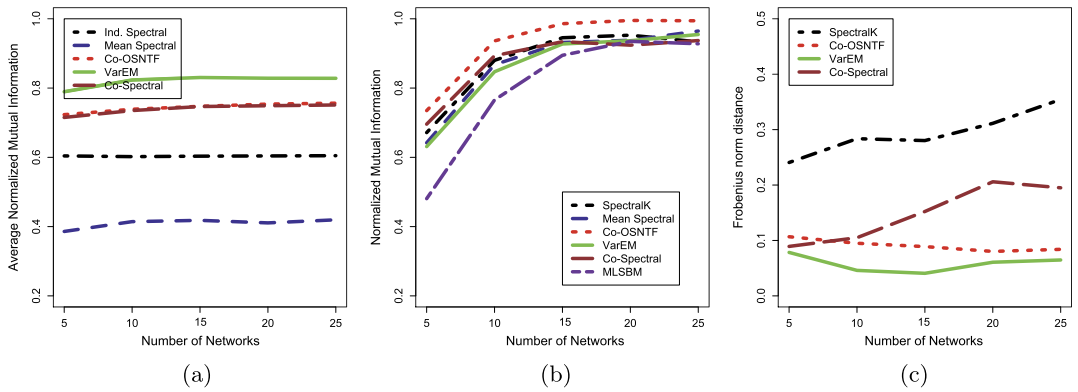


FIG. 5. *Performance of various methods across three metrics*: (*a*) *Average clustering performance across all member networks* (*in NMI*), (*b*) *performance in detecting the mean community structure* (*in NMI*) *and* (*c*) *accuracy in estimating the transition probability matrix* (*in Frobenius norm*) *with increasing number of member networks from five to* 25. *The number of nodes is* 300, *the number of communities is* 3, *the average degree per networks is* 25 *and the variation factor is* 0.20.

flat with increasing variation factor. The performance of Co-Spectral is poor, initially, at low variation factor but quickly improves as the variation factor increases. The performance of Co-OSNTF is slightly worse than that of VarEM and also stays flat with increasing variation factor. The estimate of T from the combination of Ind. Spectral and SpectralK (labeled as SpectralK in Figure 4(c)) performs poorly compared to the other three methods throughout.

4.3. *Increasing size of sample.* This simulation is to assess the effect of increasing the sample size, that is, the number of member networks in the sample on the performance of the methods. We fix n at 300, k at 3, the average degree per network at 25 (about 8% degree density) and the variation factor at 0.20, while we vary the number of networks from five to 25. The three plots in Figure 5 display the results on the three metrics of comparison. The VarEM outperforms all competing methods in terms of average performance in the individual networks (Figure 5(a)) and the accuracy of estimating the T matrix (Figure 5(c)). However, it underperforms in estimating the mean community assignments as compared to Mean Spectral, SpectralK, Co-Spectral and Co-OSNTF (Figure 5(b)). Among the two-step algorithms the newly proposed Co-OSNTF performs better than Co-Spectral algorithm in estimating both the mean community assignment and T matrix while remaining competitive with Co-Spectral in memberwise performance. The SpectralK algorithm is competitive to Co-Spectral in detecting the mean community structure. However, its memberwise counterpart, the Ind. Spectral, performs poorly in detecting the memberwise community structures. Consequently, the estimates of T, derived from a combination of Ind. Spectral and SpectralK, are also less accurate compared to other methods.

In the Supplementary Material (Paul and Chen (2020b)), we present another simulation where we vary the density of the member networks while fixing the variation factor, number of nodes, number of communities and number of member networks.

4.4. *Performance of hypothesis testing procedures on synthetic networks.* We next numerically compare the MUV tests using the estimates from the various methods developed in this article in a hypothesis testing problem on synthetic networks generated from the RESBM. We generate two samples of sizes 20 and 25 from a RESBM with 100 nodes, three communities and a T matrix whose diagonal elements are 0.8 and the off diagonal elements are 0.2. Hence, there is a 20% chance that a node will not retain its $\bar{Z}$ community in $Z^{(m)}$. The putative mean community assignment matrices for the two populations $\bar{Z}_A$ and $\bar{Z}_B$ are changed

TABLE 2

Network level tests: *p-values of various test statistics on two synthetic network samples of sizes* 20 *and* 25 *drawn from a* 100*-node, three-community RESBM. The columns represent results for different fraction of nodes that were changed to obtain the second* $\bar{Z}$ *from the first.* 10,000 *resamples were used to compute the p-values, a * indicates significant at* 1% *level*

Test	0	0.05	0.10	0.15	0.20	0.25	0.30
MUV test (VarEM)	0.6797	0.0013*	0.0000*	0.0000*	0.0000*	0.0000*	0.0000*
MUV test (Co-Spectral)	0.3437	0.0087*	0.0021*	0.0000*	0.0000*	0.0000*	0.0000*
MUV test(Co-OSNTF)	0.1316	0.0051*	0.0012*	0.0000*	0.0000*	0.0000*	0.0000*
MUV test (SpectralK)	0.2341	0.0037*	0.0061*	0.0000*	0.0000*	0.0000*	0.0000*

from being identical to being different in approximately 30% of nodes at intervals of 5%. A good test should not detect any difference either at network level or node level, when the samples come from the same population RESBM, but detect differences as the populations from which the samples are drawn differ. The distribution of all the test statistics under the null hypothesis was computed by resampling from the combined sample 10,000 times. The computation time was about 300 CPU core-hours (using HP X5650 2.66 GHz cores in a computing cluster) for each of the seven cases. Due to large number of function calls to the VarEM method, this test is computationally expensive; however, evaluation in the resampled datasets can be performed in parallel. The results from the network level tests are presented in Table 2 and those from the node level tests are presented in Table 3.

First we note from Table 2 that all the MUV tests can detect statistically significant difference at the 1% level when the $\bar{Z}$ matrices differ by 5% nodes. When the $\bar{Z}$ matrices differ by 10% or more nodes, all MUV tests give extremely small p-values. Given that each component network community structure is expected to differ from its group mean in about 20% of the nodes, the tests seem to be quite powerful. We also note that all tests correctly give large p-values when there is no difference between the $\bar{Z}$ matrices.

Table 3 presents the performance of the tests to detect node level differences. We correct for multiple comparison using a threshold of 0.05 FDR and present results for thresholds of 0.05 FWER and 0.10 FDR in the Supplementary Material (Paul and Chen (2020b)). All MUV tests correctly fail to detect any node-level differences when no nodes were changed. However, the procedures start and continue to make errors as the number of nodes changed increases from zero to 12 but then quickly reduce to almost no error as the number of nodes changed increases from 23 to 32. Among the MUV tests, both the VarEM based test and the Co-OSNTF based test perform particularly well. Moreover, the Co-OSNTF test detects all

TABLE 3

Performance of node level tests (*total errors and false positives*) *on testing between two samples of synthetic networks of sizes* 20 *and* 25 *drawn from a* 100*-node, three-community RESBM at* 0.05 *FDR threshold. "Number of nodes changed" is the actual number of nodes whose putative mean community assignment varied between the two groups. The best performance in terms of least total errors in each column is indicated in bold*

	False positives							Total errors						
Number of nodes changed	0	6	7	12	23	24	32	0	6	7	12	23	24	32
MUV test (VarEM)	0	0	2	1	1	5	2	0	**4**	**5**	**4**	3	5	2
MUV test (Co-Spectral)	0	0	0	1	0	1	1	0	6	7	10	**0**	1	1
MUV test(Co-OSNTF)	0	0	0	1	0	0	0	0	5	7	7	**0**	**0**	**0**
MUV test (SpectralK)	0	0	0	1	0	2	1	0	5	6	11	1	2	1

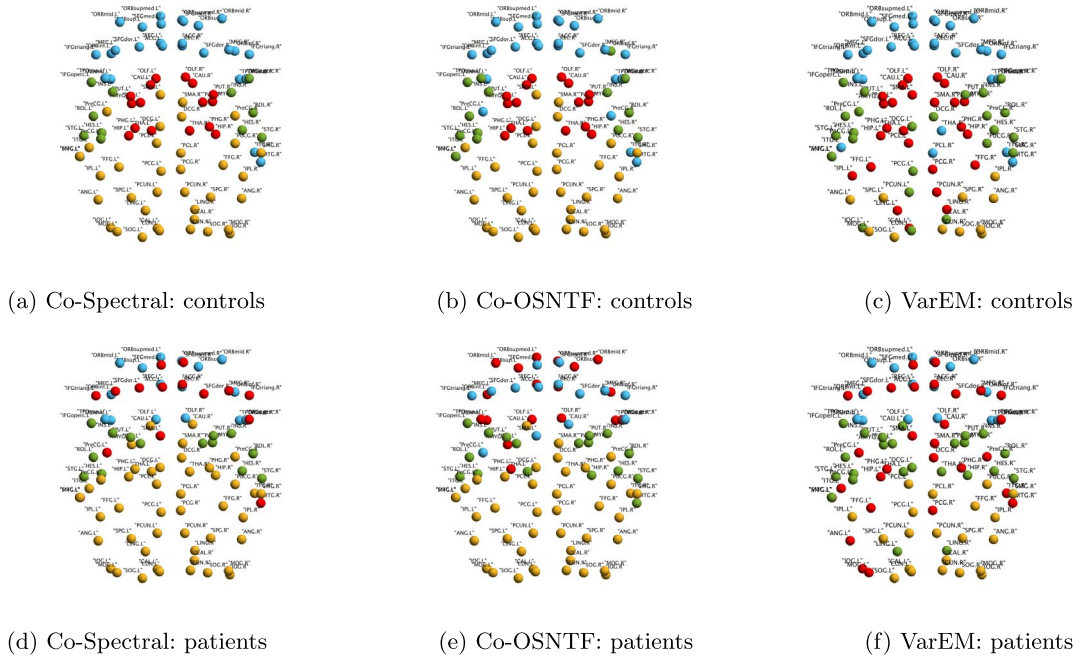
the node-level differences correctly when the number of nodes is 23 or more. However, the VarEM test suffers from increased false positives as the number of nodes changed becomes high, despite FWER and FDR controls. We note that switching from 0.05 FDR to 0.10 FDR does not increase the number of false positives much but improves performance, especially when the number of nodes truly changed is low (see Supplementary Material (Paul and Chen (2020b))). Based on the observations from the performance of the competing methods and test statistics on the synthetic networks, we recommend the MUV test with VarEM and Co-OSNTF as the two best performing tests.

5. Application of the methods to COBRE data. From the exploratory analysis in Section 2, it is clear that the control group has higher modularity values, as compared to the patient group, irrespective of our choice of threshold (and statistically significant at most of the thresholds). The average number of communities detected in each of the groups is not significantly different at any of the thresholds (Table 1). Both observations in terms of modularity and number of communities are consistent with previously reported results in the context of childhood-onset schizophrenia (Alexander-Bloch et al. (2012)).

We primarily focus our analysis on the threshold of 0.2 and, later in Section 5.8, we repeat some of our analyses for two other thresholds 0.3 and 0.4 as robustness checks. Note that the methods developed in this paper require the number of communities to be supplied as input. We obtain the number of communities in each case from the average number of communities detected in Table 1. For the threshold of 0.2, we observe that the average number of communities in both groups is about four. The histogram of number of communities in Figure 2 also shows four as the most common number of communities for subjects in both control and patient groups. Hence, we fit RESBMs with four communities to the networks from the control group and patient group subjects using both the VarEM and the two-step methods.

Although Alexander-Bloch et al. (2012) have shown that group differences in both modularity value and modular organization become more pronounced at higher values of the threshold, we use a relatively smaller value of threshold since, at higher values, the optimum number of communities is quite high which are difficult to interpret and visualize. For example, for thresholds between 0.5–0.6, the network divides into 30–60 communities (Table 1). In a network of 90 nodes, they are a really high number of communities, and the network is essentially disintegrated in very local submodules. Consequently, large differences are expected between the community structures of any two subjects, no matter which populations they are from and, therefore, a consistent picture of the modular disruption in schizophrenia cannot be obtained.

5.1. *Group differences in community structure*. The group putative community structure obtained for each group from each of the three methods is illustrated on a brain surface template in Figure 6. In addition, Figure 7 displays a visualization of the locations of the ROI groups, in a three dimensional eight-view layout of the brain. All visualizations have been produced using the BrainNet Viewer software (http://www.nitrc.org/projects/bnv/) (Xia, Wang and He (2013)). In Figures 6 and 7, the ROIs are colored according to their community membership detected in the respective group putative community structures $\bar{Z}$. In each of the three methods, the community labels between controls and patients are matched by maximizing overlap through an algorithm that solves the Linear Sum Association Problem (LSAP) (Papadimitriou (2003), Kuhn (1955)). This is necessary because community labels are recovered by any method only up to the ambiguity of label permutations, and, hence, two community partitions need to be aligned in terms of labels before they can be compared. We first note that community structure for the control group from Co-Spectral and Co-OSNTF almost agree with each other. The four modules obtained roughly correspond to the modules detected consistently across different subjects in resting state in Moussa et al. (2012)

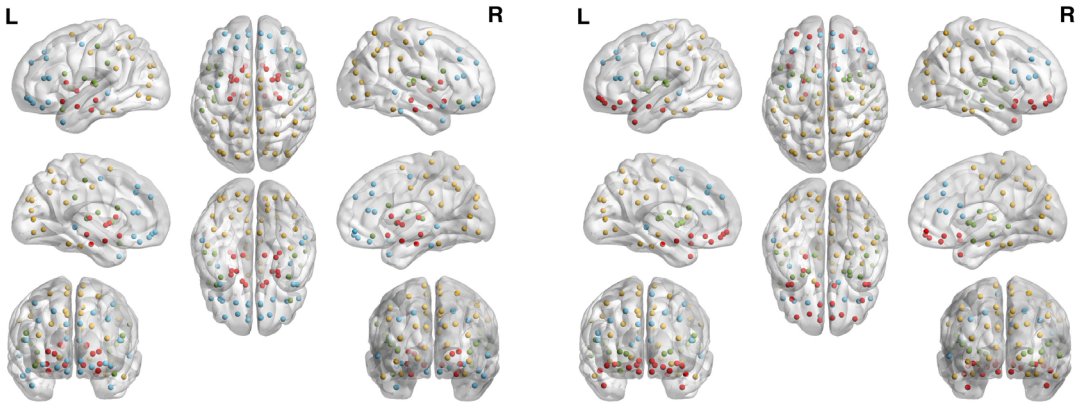


(a) Co-Spectral: controls (b) Co-OSNTF: controls (c) VarEM: controls

(d) Co-Spectral: patients (e) Co-OSNTF: patients (f) VarEM: patients

FIG. 6. *Group putative community structure of resting state network based on AAL ROIs in* (*a*), (*b*), (*c*) *healthy controls*, *and* (*d*), (*e*), (*f*) *patients with schizophrenia. Nodes are colored according to their group putative community obtained from the following methods*: *Co-Spectral for the first column* (*a*), (*d*), *Co-OSNTF for the second column* (*b*), (*e*) *and VarEM for the third column* (*c*), (*f*).

using a voxel level network approach on a very large scale dataset. Specifically, in the control group our modules blue (module 1), red (module 2), yellow (module 3) and green (module 4) have large overlaps with the default mode module, basal ganglia module, visual module and sensory/motor module, respectively, of Moussa et al. (2012). We refer the reader to Moussa et al. (2012) for a comprehensive list of modules detected in other works in the literature using both a network approach and an Independent Component Analysis (ICA) approach in resting state, to which these modules are equivalent or related to. The modules detected by the VarEM algorithm differs from the ones detected by Co-Spectral and Co-OSNTF; however, there is a large overlap. Later in Sections 5.4 and 5.5 on prediction and validation, we note



(a) Controls (b) Patients

FIG. 7. *A visualization of the group putative community structure in healthy controls and patients. The ROIs are colored according to the community memberships detected using Co-OSNTF method.*

TABLE 4
Nodes of Interest: *Nodes which are significant at* 0.1 *FDR correction*

Node	Uncorrected	FDR corrected	FWER
(A) Co-Spectral: Network level p-value: 0.0105			
Pallidum.L (PAL.L)	0.0004	0.0360	0.0360
Amygdala.R (AMYG.R)	0.0020	0.0774	0.1780
Pallidum.R (PAL.R)	0.0035	0.0774	0.3080
Putamen.L (PUT.L)	0.0039	0.0774	0.3393
Thalamus.L (THA.L)	0.0043	0.0774	0.3698
(B) Co-OSNTF: Network level p-value: 0.0042			
Caudate.R (CAU.R)	0.0013	0.0742	0.1170
Temporal.Pole.Sup.L (TPOsup.L)	0.0017	0.0742	0.1513
Hippocampus.R (HIP.R)	0.0031	0.0742	0.2728
Pallidum.L (PAL.L)	0.0033	0.0742	0.2871

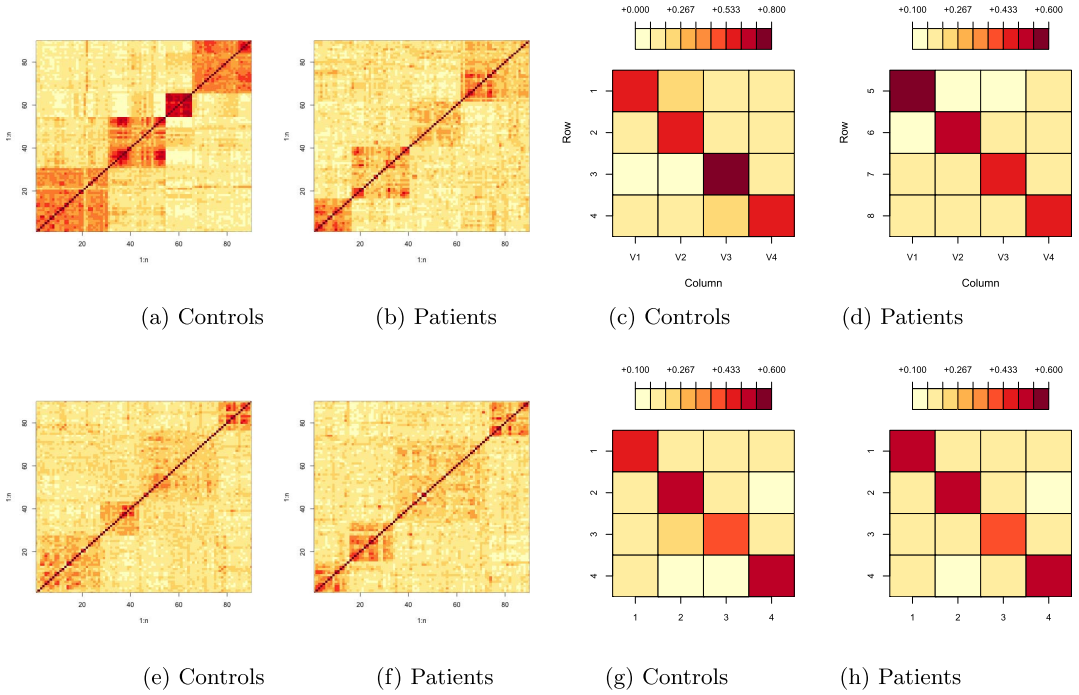
the performance of the VarEM algorithm is poor, compared to Co-OSNTF and Co-Spectral, for predicting community structure of held out subjects in this data. Therefore, we primarily focus on the Co-OSNTF method in interpreting the results.

For both Co-Spectral and Co-OSNTF the red group in controls gets disrupted in patients, and nodes belonging to the module get distributed to the yellow, green and blue modules in patients indicating disruption of the functional cohesion of these nodes. The blue module in controls also gets divided into two modules in patients, while the yellow and green groups remain relatively intact. Disruption of module structures has been reported in the literature on schizophrenia before; however, our methods provide visual identification of the disruption. In the subsequent analysis we investigate the disruption at the ROI level.

To statistically test the disruption in community structure, we compute the p-values for the MUV test with 10,000 permutation resamples. In both Co-Spectral and Co-OSNTF, the MUV test rejects the null hypothesis of no difference in the community structure between the controls and patients with p-values of 0.0105 and 0.0042, respectively (Table 4). The node level MUV test with Co-Spectral and Co-OSNTF found a number of ROIs to be statistically significantly altered at 0.1 FDR corrected p-values, which we call nodes of interest in Table 4. All the ROIs in cases of Co-Spectral and Co-OSNTF, except Temporal.Pole.Sup.L from Co-OSNTF, belong to the red module, as can be seen from Figure 6.

Our findings add to the growing evidence that module structure in brain functional networks gets disrupted in schizophrenia. Not only are we able to highlight the nature of the disruption at the global scale but also more locally infer the ROIs where the differences are the most.

5.2. *Consistency of community structure across subjects*. It is also important to understand the variability of the different modules in the group putative community structure across the different subjects within the group. In our model it can be assessed through the estimated transition probability matrices as well as a module consistency matrix that measures the fraction of subjects for which two nodes or ROIs are in the same community. The module consistency matrix is similar to the module allegiance matrix in Braun et al. (2015) and Scaled Inclusivity measure in Steen et al. (2011) and Moussa et al. (2012). Figure 8 displays the two measures for the results from VarEM (Figure 8(a)–(d)) and Co-OSNTF (Figure 8(e)–(h)) methods. In Figure 8(a), (b), (e) and (f) the ROIs are sorted according to their community label in increasing order (i.e., module 1 (blue module) in the bottom left corner and module 4 (green module) in the top right corner). In Figure 8(c), (d), (g) and (h), the 4×4 module



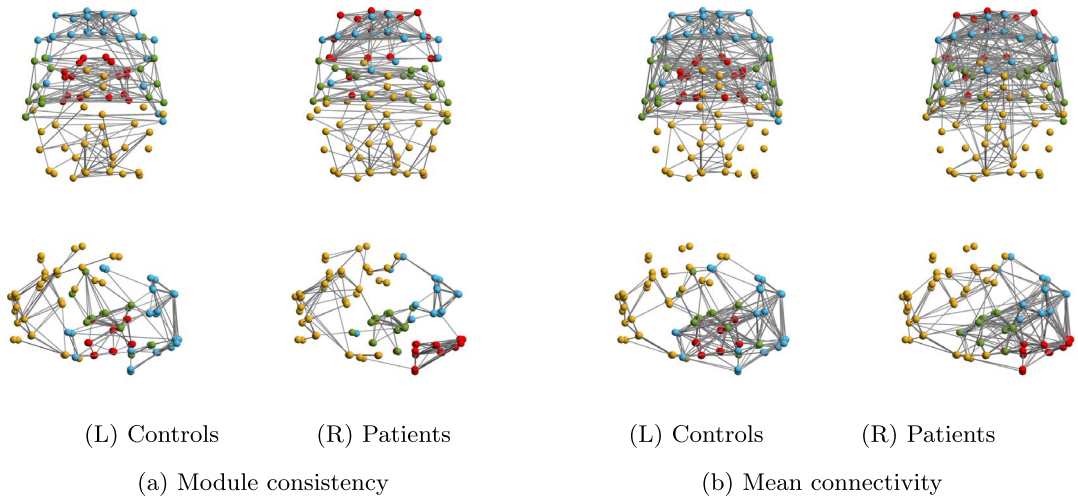
(a) Controls (b) Patients (c) Controls (d) Patients

(e) Controls (f) Patients (g) Controls (h) Patients

FIG. 8. *Stability of the group putative community assignments for VarEM* (*upper row* (*a*)–(*d*)) *and Co-OSTNF* (*lower row* (*e*)–(*h*)): (*a*), (*b*), (*e*) *and* (*f*) *are the matrices with elements as fraction of subjects for which two nodes* (*ROIs*) *are in the same community sorted according to the putative community structure for healthy controls and patients*, *respectively*; (*c*), (*d*), (*g*) *and* (*h*) *are the estimated transition matrices among modules for healthy controls and patients*, *respectively.*

transition matrix is plotted with the putative modules arranged from 1 to 4 along the row and column (i.e., module 1 (blue module) in the top left corner and module 4 (green module) in the bottom right corner). From Figure 8(a)–(d) the module structure appears to be more consistent in controls than in patients for VarEM. The patients show greater variability in the functional connectivity between any two regions, leading them to be classified into different modules more often than controls. This is possibly because of the variation in severity of the underlying conditions in the patient group. In particular, it can be seen from both the metrics in Figure 8 that the yellow module (module 3) is highly consistent in controls, as has been observed in many previous resting state studies (see Moussa et al. (2012) and references therein); however, in patients it is much less consistent. We do not see much difference in consistency of module structure between controls and patients using Co-OSNTF method (Figure 8(e)–(f)).

To better understand the community structure for the entire group, we visualize the structure obtained from Co-OSNTF algorithm in Figure 9 with the help of two measures: module consistency and mean connectivity. While the community structure is the same (the group putative community structure for controls and patients), the edges in Figure 9(a) represent the fraction of subjects for which two nodes are in the same community. The edges in Figure 9(b) represent the average connectivity between two nodes across all subjects. In both cases the edges are thresholded at a certain level (i.e., the edge appears if the quantity it represents exceeds a certain value) and are weighted, with thicker edges representing larger values. Similar to the observation in Moussa et al. (2012), we also have the visual module (yellow colored), containing both the primary and secondary visual cortices, as the most consistent module. We also observe that this module is relatively consistent in patients as well, confirming the observation from previous studies (Yu et al. (2012)). The red module in the control group,

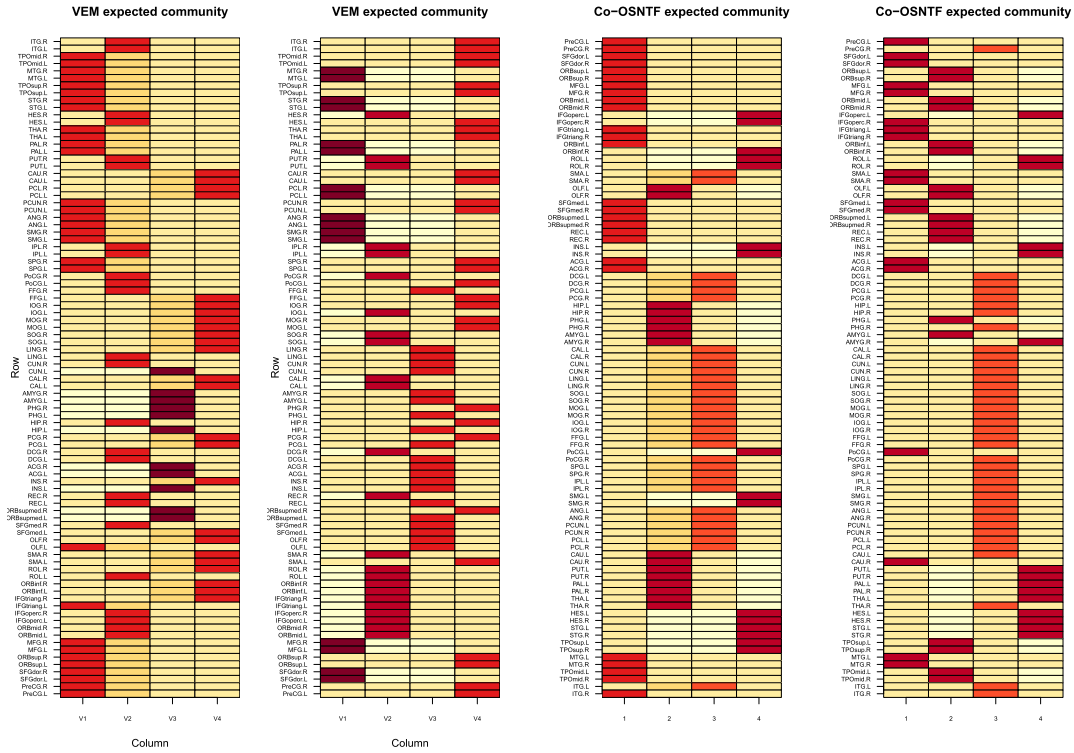


(a) Module consistency (b) Mean connectivity

FIG. 9. *Group putative community structure of resting state network from Co-OSNTF visualized through axial (top row) and sagittal (bottom row) brain views. In (a) the edges represent fraction of subjects for which two nodes are in the same community with a threshold of* 0.40, *while in (b) the edges represent the average connectivity (correlation) between two nodes across all subjects with a threshold of* 0.50. *In each case nodes are colored according to their group putative community in (L) Controls and (R) Patients.*

which roughly corresponds to the default mode network in Moussa et al. (2012), is split into two parts with some nodes being part of another module. From Figure 9(a) it is clear that the blue group in controls is split in two groups in patients (blue and red) which are almost disjoint in terms of module consistency thresholded at 0.40. The nodes in the red group in controls have lost the tendency to be grouped together in the patients, and, instead, they are more consistently grouped with different modules. The nodes that belonged to the yellow and green groups in controls appear to be unchanged in their comodule relation with the rest of the network.

5.3. *Estimates of expected communities of the ROIs.* We plot the estimated mean community assignments (mixed membership or soft assignments), that is, estimates for $\bar{Z}_i T$ for each ROI i for the control and patient groups in Figure 10. These mean estimates give us estimated probabilities that a given ROI will belong to one of the four communities for a randomly selected subject from the control group or the patient group. We note that, for both VarEM and Co-OSNTF, the community labels are separately aligned between controls and patients, and the module numbers are the same as the previous section. Therefore, each row of the four matrices in Figure 10 represents the estimated community "profile" for a ROI, with the colors being proportional to the estimated probabilities that the ROI belongs to a community. This enables us to visually detect the ROIs for which there are differences between the estimated community probabilities. For example, among the ROIs shown in Table 4(B) which have significantly altered expected community assignments using the node level MUV test, we note HIP.R has a very high probability of being in module 2 (red module) in controls and low probability of being in other modules, while it has a medium high probability of being in module 3 in patients and low probability of being in other modules. Similarly, CAU.R and PAL.L have high probability of being in module 2 (red module), low probability of being in module 4 and low to medium probability of being in modules 1 and 3 in controls. However, in patients their probability profiles are different. CAU.R has high probability of being in module 1, low-medium probability of being in the other three modules, while PAL.L has high probability of being in module 4, low probability of being in module 2 and low-medium

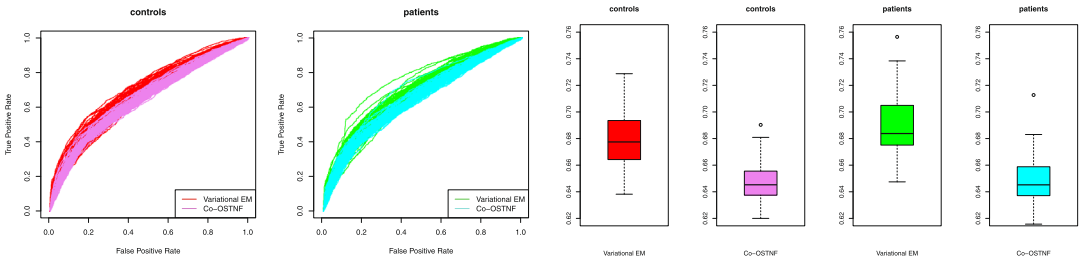


(a) Controls (b) Patients (a) Controls (b) Patients

FIG. 10. *Community estimate for AAL ROIs in controls and patients*: (*a*) *and* (*b*) *using the VarEM method*, (*c*) *and* (*d*) *using the Co-OSNTF method. The colors are proportional to the estimated probabilities that a ROI belongs to one of the four communities. The color scale goes from light yellow* (*lowest*) *to dark red* (*highest*).

probability of being in modules 1 and 3. The remaining ROI with a significant difference according to our node-level test is TPOsup.L, which gets a very high probability of being in module 4, low-medium probability of being in module 1 and low probability of being in modules 2 and 3 in the control group. In patients it restructures to get high probability of being in module 2, low-medium probability of being in modules 1 and 3 and low probability of being in module 4. We can also see how the probability profiles of ROIs with high probability in module 2 (red module in previous section) in the control group has got disrupted and changed in patient group. While the node level test did not detect them to be significant at the prescribed FDR level, we note the probability profiles of several ROIs, including HIP.L, PHG.R, AMYG.R, CAU.L, PUT.L, PUT.R, THA.L, THA.R, have changed between control and patient groups. We note that some of those ROIs have appeared in Table 4(A) as part of the ROIs detected by the Co-Spectral method. This is an evidence that Co-OSNTF and Co-Spectral detect somewhat consistent set of ROIs as regions where the module structure is disrupted in schizophrenia, even though the ROIs that are found to be statistically significant are not exactly the same. Taken together, Figure 10 gives a detailed picture and compelling evidence of the disruption and reorganization of module structure in schizophrenia.

5.4. *Prediction of community structure for new subjects and model fit.* Next, we study how the model fits the data and the predictive power of the model for new subjects. Assessing and comparing the methods in terms of their predictive ability (and model fit) is very important, especially since the methods lead to somewhat different results. In addition, good predictive ability of the model and the methods for new subjects give us confidence about the generalization of the results we have obtained in this study to the greater population.



(a) ROC curves (b) AUC values

FIG. 11. *(a) A plot of ROC curves for all subjects in control and patient groups and (b) a boxplot of distribution of AUC values across subjects.*

We first assess the in-sample model fit of the methods by plotting the Receiver Operating Characteristics (ROC) curves for each of the subject networks in the control and patient groups for VarEM and Co-OSNTF methods in Figure 11(a). The ROC is a plot between the false positive rate and true positive rate for fitting a binary data, where a high area under the curve (AUC) indicates a good model fit. A random guess will have a ROC curve close to the diagonal line and an AUC of 0.5. From the ROC curves it appears that there are a number of ROC curves of the VarEM method which are above the ROC curves of the Co-OSNTF method in both controls and patients. However, we also note that the ROC curves of the VarEM method have a greater variation. This can be seen more clearly from the boxplot of the AUC values across the subject networks in Figure 11(b). For both controls and patients the AUC values are generally higher using the VarEM method compared to the Co-OSNTF method; however, the spread in the AUC values is larger for VarEM method. We note the AUC values are in the range of 0.64 to 0.70, indicating a reasonably good model fit for both the methods. However, the VarEM method fits the data better in both control and patient networks.

While the ROC curves above give us indication of in-sample model fit, our primary tool of assessing the accuracy of the model and the methods is out of sample predictive ability of community structure. We perform a cross-validation by splitting both the control and the patient samples into two parts: a test data that holds out 15 samples from each of the groups and a training data consisting of the remaining subjects to which the model is fit. For both control and patient groups, we predict the community structure in the test data using our estimated community structure $\bar{Z}T$ from the training data. Then, in each of the subjects of the test data we estimate the community structure separately (independently) using two methods for single network community detection: (a) Spectral clustering with normalized Laplacian matrix and row normalization (Lei and Rinaldo (2015), Qin and Rohe (2013)) and (b) the OSNTF method (Paul and Chen (2016c)). The estimated community labels in each of the subjects are aligned with labels in the predicted community structure, by solving the LSAP problem as before. We assess the predictive performance using two metrics. Let the k dimensional community assignment vector for ROI i in test subject j be $u_i^{(j)}$. Then, the average community assignment for ROI i across the J test subjects is $\bar{u}_i = \frac{1}{J}\sum_j u_i^{(j)}$. Then, we compute a predictive accuracy by comparing this average community assignment with the predicted community assignment $(\bar{Z}T)_i$: $\text{median}_{ik}(|\bar{u}_{ik} - (\bar{Z}T)_{ik}|)$. This is our first metric for assessing predictive accuracy. In Figure 12 we display the boxplot of this predictive accuracy over 10 such train-test random sample splits for VarEM and Co-OSNTF methods for control and patient groups. We note that, for control group, Co-OSNTF method is more accurate with about 7–9% median absolute difference between the predicted and observed community assignments compared to VarEM with 9–12% median absolute difference. We make a similar observation in the patient group as well.

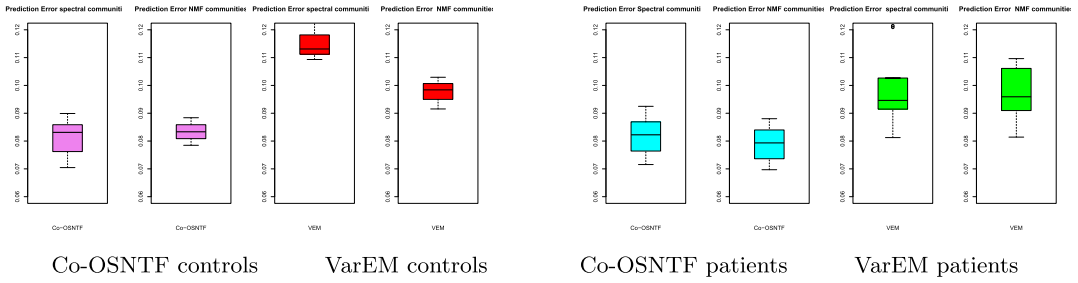


FIG. 12. *Predictive performance*: *Error of community prediction over* 10*-fold cross-validation using Co-OSNTF and VarEM methods in control and patient groups.*

The second metric considers the overlap of the mean putative community structure $\bar{Z}$ with the observed community structures in the test subjects again obtained using the spectral and OSNTF approaches. In Figure 13 we display the boxplot of the overlaps (correct classification rate) for VarEM and Co-OSNTF methods for control and patient groups. The correct classification rate for the Co-OSNTF method is better than that of the VarEM method. We note that in both control and patient group, the correct classification rate is around 0.40–0.44 for Co-OSNTF method which is substantially better than the random guessing rate with four communities of 0.25.

Finally, in Figure 14 we present two examples from each of the control and patient groups, where the sample has been divided into a test set comprising of 15 subjects and a training set comprised of the remaining subjects. For each of the examples (a)–(d), the leftmost figure displays the estimated community membership $(\bar{Z}T)_i$ for each ROI i using the Co-OSNTF method. This estimated community membership is also our prediction for a new subject. The middle and the rightmost figures then display $\bar{u}_i$, the average community membership for each ROI over the 15 test subjects obtained using the NMF and spectral methods for single networks. We note that in each case there is strong indication that $\bar{Z}_i T$ from the training sample helps us predict the average community memberships $\bar{u}_i$ from the test samples.

Taken together, the low absolute error in predicting the ROI community assignments and overlap of the actual community assignments with the putative mean assignment are testaments to the predictive value of the model and the methods. In particular, the Co-OSNTF method displays a superior predictive performance, and, therefore, we primarily present the results from Co-OSNTF method as our findings from this study. This is slightly surprising given the superior performance of the VarEM method in the simulations. However, since the data in the simulations were generated from the RESBM, it is not too surprising that the method which fits the model to the data performed better over a model-free approach. The model is based on SBM and as such inherits the limitations of the SBM, namely, a failure to

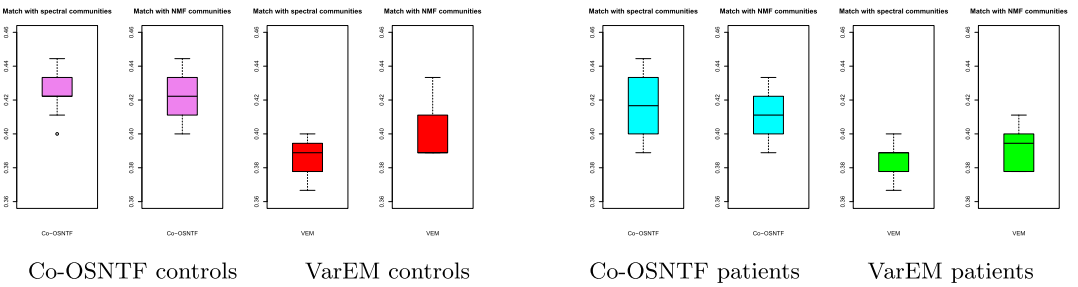
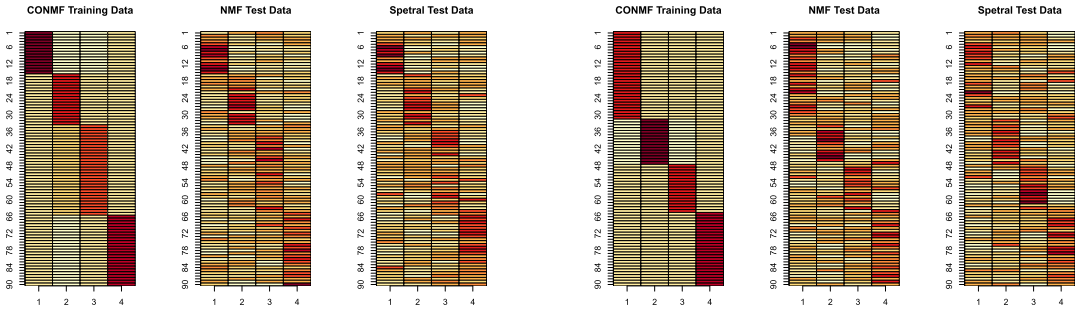
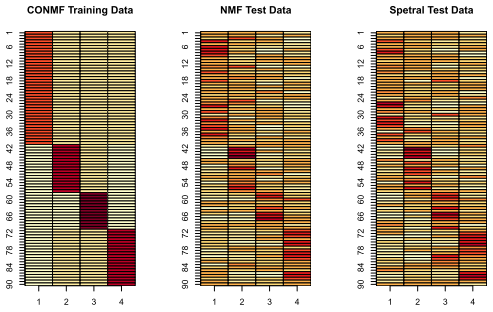


FIG. 13. *Predictive performance*: *Overlap* (*correct classification rate*) *with mean group putative community structure over* 10*-fold cross-validation using Co-OSNTF and VarEM methods in control and patient groups.*

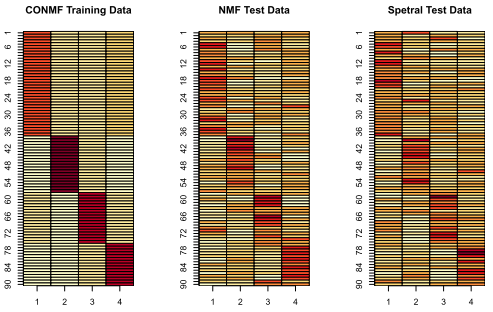


(a) Control group example 1

(b) Control group example 2



(c) Partient group example 1



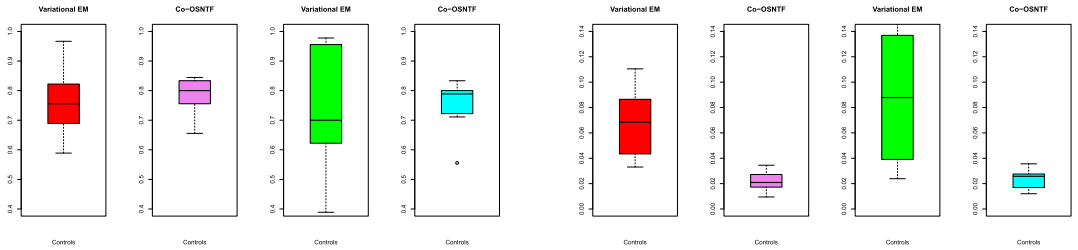
(d) Patient group example 2

FIG. 14. *Examples of predictive performance of Co-OSNTF method. In each case the data is split into a test set comprising of* 15 *subjects and a training set with the remaining subjects. The Co-OSNTF method is run on the training set to obtain estimated community assignments* $\bar{Z}T$ *for the ROIs* (*plotted in left panel in each case*). *The average of the community assignments for the ROIs obtained separately for each subject using an NMF and a spectral method are plotted side by side. The plots* (*a*) *and* (*b*) *are for control group*, *while* (*c*) *and* (*d*) *are for patient group.*

model degree heterogeneity. On the other hand, it has been shown in Paul and Chen (2016c) that the OSNTF method is consistent for community detection under degree heterogeneity (degree corrected stochastic block model) as well. The Co-OSNTF method likely inherits this property from OSNTF and is, therefore, robust against some of the limitations that SBM poses.

5.5. *Validation and robustness checks.* The predictive performance validates the use and value of our model and methods. Next, we further perform a split sample validation of the results where we split the sample into two roughly equal parts. Then, we estimate the parameters $(\bar{Z}, T)$ on the two parts separately using VarEM and Co-OSNTF methods. We compare the two sets of estimates. The $\bar{Z}$ estimates are compared using the correct classification rate, while the T estimates are compared using the median absolute difference between the corresponding elements. As before, the community labels in the two splits are aligned by solving a LSAP problem. We repeat this procedure 10 times, every time randomly splitting the sample into two parts. Boxplots of the correct classification rate for comparing $\bar{Z}$ estimates and the median absolute difference for comparing T estimates are presented in Figure 15. We find that there is an extremely large overlap with correct classification rate reaching 80–90% using Co-OSNTF (Figure 15(a)) and the median absolute difference between elements of T matrix being around 0.02–0.03 (Figure 15(b)). This validates that our estimates are consistent across splits of the sample.

In our first robustness check we fit RESBM to networks obtained by thresholding the residualized correlation matrices obtained by regressing out the effects of a number of co-



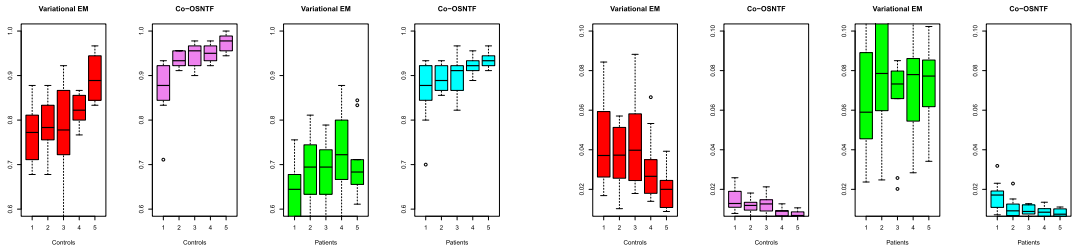
(a) comparing $\bar{Z}$ (b) comparing T

FIG. 15. *Split sample validation*: *The samples are divided into two parts randomly*; *Co-OSNTF and VarEM are fitted to each part. The* (*a*) $\bar{Z}$ *and* (*b*) T *estimates from the two splits are compared for control and patient groups*.

variates, as outlined in Section 2.1.2. We obtain $(\bar{Z}, T)$ estimates using Co-OSNTF method and compare those with the estimates obtained without residualizing. We find the estimate for $\bar{Z}$ matches for about 95% of the ROIs (a mismatch of four out of 90 ROIs) for both controls and patients. The median absolute difference for the estimate of T is 0.0063 and 0.0049 for controls and patients, respectively. This suggests the estimates obtained from residualized correlation matrices are not much different from the one obtained without residualizing.

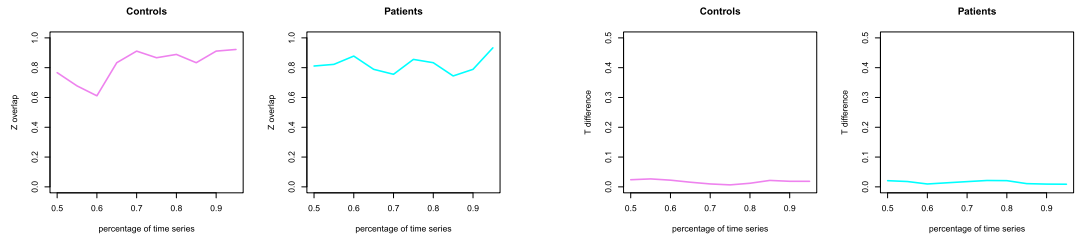
Our second robustness check involves sampling a subset of subjects in each group and fitting the model on the subset as if the remaining data did not exist. We then compare the $(\bar{Z}, T)$ estimates from this reduced dataset with those obtained from the full dataset. In Figure 16 we progressively sample (50%, 60%, 70%, 80%, 90%) of the data and compare the correct classification rate of $\bar{Z}$ estimate and median absolute difference of T estimate with the full data. In each case the boxplot represents the distribution of these values over 10 random samples. We note that the estimates for both $\bar{Z}$ and T from the reduced samples become closer to those obtained from the full sample as the size of the reduced sample increases for both controls and patients. Across different scenarios, Co-OSNTF method gives much better proximity to full sample results compared to VarEM. For the Co-OSNTF method, the $\bar{Z}$ estimate from the reduced sample has an overlap of 90% with 50% data, and it increases to almost 95% with increasing sample. On the other hand, the estimate for T has a median absolute error of 0.01 with 50% data and decreases even further to almost no error with 90% of the data. These results can be thought of as another evidence toward validation of the results and an important check for the robustness of our findings in the study.

The final robustness check involves availability of smaller time series data for estimating the correlation matrices. Clearly, the accuracy with which the correlation matrices are measured is a function of the sample size of the time series and, hence, is expected to have an impact in our analysis. In Figure 17 we start with roughly half the length of the time series



(a) accuracy of estimating $\bar{Z}$ (b) accuracy of estimating T

FIG. 16. *Robustness check*: *Overlap of results on full sample with results from a smaller subsample* (*a*) *accuracy of estimating* $\bar{Z}$ *using correct classification rate and* (*b*) *accuracy of estimating* T *using median absolute error.*



(a) $\bar{Z}$ overlap (b) Difference in T

FIG. 17. *Robustness check*: *Overlap between results from shorter time series with those obtained from the full time series*.

and increase the length progressively. We assess the similarity of the estimates for $\bar{Z}$ and T obtained from the shorter time series with the full series in Figure 17(a) and (b), respectively. It appears that the length of the time series, despite having some impact, does not drastically change our estimates. The $\bar{Z}$ estimates have strong overlap with the estimate from full sample, while the median absolute difference of the T estimates from that of the full sample is small.

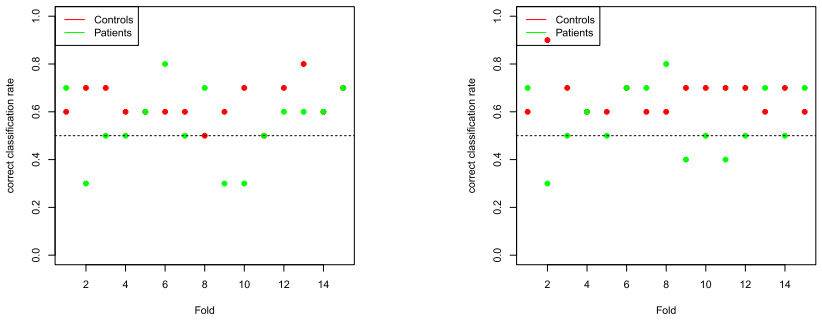
5.6. *Discrimination analysis with new subjects*. Next, we assess the utility of our model and Co-OSNTF method in classifying an unlabeled subject to one of the two groups: controls or patients. Classification of subjects in one of the groups is a fundamental interest in fMRI due to obvious diagnostic and clinical implications. Recently, there has been an enhanced interest in classifying subjects in schizophrenia, with many studies reporting high accuracy. We show that the proposed methods can uncover information that are useful in discrimination.

We use two discrimination functions, one based on the MUV statistic described earlier and the other one based on ratio of log-likelihoods. Suppose as before, u_i denotes a k dimensional community assignment vector for ROI i in a test subject obtained using either the spectral method or the OSNTF method. Note from equation (3.1) that the likelihood for ROI i in a test subject to belong to the population A is $u_i \log(\bar{Z}_A T_A)$. Therefore, the log-likelihood that the test subject belongs to population A is $\sum_i u_i \log(\bar{Z}_A T_A)$ due to the conditional independence assumption. Similarly, the log-likelihood that the subject belongs to population B is $\sum_i u_i \log(\bar{Z}_B T_B)$. Therefore, the log-likelihood based discrimination rule is to classify a subject to either population A or B based on whether $\sum_i u_i \log(\bar{Z}_A T_A) > \sum_i u_i \log(\bar{Z}_B T_B)$ or not. The MUV discrimination function is a simple variant of the MUV test statistic defined earlier,

$$\text{MUV} = \| \bar{Z}_{(A)} T_{(A)} T_{(A)}^T \bar{Z}_{(A)}^T - U U^T \|_F^2.$$

In Figure 18 we present accuracy of classifying subjects from a 15% holdout test set on the basis of estimates obtained from training data over 15 repetitions. In the majority of cases in both controls and patients, the accuracy of classification is above the random guessing line of 0.5. We note that the classification accuracy is low compared to recent reports. However, given this is a discrimination analysis and we do not train a classifier for the purpose of supervised classification, it is not so surprising.

5.7. *Computation time and choice of Co-OSNTF tuning parameter* λ. Next, we comment on the computing time each of our methods take on this data and the choice of the tuning parameter λ in the Co-OSNTF method. In Figure 19(a) and (b) we compare the computation time of VarEM, Co-OSNTF and Co-Spectral with increasing number of subjects from each of the control and patient groups. These computations have been carried out using a 4.2 GHz



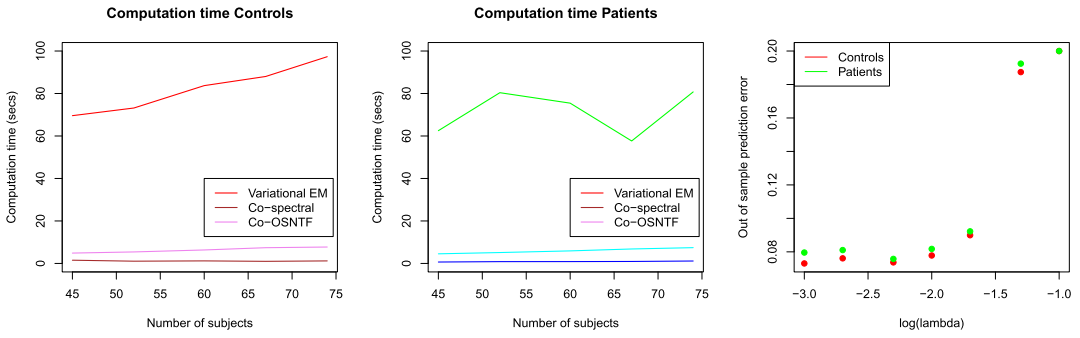
(a) Discrimination with Log-likelihood statistic (b) Discrimination with MUV statistic

FIG. 18. *Discrimination power of Co-OSNTF method. We use the* (*a*) *log-likelihood and the* (*b*) *MUV statistics to classify a new subject to one of the two groups.*

Intel Core i7, eight core, 32 GB memory system (however, no parallel computing is involved). Generally, the computing time increases as the number of subjects increases in both cases. We also note that both Co-Spectral and Co-OSNTF are faster compared to the VarEM method, with both of them computing their solutions for the whole dataset under 10 seconds, while VarEM requires around 60–80 seconds depending upon the number of subjects. While all the methods for estimation are reasonably fast, the hypothesis testing, however, is computationally intensive because of the requirement of running the methods on 10,000 resamples for an accurate estimate of the p-values (this step however is executed in parallel).

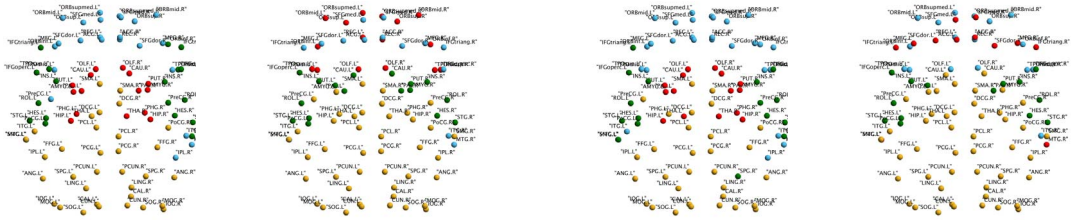
While the Co-OSNTF tuning parameter is a user given parameter, it is possible to choose the parameter using cross-validation by treating the prediction accuracy as a loss function. We consider a wide range of λ values between $[0.001, 0.1]$, fit Co-OSNTF with the λ values and compute the out-of-sample prediction error in each case. We display the cross-validation prediction error for both controls and patients with log of λ (base 10) in Figure 19(c). It appears that, for λ between 10^{-3} to 10^{-2}, there is not much difference in the prediction accuracy, and as λ increases the accuracy drastically decreases. We also note that the prediction error is least at a point somewhere between -2.5 and -2. Therefore, our choice of 0.01 for λ is very close to the optimum.

5.8. *Choice of a different threshold.* We close this section by repeating some of the analyses with two different thresholds for the correlation matrices, 0.3 and 0.4. We fit the model using the Co-OSNTF method with $k = 4$ communities in each case. The community colors



(a) (b) (c) CV λ

FIG. 19. (*a*) *and* (*b*) *Computation time in controls and patients*, (*c*) *Cross-validation for the choice of* λ.

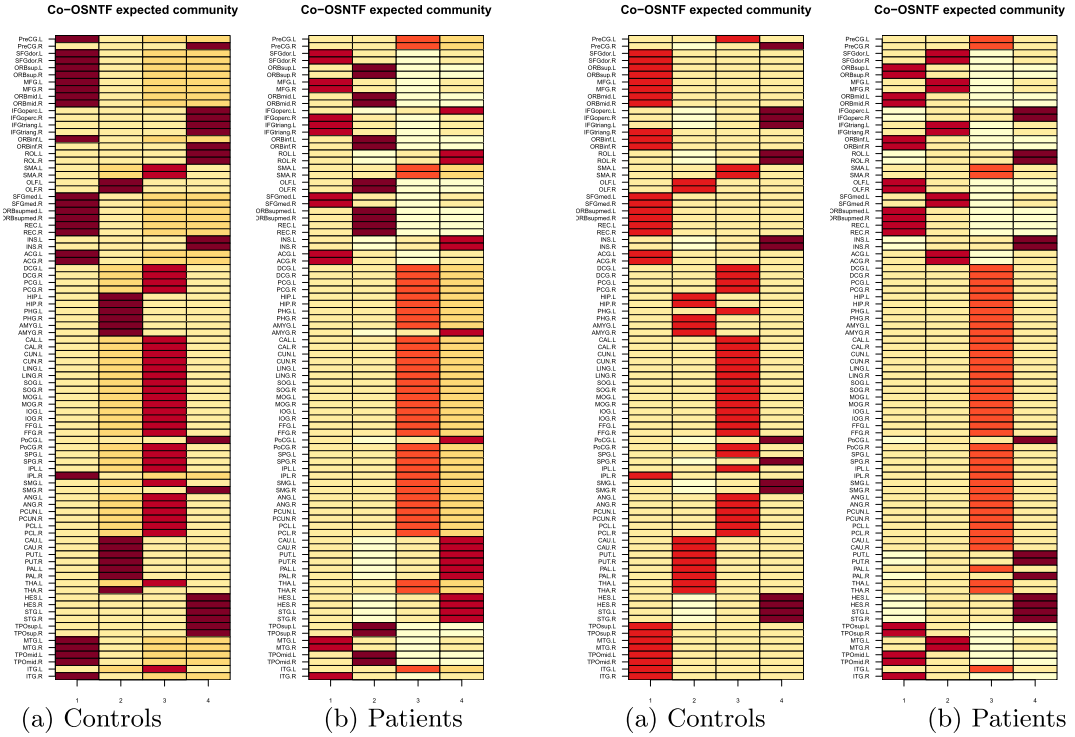


(a) Control (b) Patient (c) Control (d) Patient

(A) threshold 0.3 (B) threshold 0.4

FIG. 20. *Group putative community structure of resting state network based on AAL ROIs for thresholds of* (*A*) 0.3 *and* (*B*) 0.4 *in* (*a*), (*c*) *healthy controls*, *and* (*b*), (*d*) *patients with schizophrenia. Nodes are colored according to their group putative community obtained from Co-OSNTF fitted with four communities*.

and labels are kept consistent with previous analysis with threshold 0.2, for ease of comparison. The group putative community structure and the estimated community assignment of the ROIs are presented in Figures 20 and 21, respectively. We note that there is a very high similarity of the results with the ones we have obtained for the threshold of 0.2. We note, as before, a disruption in the community structure with nodes from red module in controls being distributed in various modules in patients, the blue module in controls getting divided into two modules in patients, and the yellow module remaining relatively unchanged between controls and patients.



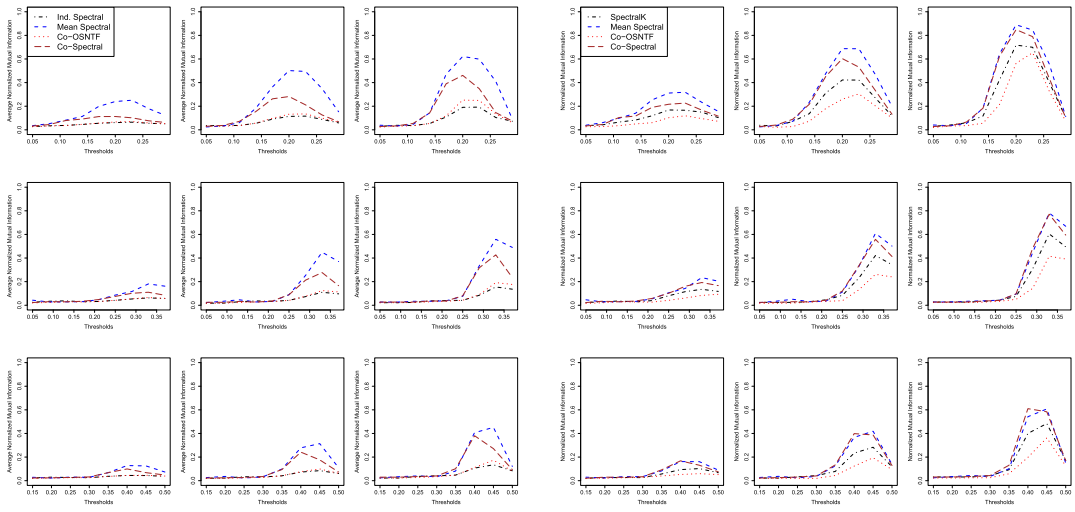
(a) Controls (b) Patients (a) Controls (b) Patients

(A) threshold 0.3 (B) threshold 0.4

FIG. 21. *Community estimate for AAL ROIs in controls and patients using the Co-OSNTF method with four communities for thresholds* (*A*) 0.3 *and* (*B*) 0.4.

6. Discussions and limitations. We have proposed a random effects model to jointly model the community structure in a sample from a population of networks. The proposed estimation and hypothesis testing methodology outperforms baseline and competing methods in simulated network samples. Our methods uncover meaningful differences between functional brain networks of healthy controls and schizophrenia patients. We hope the principled approach developed here will be useful in modeling and contrasting network valued samples in terms of their low-dimensional latent structures. We conclude by discussing a limitation of our methods and analysis. Another limitation of the methods is discussed in the Supplementary Material (Paul and Chen (2020b)).

Dependence on selection of thresholds and length of time series. In Section 4 we have performed comprehensive simulations on the performance of the proposed methods when the networks are directly generated through a RESBM. However, in fMRI application problems the raw data is in the format of a multivariate time series, and one needs to estimate the networks from the time series. While thresholding the correlation matrices at an appropriate level is a widely used method, not much attention has been given on how a thresholded estimator performs in estimating the networks. Clearly both the choice of the threshold and the length of the available time series are important considerations for performance of our methods. Therefore, we present a limited simulation where we generate multivariate time series data first and then estimate the adjacency matrices. For this purpose we combine the proposed RESBM with a model presented in Brownlees, Gudmundsson and Lugosi (2017), called the Stochastic Block Partial Correlation Model (SBPCM). As in Section 4, we generate a sample of $M = 10$ networks from the RESBM with $n = 100$ vertices and $k = 3$ communities. However, instead of directly applying our methods to the sample of networks, we first obtain n dimensional multivariate time series of length t for each of the samples, according to the SBPCM. In the terminology of Brownlees, Gudmundsson and Lugosi (2017), we consider three levels of the parameter ϕ, which controls the relative weight of the graph Laplacian matrix in creating the inverse covariance (concentration or precision) matrix. The three levels are $25, 50$ and 75, respectively. We also consider three levels of the length of the time series t which are $200, 500$ and 1000. The accuracy of estimating $\bar{Z}$ and the average accuracy of estimating $Z^{(m)}$s are presented in Figure 22. We note that there is no uniform "optimum threshold," and the accuracy is highest for different thresholds depending upon the ϕ parameter. The accuracies are highly sensitive to the choice of the threshold as well; however, there is usually a range of thresholds for which the accuracies are reasonable. In each case the accuracies are lower for a very low threshold when the network is dense and full of possibly spurious correlations, gradually increase with increasing threshold and, finally, start decreasing as the network becomes too sparse in high thresholds. Therefore, the sparse networks obtained at very high thresholds are not necessarily the best ones. We also note from Figure 22 that the thresholds at which the accuracies are higher generally increase with increasing ϕ. This makes intuitive sense, since the parameter ϕ effectively controls how much network structure is passed on to the correlation matrix from which the multivariate time series data is generated. Therefore, a higher ϕ means the network structure is strongly present; a high cutoff for the correlations produces the most informative networks. Across all values of ϕ, we note that the performance is poor for the length of the time series $t = 200$ and increases as t increases to 1000. This is, of course, also not surprising and underscores the dependence of accuracy of our methods on the estimation accuracy of the correlation matrices which in turn is dependent on the length of the time series.



(a) $t = 200$ (b) $t = 500$ (c) $t = 1000$ (d) $t = 200$ (e) $t = 500$ (f) $t = 1000$

(A) Average accuracy in estimating $Z^{(m)}$s (B) Accuracy of estimating $\bar{Z}$.

FIG. 22. *Performance of various methods in terms of* (*A*) *average accuracy in estimating the community assignments of the member networks* $Z^{(m)}$*s and* (*B*) *accuracy of estimating the mean putative community assignment* $\bar{Z}$ *from a sample of multivariate time series data of sizes* $200, 500, 1000$ *generated using a model which combines RESBM and SBPCM. The three rows of figures correspond to* $\phi = 25, 50, 75$*, respectively.*

Acknowledgments. This work was supported in part by National Science Foundation grants DMS-1406455 and DMS-1830412. This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA), which is supported by funds from the University of Illinois at Urbana-Champaign.

SUPPLEMENTARY MATERIAL

Supplement to "A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging" (DOI: 10.1214/20-AOAS1339SUPP; .pdf). The supplementary file contains the details of derivation of the methods, their convergence and implementation details, some additional simulations mentioned in Section 4, and a discussion of a limitation of the algorithms.

REFERENCES

ALEXANDER-BLOCH, A. F., GOGTAY, N., MEUNIER, D., BIRN, R., CLASEN, L., LALONDE, F., LENROOT, R., GIEDD, J. and BULLMORE, E. T. (2010). Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. *Front. Syst. Neurosci.* **4** 147. https://doi.org/10.3389/fnsys.2010.00147

ALEXANDER-BLOCH, A., LAMBIOTTE, R., ROBERTS, B., GIEDD, J., GOGTAY, N. and BULLMORE, E. (2012). The discovery of population differences in network community structure: New methods and applications to brain functional networks in schizophrenia. *NeuroImage* **59** 3889–3900.

BACCO, C. D., POWER, E. A., LARREMORE, D. B. and MOORE, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. E* **95** 042317. https://doi.org/10.1103/PhysRevE.95.042317

BARBILLON, P., DONNET, S., LAZEGA, E. and BAR-HEN, A. (2017). Stochastic block models for multiplex networks: An application to a multilevel network of researchers. *J. Roy. Statist. Soc. Ser. A* **180** 295–314. MR3600512 https://doi.org/10.1111/rssa.12193

BASSETT, D. S., WYMBS, N. F., PORTER, M. A., MUCHA, P. J., CARLSON, J. M. and GRAFTON, S. T. (2011). Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. USA* **108** 7641–7646.

BASSETT, D. S., NELSON, B. G., MUELLER, B. A., CAMCHONG, J. and LIM, K. O. (2012). Altered resting state complexity in schizophrenia. *NeuroImage* **59** 2196–2207.

BAZZI, M., PORTER, M. A., WILLIAMS, S., MCDONALD, M., FENN, D. J. and HOWISON, S. D. (2016). Community detection in temporal multilayer networks, with an application to correlation networks. *Multiscale Model. Simul.* **14** 1–41. MR3439753 https://doi.org/10.1137/15M1009615

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BETZEL, R. F., BERTOLERO, M. A., GORDON, E. M., GRATTON, C., DOSENBACH, N. U. F. and BASSETT, D. S. (2019). The community structure of functional brain networks exhibits scale-specific patterns of inter- and intra-subject variability. *NeuroImage* **202** 115990. https://doi.org/10.1016/j.neuroimage.2019.07.003

BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853 https://doi.org/10.1214/13-AOS1124

BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008** P10008.

BOCCALETTI, S., BIANCONI, G., CRIADO, R., DEL GENIO, C. I., GÓMEZ-GARDEÑES, J., ROMANCE, M., SENDIÑA-NADAL, I., WANG, Z. and ZANIN, M. (2014). The structure and dynamics of multilayer networks. *Phys. Rep.* **544** 1–122. MR3270140 https://doi.org/10.1016/j.physrep.2014.07.001

BRAUN, U., SCHÄFER, A., WALTER, H., ERK, S., ROMANCZUK-SEIFERTH, N., HADDAD, L., SCHWEIGER, J. I., GRIMM, O., HEINZ, A. et al. (2015). Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proc. Natl. Acad. Sci. USA* **112** 11678–11683.

BROWNLEES, C. T., GUDMUNDSSON, G. and LUGOSI, G. (2017). Community detection in partial correlation network models. Available at SSRN 2776505.

BULLMORE, E. and SPORNS, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nat. Rev., Neurosci.* **10** 186–198.

CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467 https://doi.org/10.1214/12-EJS729

CHEN, S., KANG, J., XING, Y. and WANG, G. (2015). A parsimonious statistical method to detect groupwise differentially expressed functional connectivity networks. *Hum. Brain Mapp.* **36** 5196–5206.

DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183. MR2390817 https://doi.org/10.1007/s11222-007-9046-7

DING, C., LI, T., PENG, W. and PARK, H. (2006). Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 126–135. ACM, New York.

DONG, X., FROSSARD, P., VANDERGHEYNST, P. and NEFEDOV, N. (2014). Clustering on multi-layer graphs via subspace analysis on Grassmann manifolds. *IEEE Trans. Signal Process.* **62** 905–918. MR3160322 https://doi.org/10.1109/TSP.2013.2295553

EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. MR1646856 https://doi.org/10.1137/S0895479895290954

FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage* **80** 426–444.

FORNITO, A., ZALESKY, A. and BREAKSPEAR, M. (2015). The connectomics of brain disorders. *Nat. Rev., Neurosci.* **16** 159–172.

FUJITA, A., TAKAHASHI, D. Y., PATRIOTA, A. G. and SATO, J. R. (2014). A non-parametric statistical test to compare clusters with applications in functional magnetic resonance imaging data. *Stat. Med.* **33** 4949–4962. MR3276511 https://doi.org/10.1002/sim.6292

GADELKARIM, J. J., SCHONFELD, D., AJILORE, O., ZHAN, L., ZHANG, A. F., FEUSNER, J. D., THOMPSON, P. M., SIMON, T. J., KUMAR, A. et al. (2012). A framework for quantifying node-level community structure group differences in brain connectivity networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 196–203. Springer, Berlin.

GINESTET, C. E., FOURNEL, A. P. and SIMMONS, A. (2014). Statistical network analysis for functional MRI: Summary networks and group comparisons. *Front. Comput. Neurosci.* **8** 51. https://doi.org/10.3389/fncom.2014.00051

GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. MR3693544 https://doi.org/10.1214/16-AOAS1015

GLEREAN, E., PAN, R. K., SALMI, J., KUJALA, R., LAHNAKOSKI, J. M., ROINE, U., NUMMENMAA, L., LEPPÄMÄKI, S., NIEMINEN-VON WENDT, T. et al. (2016). Reorganization of functionally connected brain subnetworks in high-functioning autism. *Hum. Brain Mapp.* **37** 1066–1079.

HAN, Q., XU, K. S. and AIROLDI, E. M. (2015). Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning* 1511–1520.

HE, Y., WANG, J., WANG, L., CHEN, Z. J., YAN, C., YANG, H., TANG, H., ZHU, C., GONG, Q. et al. (2009). Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS ONE* **4** e5226.

HUTCHISON, R. M., WOMELSDORF, T., ALLEN, E. A., BANDETTINI, P. A., CALHOUN, V. D., CORBETTA, M., DELLA PENNA, S., DUYN, J. H., GLOVER, G. H. et al. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage* **80** 360–378.

JONES, D. T., VEMURI, P., MURPHY, M. C., GUNTER, J. L., SENJEM, M. L., MACHULDA, M. M., PRZYBELSKI, S. A., GREGG, B. E., KANTARCI, K. et al. (2012). Non-stationarity in the "resting brain's" modular architecture. *PLoS ONE* **7** e39731. https://doi.org/10.1371/journal.pone.0039731

KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.

KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. MR0075510 https://doi.org/10.1002/nav.3800020109

KUJALA, R., GLEREAN, E., PAN, R. K., JÄÄSKELÄINEN, I. P., SAMS, M. and SARAMÄKI, J. (2016). Graph coarse-graining reveals differences in the module-level structure of functional brain networks. *Eur. J. Neurosci.* **44** 2673–2684. https://doi.org/10.1111/ejn.13392

KUMAR, A., RAI, P. and DAUME, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems* 1413–1421.

LEE, D. D. and SEUNG, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems* 556–562.

LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

LIU, Y., LIANG, M., ZHOU, Y., HE, Y., HAO, Y., SONG, M., YU, C., LIU, H., LIU, Z. et al. (2008). Disrupted small-world networks in schizophrenia. *Brain* **131** 945–961.

LIU, J., WANG, C., GAO, J. and HAN, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM* **13** 252–260. SIAM, Philadelphia.

LYNALL, M.-E., BASSETT, D. S., KERWIN, R., MCKENNA, P. J., KITZBICHLER, M., MULLER, U. and BULLMORE, E. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* **30** 9477–9487.

MANKAD, S. and MICHAILIDIS, G. (2013). Structural and functional discovery in dynamic networks with non-negative matrix factorization. *Phys. Rev. E* **88** 042812.

MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311 https://doi.org/10.1111/rssb.12200

MEUNIER, D., LAMBIOTTE, R. and BULLMORE, E. T. (2010). Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* **4** 200. https://doi.org/10.3389/fnins.2010.00200

MIRZAL, A. (2014). A convergent algorithm for orthogonal nonnegative matrix factorization. *J. Comput. Appl. Math.* **260** 149–166. MR3133336 https://doi.org/10.1016/j.cam.2013.09.022

MOUSSA, M. N., STEEN, M. R., LAURIENTI, P. J. and HAYASAKA, S. (2012). Consistency of network modules in resting-state FMRI connectome data. *PLoS ONE* **7** e44428. https://doi.org/10.1371/journal.pone.0044428

MUCHA, P. J., RICHARDSON, T., MACON, K., PORTER, M. A. and ONNELA, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328** 876–878. MR2662590 https://doi.org/10.1126/science.1184819

NARAYAN, M. and ALLEN, G. I. (2016). Mixed effects models for resampled network statistics improves statistical power to find differences in multi-subject functional connectivity. *Front. Neurosci.* **10** 108. https://doi.org/10.3389/fnins.2016.00108

NICOSIA, V. and LATORA, V. (2015). Measuring and modeling correlations in multiplex networks. *Phys. Rev. E* **92** 032805.

PAPADIMITRIOU, C. H. (2003). *Computational Complexity*. Wiley, New York.

PAUL, S. and CHEN, Y. (2016a). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Stat.* **10** 3807–3870. MR3579677 https://doi.org/10.1214/16-EJS1211

PAUL, S. and CHEN, Y. (2016b). Null models and modularity based community detection in multi-layer networks. Preprint. Available at arXiv:1608.00623.

PAUL, S. and CHEN, Y. (2016c). Orthogonal symmetric non-negative matrix factorization under the stochastic block model. Preprint. Available at arXiv:1605.05349.

PAUL, S. and CHEN, Y. (2020a). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.* **48** 230–250. MR4065160 https://doi.org/10.1214/18-AOS1800

PAUL, S. and CHEN, Y. (2020b). Supplement to "A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging." https://doi.org/10.1214/20-AOAS1339SUPP

PEIXOTO, T. P. (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Phys. Rev. E* **92** 042807.

PENNY, W. D., FRISTON, K. J., ASHBURNER, J. T., KIEBEL, S. J. and NICHOLS, T. E. (2011). *Statistical Parametric Mapping*: *The Analysis of Functional Brain Images*. Academic Press, San Diego.

PERCIVAL, D. B. and WALDEN, A. T. (2006). *Wavelet Methods for Time Series Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **4**. Cambridge Univ. Press, Cambridge. MR2218866

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.

REICHARDT, J. and BORNHOLDT, S. (2006). Statistical mechanics of community detection. *Phys. Rev. E* (3) **74** 016110, 14. MR2276596 https://doi.org/10.1103/PhysRevE.74.016110

REYES, P. and RODRIGUEZ, A. (2016). Stochastic blockmodels for exchangeable collections of networks. Preprint. Available at arXiv:1606.05277.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52** 1059–1069.

SIMPSON, S. L., BOWMAN, F. D. and LAURIENTI, P. J. (2013). Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. *Stat. Surv.* **7** 1–36. MR3161730 https://doi.org/10.1214/13-SS103

SIMPSON, S. L., LYDAY, R. G., HAYASAKA, S., MARSH, A. P. and LAURIENTI, P. J. (2013). A permutation testing framework to compare groups of brain networks. *Front. Comput. Neurosci.* **7** 171. https://doi.org/10.3389/fncom.2013.00171

SPORNS, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nat. Neurosci.* **17** 652–660.

STAM, C. J. (2014). Modern network science of neurological disorders. *Nat. Rev., Neurosci.* **15** 683–695.

STANLEY, N., SHAI, S., TAYLOR, D. and MUCHA, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **3** 95–105. MR3515211 https://doi.org/10.1109/TNSE.2016.2537545

STEEN, M., HAYASAKA, S., JOYCE, K. and LAURIENTI, P. (2011). Assessing the consistency of community structure in complex networks. *Phys. Rev. E* **84** 016111.

STEVENS, A. A., TAPPON, S. C., GARG, A. and FAIR, D. A. (2012). Functional brain network modularity captures inter- and intra-individual variation in working memory capacity. *PLoS ONE* **7** e30468. https://doi.org/10.1371/journal.pone.0030468

STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. *Computer Science and Scientific Computing*. Academic Press, Boston, MA. MR1061154

SWEET, T. M., THOMAS, A. C. and JUNKER, B. W. (2015). Hierarchical mixed membership stochastic blockmodels for multiple networks and experimental interventions. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 463–487. CRC Press, Boca Raton, FL. MR3380043

TANG, W., LU, Z. and DHILLON, I. S. (2009). Clustering with multiple graphs. In *Proceedings of the Ninth IEEE International Conference on Data Mining* 1016–1021. IEEE Press, New York.

TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.

VALLES-CATALA, T., MASSUCCI, F. A., GUIMERA, R. and SALES-PARDO, M. (2016). Multilayer stochastic block models reveal the multilayer structure of complex networks. *Phys. Rev. X* **6** 011036.

VAN DEN HEUVEL, M. P. and POL, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20** 519–534.

VAN DEN HEUVEL, M. P., MANDL, R. C., STAM, C. J., KAHN, R. S. and POL, H. E. H. (2010). Aberrant frontal and temporal complex network structure in schizophrenia: A graph theoretical analysis. *J. Neurosci.* **30** 15915–15926.

VAN DEN HEUVEL, M. P., SPORNS, O., COLLIN, G., SCHEEWE, T., MANDL, R. C., CAHN, W., GOÑI, J., POL, H. E. H. and KAHN, R. S. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA Psychiatry* **70** 783–792.

WANG, J., ZUO, X., DAI, Z., XIA, M., ZHAO, Z., ZHAO, X., JIA, J., HAN, Y. and HE, Y. (2013). Disrupted functional brain connectome in individuals at risk for Alzheimer's disease. *Biological Psychiatry* **73** 472–481.

WEBER, M. J., DETRE, J. A., THOMPSON-SCHILL, S. L. and AVANTS, B. B. (2013). Reproducibility of functional network metrics and network structure: A comparison of task-related BOLD, resting ASL with BOLD contrast, and resting cerebral blood flow. *Cognitive*, *Affective*, *& Behavioral Neuroscience* **13** 627–640.

WHITFIELD-GABRIELI, S. and NIETO-CASTANON, A. (2012). CONN: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect*. **2** 125–141.

XIA, M., WANG, J. and HE, Y. (2013). BrainNet viewer: A network visualization tool for human brain connectomics. *PLoS ONE* **8** e68910.

YU, Q., PLIS, S. M., ERHARDT, E. B., ALLEN, E. A., SUI, J., KIEHL, K. A., PEARLSON, G. and CALHOUN, V. D. (2011). Modular organization of functional network connectivity in healthy controls and patients with schizophrenia during the resting state. *Front*. *Syst*. *Neurosci*. **5** 103. https://doi.org/10.3389/fnsys.2011.00103

YU, Q., ALLEN, E. A., SUI, J., ARBABSHIRANI, M. R., PEARLSON, G. and CALHOUN, V. D. (2012). Brain connectivity networks in schizophrenia underlying resting state functional magnetic resonance imaging. *Current Topics in Medicinal Chemistry* **12** 2415–2425.

ZALESKY, A., FORNITO, A. and BULLMORE, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage* **53** 1197–1207.