A framework of change detection for dynamic networks

Fuchen Liu

Advisor: Kathryn Roeder and David Choi
Department of Statistics and Data Science, Carnegie Mellon University

March 2, 2018

Abstract

Understanding the development of intrinsic brain network organization based on resting state functional Magnetic Resonance Imaging (rs-fMRI) is of increasing importance and interest. However, given a sample of rs-fMRI images from different aged subjects, the problem of how to test for changes in network organization across ages is not well understood. We propose two new approaches, eigen test (ET) and likelihood ratio test (LRT), to fill in the gap. As the names suggest, ET utilizes the eigenvectors and LRT compares the log likelihood of competing models. The former is sensitive only to changes in community assignment, while the latter detects both changes in community assignments and connectivity between communities. We are interested detecting various changes in network evolution: smooth trends over time, multiple change points and outliers. We propose to develop a theoretical framework for LRT, especially for models with multiple change points. We will evaluate our methods in a wide variety of settings: simulated data, rs-fMRI data and dynamic gene co-expression networks, which can be analyzed with structure similar to rs-fMRI networks. Thus far, with preliminary analysis we have discovered that the network organization in rsfMRI is stable over time, which sheds light on competing hypotheses in the literature. Gene co-expression data, which is available from conception to adulthood, shows a marked change point at birth.

1 Introduction

Networks or graphs are used to display connections within a complex system. The vertices in a network often reveal clusters with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, could arise from functionality of distinct components of the network, e.g., brain regions cooperate a specific function or genes co-regulating a cellular process.

Statistical theory (Kolaczyk [2009], Newman [2010], Lei and Rinaldo [2015]) has mostly focused on static networks, observed as a single snapshot in time or developmental epoch. In reality, networks are generally dynamic, and it is of substantial interest to visualize and model their persistency. Applications abound, e.g., social networks in Twitter, dynamic diffusion networks in physics and gene co-expression networks or functional connectivity networks for developing brains. Recent works have sought to extend community detection to dynamic networks (Matias and Miele [2016], Ghasemian et al. [2016], Cribben and Yu [2017a], Nguyen et al. [2014], Xu and Hero [2014], Liu et al. [2018]),

to centrality (Taylor et al. [2017]), and to extend clustering to dynamic data (Chi et al. [2007]). Community detection is allowed to perform more precisely within dynamic network series, even if the single is very weak in each single network.

However, how to perform further analysis is still not well discussed. One vital scientific question for dynamic networks is whether the community structure is persistent over time, or if not, where is the change point. This problem is especially important in understanding brain functions. Saying we have a list of functional connectivity networks from different aged brains, then if we can detect that some brain regions belongs to a same community in children's brain networks but break down into different communities for adult, it would be a very meaningful scientific finding. On the contrary, it would also be significant if we can prove that the community structure of brain regions is persistent over ages. Recent scientific papers have been focus on testing such changes over grouped networks (Marek et al. [2016]), but there is still no statistical method to detect and test the change point among dynamic networks.

To fill in the gap, we proposed a framework to detect and visualize the changes of a list of networks. The changes of networks includes both changes of communities and changes of connectivities between communities. There also could be three different kinds of changes as listed below:

1 Change point. The network structure (community assignments) or connectivity change significantly at one or several times (age in our setting)

2 Outlier. The network structure or connectivity is dramatically different at one or several times. But the structure or connectivity are similar for the matrices before and after that point.

3 Smooth trend over time. For dynamic networks, a more reasonable assumption may should be that the changes are smooth over time. There is no significant change point but the network structure or functional connectivity are varying smoothly over time.

The major component of this proposal centers around change point detection methods for dynamic networks. Our algorithms are all based on hypothesis testing, with "no change" as our null hypothesis. Both spectral based method, eigen test (ET) and likelihood ratio test (LRT) are proposed as possible approaches. ET employs the difference of eigenvectors between networks as the test statistics based on the theories of spectral clustering under stochastic block model (Holland et al. [1983]). While LRT is performed by comparing the performance with or without a partition of the network series. Significant test for both methods would be relying on data permutation.

We also propose outlier detection and smooth trend detection algorithms to support the whole framework. The test statistics of these two detections are still based on ET and LRT. As a whole procedure, outlier detection is our first step to remove "noise" networks from the series. Then a change point detection can be performed to detect signification change. Assume we have already divided the networks into different small groups without significant change, a trend detection can then be performed to find out whether the community structure is smoothly varying over time. As an extreme case, if the network structure is varying rapidly between every consecutive networks, then every points are change points.

We applied the framework into two different dataset: functional connectivity networks from resting state fMRI (rsfMRI) and gene co-expression networks from monkey brain. rs-fMRI is a method of functional brain imaging that can be used to evaluate regional interactions that occur when a subject is not performing an explicit task. Time series of blook-oxygen-level dependent signal is observed for every voxels or regions. Functional connectivity (FC) networks can be formed with rsfMRI data with brain regions as vertices and correlation of paired time series as edges. We got FC networks from subject with different ages to form a list of dynamic networks. With this dataset,

we prove that the community structure is consistent over ages. Gene co-expression networks have a similar structure as rsfMRI with vertices as genes and edges as correlation of gene expressions.

The remaining of this proposal is organized as follows. In section 2, we provide a background on stochastic block model and its eigen-decomposition structure, as well as one global spectral clustering algorithm for dynamic networks. In section 3, we give a brief introduction of related change point detection algorithms and also develop our own change point detection methods. And in section 4, we discuss outlier detection and trend detection procedures. In section 5, we give application results in both rsfMRI data and gene co-expression networks. In section 6, we map out our future plan. Additionally, in Appendix A1, we give simulations results of proposed methods. In Appendix A2, we give the proof of related theoretical results.

2 Background information

2.1 Stochastic block model (SBM)

To support our change point detection, we want to find models which are both realistic and mathematical tractable, and can lead to a consistent community detection results. There have been a lot of recent works focus on statistical modeling of networks [Newman paper]. We used stochastic block model (SBM) as a fundamental model because of its simplicity and expressive power. Especially, the eigen structure of SBM has already been well explained [cite jing], which allows eigen decomposition becomes a powerful and consistent way to convey information and apply community detection. But our method can also be easily extended to other networks models, like degree corrected block model.

A network is formed by its nodes and edges. If we further assume the nodes are belong to different communities, SBM gives a way to generate edges based on nodes and their communities. A SBM with n nodes and K communities is parameterized by a pair of matrices (Θ, B), where $\Theta \in \mathbb{M}_{n,K}$ is the *membership matrix* and $B \in \mathbb{R}^{K\times K}$ is a symmetric *connectivity matrix*. For each node i, let g_i be its community label, such that the ith row of Θ is 1 in column g_i and 0 elsewhere. The entry $B_{k,l}$ in B is the edge probability between any node in community k and any node in community l. Given (Θ, B), the adjacency matrix $A = (a_{ij})_{1\le i,j\le n}$ is a symmetric matrix with diagonal equals to 0. The entries of A are generated independently from Bernoulli distribution as follows.

$$P(a_{ij}=1) = \begin{cases} B_{g_i,g_j} & \text{if } i<j \\ 0 & \text{if } i=j \end{cases} \quad , a_{ij} = a_{ji} \text{ for } i>j \tag{1}$$

In a SBM, the heuristic of spectral clustering is to relate the eigenvectors of A to those of $P := \Theta B \Theta^T$ using the fact that $\mathbb{E}(A) = P - diag(P)$. Let $P = UDU^T$ be the eigen-decomposition of P with $U^TU = I_k$ and $D \in \mathbb{R}^{K\times K}$ diagonal, then it is easy to see that U has only K distinct rows since P has only K distinct rows. Under mild conditions, two nodes are in the same community if and only if their corresponding rows in U are the same. Lei and Rinaldo [2015] gives a following lemma for the decomposition of P.

Lemma 2.1. *(Basic eigen-structure of SBM) Let the pair (Θ, B) parametrize a SBM with K communities, where B is full rand. Let UDU^T be the eigen-decomposition of $P = \Theta B \Theta^T$. Then $U = \Theta X$ where $X \in \mathbb{R}^{K\times K}$ and $||X_{k*} - X_{l*}|| = \sqrt{n_k^{-1} + n_l^{-1}}$, for all $1 \le k < l \le K$.*

Based on this observation, spectral clustering tries to estimate U and its row clustering using a spectral de-composition of A. Letting $A = \hat{U}\hat{D}\hat{U}^T$ be the eigen-decomposition of A and $\hat{U}_K$ denote

K eigenvectors corresponding to the K largest absolute eigenvalues. $\hat{U}_K \hat{D}_K \hat{U}_K^T$ could be a good approximation of $P = UDU^T$.

2.2 Global spectral clustering for dynamic networks

Community detection with one network could be challenging if the signal is not strong enough. For example, we can not detect proper communities with only one imaging in rs-fMRI data. Getting the mean of several networks and then performing community detection is a simple and powerful solution. However, in the analysis of smooth trend detection, we want to compare the community structure of each specific network. In order to make it possible, we introduce one previous work of our group, PisCES (Liu et al. [2018]).

Our method, PersIStent Communities by Eigenvector Smoothing (PisCES), implements degree-corrected spectral clustering, with a smoothing term to promote similarity across time periods, and iterates until a fixed point is achieved. Specifically, this global spectral clustering approach combines the current network with the leading eigenvector of both the previous and future results. The combination is formed as an optimization problem that can be solved globally under moderate levels of smoothing when the number of communities is known. We also utilized data-driven method to choose appropriate levels of both smoothing and model order, as well as to balance regularization with "letting the data speak".

The basic idea of PisCES is to perform eigenvector smoothing. Let $A_1, \ldots, A_T$ denote a time series of symmetric adjacency matrices, and for $t = 1, \ldots, T$, let L_t denote the Laplacianized version of A_t. Let K be fixed, and let $V_t \in \mathbb{R}^{n \times K}$ denote the matrix whose columns are the K leading eigenvectors of L_t. Let $U_t = V_t V_t^T$, the projection matrix onto the column space of V_t.

In static spectral clustering, one would apply K-means clustering to $V_1, \ldots, V_T$ separately. To share signal strength over time, a simplified form of PisCES would solve the following optimization problem, which returns a sequence of matrices $\bar{U}_1, \ldots, \bar{U}_T$ that are smoothed versions of $U_1, \ldots, U_T$:

$$\begin{aligned} &\min_{\bar{U}_1, \ldots, \bar{U}_T} \sum_{t=1}^{T} \|U_t - \bar{U}_t\|_F^2 + \alpha \sum_{t=1}^{T-1} \|\bar{U}_t - \bar{U}_{t+1}\|_F^2 \\ &\text{subject to } \bar{U}_t \in \{VV^T : V \in \mathbb{R}^{n \times K}, V^T V = I\} \, \forall \, t, \end{aligned} \tag{2}$$

and then apply K-means clustering to the eigenvectors of each smoothed matrix $\bar{U}_1, \ldots, \bar{U}_T$ separately.

3 Change point detection algorithms

3.1 Related works

Even though there is no change point detection method particularly for network series, the change point detection for long-term multivariate data is well discussed. Ivor Cribben and Yi Yu (Cribben and Yu [2017b]) proposed a Network Change Points Detection (NCPD) method based on spectral clustering. Start from n time series for n ROIs, each time series have T time points. They want to find possible partitions of $t = \{1, 2, \cdots, T\}$, say $1 < \tau_1 < T$. And then divided the time series into 2 part $T_L = \{1, 2, \cdots, \tau_1\}$ and $T_R = \{\tau_1 + 1, \cdots, T\}$ where the correlation matrices R_L and r_R formed with T_L and T_R have different network structure. Their test statistics is the singular value

between centralized eigenvectors of R_L and R_R (Figure 2), where the centralized index are results from community detection by spectral clustering. They also performed stationary bootstrap to do significant testing. Since their test statistics is the singular value between centralized eigenvectors, this test only focus on the community assignments. On the other words, different connectivity within or between communities would not influence the results.

This method can easily be extended to our settings, we just need to change R_L and R_R to averaged correlation matrices for younger ages and older ages. However, since NCPD rely on the results of spectral clustering, it is not stable in networks series. We will use it as a baseline and discuss the results in Appendix A1.

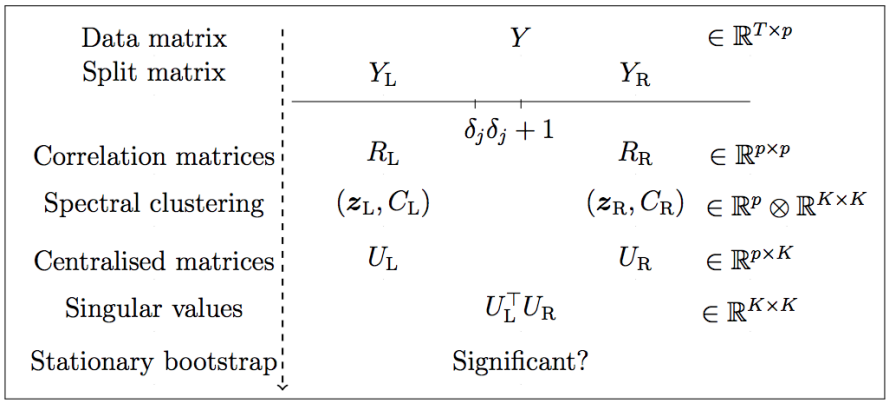

Figure 1: Work flow of NCPD

There are also several papers which focus on the change of the covariance (correlation) matrix. DCR (Cribben et al. [2012, 2013]) use graphical lasso to estimate the precision matrix and a partition is significant if the BIC of graphical lasso rise significantly after that partition. Moving window (Lindquist et al. [2014]) approaches are also well used in fMRI related papers. Many statistical papers (Avanesov and Buzun [2016], Matteson and James [2014]) also proposed many method and series based on multivariate normal distribution. The problem setting of those paper are pretty different from ours, but we still can borrow some ideas, like model based likelihood ratio test, heuristic searching and sliding window searching, from those papers.

3.2 Problem setting

Let $\{A_1, A_2, \cdots, A_T\} \in R^{n\times n}$ denote T symmetric matrices from subject with time $t = \{1, \cdots, T\}$. Each matrix have the same n nodes and those n nodes can be divided into K_t communities by community detection of A_t. To make things easier to explain, we assume K_t is constant for different t. But our method can be easily extended to time varying K_t. And let $\Theta_1, \cdots, \theta_T \in M_{n,K}$ denote the membership matrix for each network.

We further assume that A_t are adjacency matrices independently generated from SBM. $\{A_t\}_{ij}$ denotes the edge between node i and node j and is generated based on their group labels and a symmetric connectivity matrix $B \in R^{K\times K}$ as showed in section 2.1: $\{A_t\}_{ij} \sim Bernoulli(B_{g_i,g_j})$.

As a hypothesis testing problem, our null hypothesis is that the networks are i.i.d generated from the same SBM model. We consider two different problems: whether the membership is consistent or whether the model, which means either membership or probability matrix, is consistent. Those

two problems are both arised from specific scientific questions. Use rsfMRI as an example, sometime we only care about the communities (or clusters) of brain regions, while sometimes the connectivity between communities are also took into account. The following two null hypothesis illustrate those two cases.

$$H_0^{(1)} : \Theta_1 = \Theta_2 = \cdots = \Theta_T \tag{3}$$

$$H_0^{(2)} : \Theta_1 = \Theta_2 = \cdots = \Theta_T \& B_1 = B_2 = \cdots = B_T \tag{4}$$

We further assume that there is only one change point (multiple change points case would be discussed in section 3.5). Then the alternate hypothesis is that there exist one change point $\tilde{\tau}$ where the model changes after $t = \tilde{\tau}$:

$$H_{single}^{(1)} : \Theta_1 = \cdots = \Theta_{\tilde{\tau}} \neq \Theta_{\tau\tilde{+}1} = \cdots = \Theta_T \tag{5}$$

$$H_{single}^{(2)} : \Theta_1 = \cdots = \Theta_{\tilde{\tau}} \neq \Theta_{\tau\tilde{+}1} = \cdots = \Theta_T \text{ or } P_1 = \cdots = B_{\tilde{\tau}} \neq B_{\tau\tilde{+}1} = \cdots B_T \tag{6}$$

We proposed 2 methods: eigen test (ET) and likelihood ratio test (LRT) to deal with these two problems respectively. ET focus on the change of membership while LRT considers both membership and connectivity.

3.3 Eigen test

As the names suggests, ET utilizes the eigenvectors of adjacency matrices to compare the membership. Assume $\tau \in (1, T)$ is one candidate change point, we get 2 matrices $A_{left} = \sum_{t=1}^{\tau} \frac{A_t}{\tau}$ and $A_{right} = \sum_{t=\tau+1}^{T} \frac{A_t}{T-\tau}$ as the averaged adjacency matrix of left part and right part. Let U_{left} and $U_{right} \in \mathbb{R}^{n \times K}$ denote the leading K eigenvectors of A_{left} and A_{right}, where K can either be per-defined or estimated from other methods (cite or ref). Based on the theories of SBM (lemma 2.1), eigenvectors contains the information of community assignments. So we define our first test statistics as the difference between $U_{left}U_{left}^T$ and $U_{right}U_{right}^T$.

$$\gamma_\tau = ||U_{left}U_{left}^T - U_{right}U_{right}^T||_2$$

Then we choose the τ which gives biggest test statistics, which means biggest difference between left part and right part, among all candidation change points as our estimated change point.

$$\hat{\tau} = \arg\max_{\tau} \gamma_\tau$$

Permutation test is then performed to test whether $\hat{\tau}$ gives a significant change. Assume $A_1^{(b)}, A_2^{(b)}, ..., A_T^{(b)}$ is the b-th trial of permutation of the original adjacency matrices, we calculate test statistics $\gamma_{\hat{\tau}}^{(b)}$ at $t = \hat{\tau}$ for the permutation trail the same as above. After N_b trials, let $\gamma_{\hat{\tau}}^{(1)}, \cdots, \gamma_{\hat{\tau}}^{(N_b)}$ denotes the corresponding test statistics. The p-value is defined as:

$$\text{p-value} = \frac{\text{cardinality}(\{b : \gamma_{\hat{\tau}}^{(b)} > \gamma_{\hat{\tau}}\})}{N_b}.$$

One thing need to be notices is that the eigenvectors are not only related to membership but also related to connectivity matrix. Based on lemma 2.1, $U = \Theta X$ where X is from the connectivity matrix B. But in practice, we find that our test statistics actually can only convey the difference between membership Θ. We will explore possible theoretical results and extensions about that as a future work.

Test statistics from NCPD As mentioned in section 3.1, NCPD can be easily extended to our settings. Apply kmeans to $U_{left} \in R^{n\times K}$ and $U_{right} \in R^{n\times K}$. With kmeans, we can get community labels g_{left} and g_{right} and K centroids for each community. Since we only interested in the community assignments, we then construct new matrices $C_{left} \in R^{n\times K}$ and $C_{right} \in R^{n\times K}$, whose rows are the corresponding centroids: $C_{left}(i,:) =$ centroid of $g_{left}(i)$. The column spaces of C_{left} and C_{right} encode the location information, so we do not impose the condition that the columns have to be orthonormal. Then the test statistics is defined as the summation of singular values of $C_{left}^T C_{right}$

$$NCPD_\tau = sum(svd(C_{left}^T C_{right}))$$

One drawback of this statistics could be that $\gamma 1$ is big for all of the candidate change points around the boundary. For example is we have $T = 10$ networks and change point at $t = 5$. Then we may get big $\gamma 1$ from $t = 3$ to $t = 7$. Further discussion would be showed in Appendix A1.

3.4 Likelihood ratio test

The likelihood ratio test (LRT) examine whether a partition of networks series would improve the likelihood fitting based on stochastic block model (SBM). Given a list of networks which ordered by time, we first fit a consistent SBM all of the networks by assuming they are i.i.d under the null hypothesis (Eq: (4)). Then for any candidate change point, we fit SBM separately to its previous part and future part, i.e. we assume different model for each side. LRT would test if the partition make a significant improving of the fitting.

Let $A_1, A_2, \cdots, A_T$ denote adjacency matrices of the network series as showed above, we use the mean matrices to fit the model if we assume they are i.i.d. For example, under the null hypothesis, those matrices are all i.i.d from the same model. We get $A_{null} = \sum_{t=1}^{T} \frac{A_t}{T} \in \mathbb{R}^{n\times n}$ as the aggregated input matrix. Then SBM is fitted to A_{null} by performing community detection. Many different methods can be used in this case [10, 11, 12]. Let $\hat{g}_{null} \in \mathbb{R}^n$ denote the corresponding estimated membership, where $g_i = k$ means the i-th node belongs to the k-th community. Then the connectivity matrices $\hat{B}_{null} \in \mathbb{R}^{K\times K}$ in SBM can be estimated as:

$$\hat{B}_{kl} = \frac{\sum_{i,j} A_{ij} I_{\hat{g}_i=k} I_{\hat{g}_j=l}}{\sum_{ij} I_{\hat{g}_i=k} I_{\hat{g}_j=l}}$$

For any adjacency matrix A and membership g, the log-likelihood can be calculated as the following,

$$\mathcal{L}(A, \hat{g}, \hat{B}) = \sum_{ij} A_{ij} log(\hat{B}_{\hat{g}_i,\hat{g}_j}) + (1 - A_{ij}) log(1 - \hat{B}_{\hat{g}_i,\hat{g}_j})$$

Then the log-likelihood under null hypothesis is defined as the summation of log-likelihood of each matrix with $\hat{g}_{null}$ and $\hat{B}_{null}$ as the following,

$$L_{null} = \sum_{t=1}^{T} \mathcal{L}(A_t, \hat{g}_{null}, \hat{B}_{null}) \tag{7}$$

For one candidate change point $\tau \in (1, T)$, we can also get the mean adjacency matrix for the previous part $A_{left} = \sum_{t=1}^{\tau} \frac{A_t}{\tau} \in \mathbb{R}^{n\times n}$ and the mean adjacency matrix for the future part $A_{right} = \sum_{t=\tau+1}^{T} \frac{A_t}{T-\tau} \in \mathbb{R}^{n\times n}$. Then $\hat{g}_{left}$, $\hat{g}_{right}$ and $\hat{B}_{left}$, $\hat{B}_{right}$ can be estimated with the same procedure. The log-likelihood with partition at $t = \tau$ would be based on those 2 models,

$$L_\tau = \sum_{t=1}^{\tau} L(A_t, g_{left}, P_{left}) + \sum_{t=\tau+1}^{T} L(A_t, g_{right}, P_{right}) \tag{8}$$

Then the test statistics is defined as the following,

$$\Delta_\tau = L_\tau - L_{null} \tag{9}$$

where larger Δ_τ shows bigger improvement. The final estimated change point is one who get the biggest test statistics among all candidate change point $\hat{\tau} = \text{argmax}_\tau \Delta_\tau$.

To test whether the estimated change point is significant, we applied the same permutation procedure as ET. Let $\tilde{\Delta}_{\hat{\tau}}^{(1)}, \cdots, \tilde{\Delta}_{\hat{\tau}}^{(N_b)}$ denote the test statistics at $\hat{\tau}$ from N_b different permutations, The p-value is defined as:

$$\text{p-value} = \frac{\text{cardinality}(\{b : \Delta_{\hat{\tau}}^{(b)} > \Delta_{\hat{\tau}}\})}{N_b}.$$

Since the log likelihood based on SBM including both membership g and connectivity matrix B, LRT is able to find the changes in both community and connectivity. In practice, LRT (ref sim figure)is also very powful in detecting both changes.

3.5 Multiple change points

In this section, we extend our methods into multiple change points. The null hypothesis is still the same as before (eq: (3), (4)), while the corresponding alternate hypothesis becomes there exist D change points $\tilde{\tau}_1, \tilde{\tau}_2, \cdots, \tilde{\tau}_D$ where model changes at those times.

$$H_{multiple}^{(1)} : \Theta_1 = \cdots = \Theta_{\tilde{\tau}_1} \neq \Theta_{\tilde{\tau}_1+1} = \cdots = \Theta_{\tilde{\tau}_2} \neq \Theta_{\tilde{\tau}_2+1} \cdots \tag{10}$$

$$H_{multiple}^{(2)} : \text{Either } \Theta \text{ or } B \text{ change at } \tilde{\tau}_1, \tilde{\tau}_2, \cdots, \tilde{\tau}_D \tag{11}$$

To estimate multiple change points, we iteratively apply the techniques of single change point as follows (Figure 2). Suppose that $d-1$ change points have been estimated at locations $0 < \hat{\tau}_1 < \cdots < \hat{\tau}_{d-1} < T$. These change points separate the matrices into d clusters of matrices $\hat{C}_1, \hat{C}_2, \cdots, \hat{C}_d$, such that $\hat{C}_i = \{A_{\hat{\tau}_{i-1}+1}, \cdots, A_{\hat{\tau}_i}\}$, in which $\hat{\tau}_0 = 0$ and $\hat{\tau}_d = T$. Given these clusters, we then apply the methods for finding a single change point to the observations within each clusters. Specifically, for the ith cluster $\hat{C}_i$ denote the proposed change point location as $\hat{\tau}(i)$ and the associated statistics. Then, let

$$i^* = \underset{i \in \{1, \cdots, d\}}{\text{argmax}} \ \hat{\gamma}(i)$$

the corresponding d-th estimated change point is $\hat{\tau}_d = \hat{i^\star}$ located within cluster $\hat{C}_{i^\star}$. This procedure has running time (DT^2) where D is the unknown number of change points.

We prove that the heuristic procedure is consistent when the number of networks goes to infinity. The estimated change point at each step would always be around the true change point even with multiple change points, which means a mixture of models. More detailed discussion about the theoretical results would be discussed in section 3.6. However, in practice, the length of network series could be very small, even thought the length of the total network series is large enough, the distance between two change points could also be relatively short. So we do find that a sliding window search is necessary in some cases. We would leave it as a future work to do in this proposal.

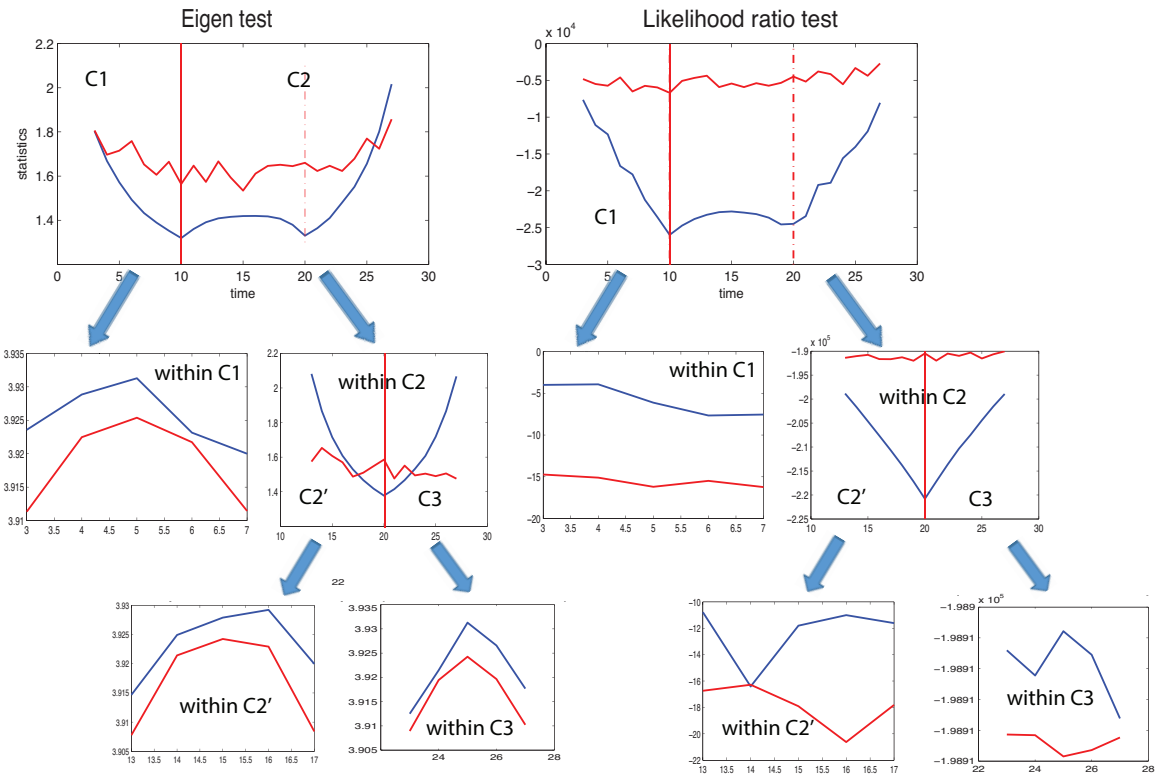

Figure 2: Bisection searching with 2 change points. The procedure of hierarchical (bisection) searching is showed for Eigen test and likelihood ratio test with 2 change points. The blue line shows original test statistics and red line shows statistics from permutation test. The vertical dotted red line shows true change point and vertical solid line shows how we divide the matrices in each step.

3.6 Theoretical results for likelihood ratio test

Here we list some theoretical results we get for likelihood ratio test. For single change point, the consistency is based on the accuracy of estimation of the model, and the sub-additivity of the metrics (log-likelihood or square error). For multiple change point, the key lemma is that the statistics can still be maximized in true change point when there are mixture of distributions, or multiple change points. Is it also because the metrics maximized with true model and the sub-additivity of the metrics.

3.6.1 Single change point

Assumption 1 and theorem 3.1 show that likelihood ratio test can find the true change point asymptotically when the sample size of each window goes to infinity.

Assumption 1. *Suppose we have a sequence of independent adjacency matrices from 2 different SBM:* $\mathbb{M}_1 : \{\Theta_1, B_1\}$ *and* $\mathbb{M}_2 : \{\Theta_2, B_2\}$*, where* $\Theta \in \mathbb{R}^{n\times n}$ *denotes the membership and* $B_i \in \mathbb{R}^{K_i\times K_i}$ *is the connectivity matrix between* $K_i, i = 1, 2$ *communities. Specifically, let* $\tau \in (1, T)$ *denote the change point and* $\pi \in (0, 1)$ *denote the fraction of the observations belongs to one of the distributions, such that* $A_1, A_2, \cdots, A_{\tau=\pi T} \sim \mathbb{M}_1$ *and* $Z_{\tilde{\tau}+1}, \cdots, Z_T \sim \mathbb{M}_2$ *for every sample size T. Let* $\{\delta_T\}$ *be a sequence of positive numbers such that* $\delta_T \to 0$ *and* $T\delta_T \to \infty$*, as* $T \to \infty$*.*

Theorem 3.1. *Suppose Assumption 1 holds. Let* $\hat{tau}_T$ *denote the estimated change point location for a sample of size T, as defined from the likelihood ratio test. Then for T large enough and* $\pi \in [\delta_T, 1 - \delta_T]$ *and further more, for all* $\epsilon > 0$

$$P(\lim_{T\to\infty} |\hat{\tau}_T/T - \tilde{\tau}/T| > \epsilon) = 0$$

To prove theorem 3.1, firstly, we assume that the true model could be estimated when sample size goes to infinity. If the adjacency matrices are all from the same model, It is easy to imagine that the mean adjacency matrix of infinite samples would converge to the true probability matrix $P = \Theta B \Theta^T$. If the adjacency matrices are from different models $\{\Theta_1, B_1\}, \cdots, \{\Theta_d, B_d\}$, then the mean of them would converge to a weighted probability matrix based on the proportion of samples from those models. So in the extreme case when sample size all goes to infinity, we can just use the mean adjacency matrix as our estimation of probability matrix P in LRT. Secondly, since the entry of adjacency matrix $[A_t]_{ij}$ is just a Bernoulli sequence based on P_t, we can easily prove that the log likelihood L_τ is increasing when τ is within $(1, \tilde{\tau}]$, and decreasing when $\tau \in [\tilde{\tau} + 1, T)$. So the maximum would occur at $\tau = \tilde{\tau}$. Detailed proof would be showed in Appendix B.

3.6.2 Multiple change points

When there are multiple change point, Theorem 3.2 shows that under similar assumption with single change point, LRT can detect the true change points by our heuristic searching procedure.

Assumption 2. *Suppose that we have heterogeneous sequence of independent adjacency matrix from* $k + 1$ *distributions, denoted* $\{\mathbb{M}_i\}_{i=0}^k$*. Specifically, let* $0 = \tau^{(0)} < \tau^{(1)} < \cdots < \tau^{(k+1)} = T$*. Then for* $i = 0, 1, \cdots, k$*, we have* $Z_{\tau^{(i)}+1}, \cdots, Z_{\tau^{i+1}} \sim \mathbb{M}_i$*, such that* $M_i \neq M_{i+1}$*. Let* $\{\delta_T\}$ *be a sequence of positive numbers such that* $\delta_T \to 0$ *and* $T\delta_T \to \infty$*, as* $T \to \infty$*.*

Theorem 3.2. *Suppose Assumption 2 holds. For number set* $\mathbb{A}_T \subset (\delta_T, 1 - \delta_T)$ *and* $x \in \mathbb{R}$*, define* $d(x, \mathbb{A}_T) = \inf\{|x - y| : y \in \mathbb{A}_T\}$*. Let* $\hat{\tau}_T$ *be the estimated change point as defined by likelihood ratio test and* $\mathbb{A}_T = \{\tau^{(1)}/T, \cdots, \tau^{(k)}/T\}$*. Then* $d(\hat{\tau}/T, \mathbb{A}_T) \xrightarrow{a.s.} 0$

The proof of Theorem 3.2 is similar with Theorem 3. since log likelihood function L_τ is concave within each interval between change points, It can only get maximized at any of boundary, which means the true change points.

4 Trend detection and outlier detection

4.1 Smooth trend detection

Other than assume there is one significant change at some specific time point, smooth trend may be a more realistic assumption in many areas. For example in gene co-expression networks, the community of genes is varying smoothly with subject ages (Liu et al. [2018]). Examine whether the smooth trend is happening in one network series and how the trend varying over time is another interesting problem.

As a first part, we want to find out whether the membership of nodes is varying smoothly over time. We focus on the membership for now, but the procedure can be easily extended to consider both membership and connectivities.

The null hypothesis is still the same as above $H_0 : \Theta_1 = \Theta_2 = \cdots = \Theta_T$. We then define the alternate hypothesis as the similarity of 2 networks: A_{t_1}, A_{t_2} is proportion to the gap of time $it = |t_2 - t_1|$.

We still use the difference of eigenvectors to measure the difference of adjacency matrices. For each gap value it, we can get a mean difference value over all t as the following.

$$\phi_t^{(it)} = ||U_t U_t^T - U_{t+it} U_{t+it}^T||_2, t \in [1, T - it]; \phi^{(it)} = \sum_t \phi_t^{(it)}/(T - it)$$

Linear regression can then be applied to test whether ϕ_{it} is proportion to it.

If there is smooth trend in the network series, the next thing we want to do is to test whether the trend is changing over time. For example, we want to know whether human brain change more rapidly at childhood than after growing to adult. Based on our definition of $\phi_t^{(it)}$, we can also test whether the difference with given it is related to t, e.g., fit $\phi_t^{(it=1)} \sim t$. If $\phi_t^{(it=1)}$ significantly increasing or decreasing with t, then we say the smooth trend is changing during time. More discussions about it can be found in Appendix A.

4.2 Outlier detection

One nature way to perform outlier detection is to compare A_t with all of the other $A_{-t} = \sum_{i \neq t} A_i/(T-1)$. After we get a difference $\beta_t = diff(A_t, A_{-t})$. Z-score of β_t can be calculated as

$$Z_t = \frac{\beta_t - \bar{\beta}}{s}$$

where $\bar{\beta}$ is the mean of all β_t and s is the standard deviation. Then Grubb's test (or any other test) can be performed with the Z-scores. It requires a large sample size to make Z-score works.

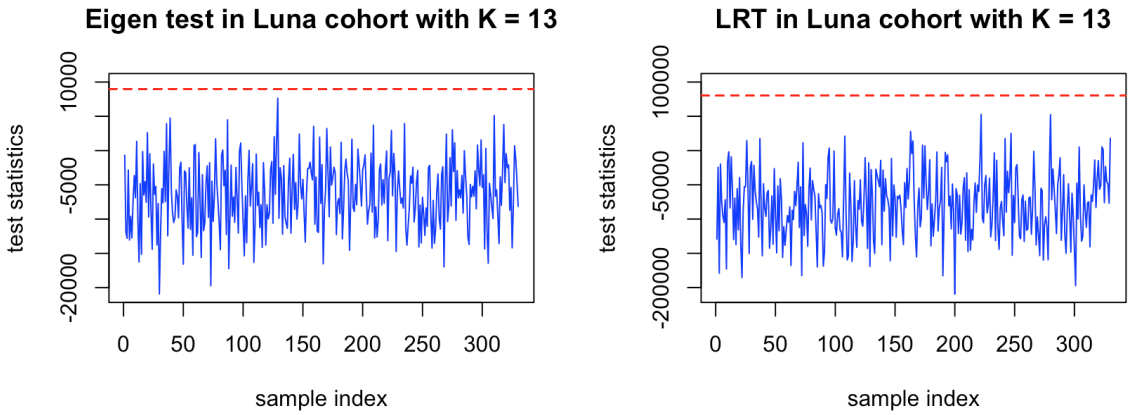

Figure 3: Change point detection on Luna cohort. The blue line shows original γ value and red line show values from bootstrap. No change point detected.

5 Application in real data

5.1 resting state fMRI data

We got rsfMRI data from 2 cohortS: longitudinal data of typically developing youth from Dr. Beatriz Luna's laboratory (Luna cohort). In Luna cohort, we have $N = 339$ the brain images from subjects with age between 13 and 26. All of the subjects in Luna cohort have typical brain developing, i.e. they do not hat any psychotic disease. Those 339 brain images are got from 223 different subjects. 50 of the subjects visit the lab twice in different ages so that we got two brain images for those 50 subjects. 57 of the them visit the lab three times so that we have three brain images for those subjects.

In PNC cohort, we got $N = 907$ brain images from 907 different subjects, i.e. there is no longitudinal data in PNC cohort. The subjects are also from ages between 13 to 26. 572 of the subjects have typical developed brain, 139 of the subjects have psychosis spectrum and 192 of the subjects have non-psychotic psychopathology. So another interesting problem is to find our the difference between typical developed brains and brains with psychopathology.

Figure 5 shows the change point detection results in Luna Cohort. There is no change point in those networks. Figure 6 and 7 also show that there is no outlier or smooth trend in those networks. PisCES is applied with window size $w = 5$ in community detection, i.e. we use previous 2 and future 2 networks to enhance the detection in specific age.

5.2 Gene co-expression networks

The transcriptional patterns of the developing primate brain are also of keen interest to neuroscientists and others interested in neurological and psychiatric disorders. Bakken et al. [25] provide a high-resolution transcriptional atlas of rhesus monkey (Macaca mulatta) built from recorded samples of gene expression, including expression of 9,173 genes that can be mapped directly to human. The samples span six prenatal ages from 40 embryonic days to 120 (E40-E120), and four postnatal ages from 0 to 48 months after birth (0M-48M). These ages represent key stages of development in the prenatal phase, and key milestones postnatally (newborn,juvenile, teen and mature). So we can

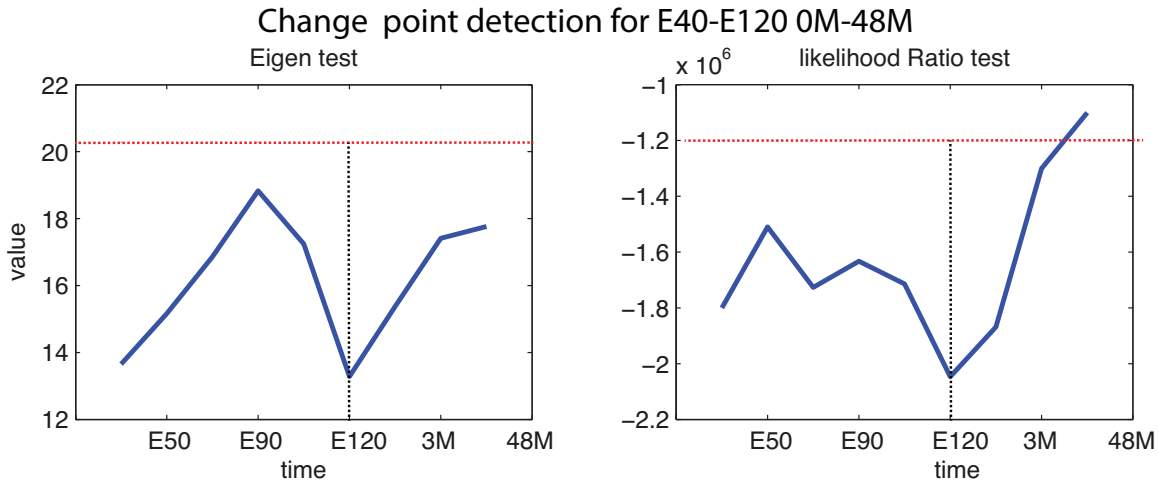

Figure 4: Change point detection of gene co-expression network. One change point is detected between E120-0M

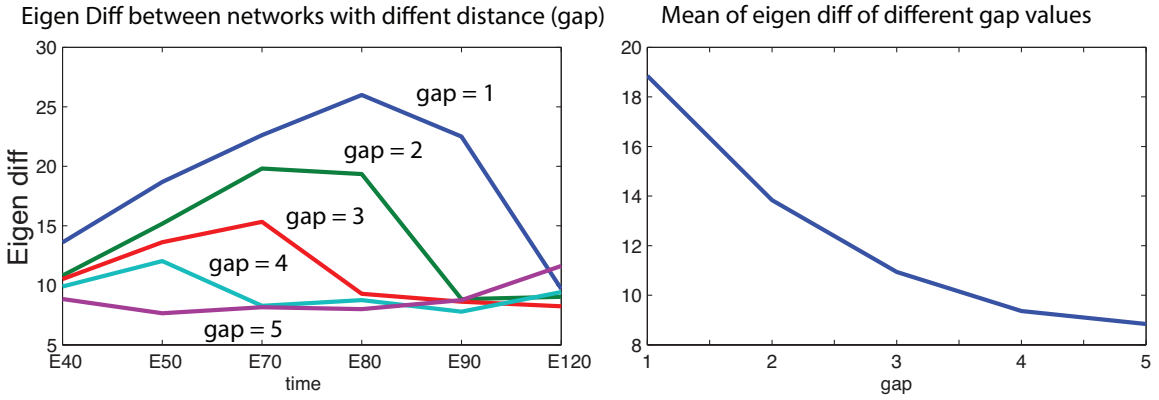

Figure 5: Smooth trend detection of gene co-expression network. The left plot shows Eigen diff at each ages with the network have specific gap with them. The right plot shows how the mean value changes over gap. "Eigen diff" is defined as $sum(SVD(U_{t1}, U_{t2}))$

also get 11 dynamic networks with this dataset.

Figure 8 shows the change point detection results of both eigen test and likelihood ratio test. We can see that birth (between E120 and 0M) is a significant change point. It also fits the ideas of scientists. But because there are only 11 networks, we are not able to detect further change point for prenatal and postnatal. One possible new permutation idea will be showed as a future plan.

Figure 9 shows the trend detection results. The difference of eigenvectors are showed with different gap of ages. It is easy to find that bigger gap results in larger difference.

6 Next steps

Further works focus on further theoretical results, sample permutation ideas and analysis of longitudinal data.

Theoretical results about Eigen test Theoretical results have already been build for likelihood ratio test. But the corresponding results for Eigen test still need more effort. Since A_{left} and A_{right} can be assumed be to averaging adjacency matrices of infinite samples, they are converging into the probability matrix P_{left} and P_{right}. Based on the theories about spectral clustering in SBM, the eigen-decomposition of $P = \Theta B \Theta^T$ where $\{\Theta, B\}$ parametrize a SBM and Θ is just the community assignment in SBM. So if all of the networks forming A_L are from the same model, i.e., there is no further change point in the left part, then the eigenvectors $U \to \Theta$ is just the true assignment. However, if the averaged matrix is formed from a mixture of different models, things could be more complex. For example if $A_1, A_2, \cdots, A_{t_1}$ are from one SBM with K_1 communities and $A_{t_1+1}, \cdots, A_{t_2}$ are generated with another model with K_2 communities, then $A_L = \sum_{t=1}^{t_2} A_t / t_2$ should have $K_1 + K_2$ communities. How it influence the test statistics would be very an interesting problem.

Moving window and stationary bootstrap Since mixture of models may cause problem with Eigen test method, a possible way to solve it to apply moving windows. The moving window strategy has been widely applied in change point detection for univariate times series () and covariance matrices (Matteson and James [2014]). By pre-defining a window size l, for each candidate change point τ, we can only use left l and right l networks to calculate A_L and A_R, i.e. $A_L = \sum_{t=\tau-l}^{\tau} \frac{A_t}{l}$, $A_R = \sum_{t=\tau+1}^{\tau+l} \frac{A_t}{l}$. By choosing one appropriate window size l, we can avoid mixture of models in one window. However, how to choose the window size would be another problem. One nature approach is to choose the one returns best test statistics. But it would be very time-consuming. Stationary bootstrap procedure would also be needed for moving window approaches to avoid mixture of models in the bootstrap sample.

Sample permutation for gene co-expression networks As mentioned in section 5, because there are only 11 ages in gene co-expression data, we are not able to perform further change point detection after finding the first change. Fortunately, if we are using correlation matrices and each correlation matrix at one age actually is formed with multiple samples. Assume $A_1, A_2, \cdots, A_3$ are the correlation matrices, one entry of each matrix A_{ij} is the correlation of gene expression levels over different samples in that age. For example, there are 48 samples for A_1, then $\{A_1\}_{ij}$ is the Pearson's correlation between gene i and gene j over those 48 samples. Since we are not able to get good significant level by permuting correlation matrices, we can permute the samples instead. This procedure is similar with what Yu's paper did for partition of long-term time series. In order to make it work, a Gaussian graphical model instead of SBM should be assumed for the samples and corresponding theoretical results can also be built. We believe the sample permutation could also be a very powerful way to deal with problems raised by lack of series size in many different areas.

Analysis of longitudinal data Longitudinal data is defined as networks from the same subject with different visits (ages). As mentioned above, we have longitudinal data for both Luna and PNC cohort. About 1/3 of subjects visit the lab two or three times in those 2 cohort. In the change point detection algorithms, we assume that the network series are independent with each other. However, those longitudinal data are of course dependent and may influence the results. The analysis of longitudinal data including 2 parts: (i) if we have 2 networks from subjects with different ages, by giving one of those networks, can we identify the other one based on their similarity. Once recent paper (Finn et al. [2015]) show that 2 rs-fMRI images scanned by consecutive days from the same subject can be identified from a list of images from different people. Based on this results, we also want to know whether images scanned by different ages from the same people can also be identified, and whether the identification rate related to subject ages. (2) find out whether the longitudinal data influence the community detection and change point detection results. For both of those 2 parts, we can also test whether the identification or influence are related to specific sample ages.

Data visualization How to visualize the dynamic networks series is another interesting and necessary task. In PisCES paper, we applied sankey plot to show the nodes flow between networks. Similar ideas may also be used in the change point or smooth trend visualization.

References

Valeriy Avanesov and Nazar Buzun. Change-point detection in high-dimensional covariance structure. *arXiv preprint arXiv:1610.03783*, 2016.

Yun Chi, Xiaodan Song, Dengyong Zhou, Koji Hino, and Belle L. Tseng. *Evolutionary Spectral Clustering by Incorporating Temporal Smoothness*. 2007.

Ivor Cribben and Yi Yu. Estimating whole-brain dynamics by using spectral clustering. *J R Stat Soc Ser C Appl Stat*, 66(3):607–627, 2017a. ISSN 1467-9876. doi: 10.1111/rssc.12169. URL `http://dx.doi.org/10.1111/rssc.12169`.

Ivor Cribben and Yi Yu. Estimating whole-brain dynamics by using spectral clustering. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):607–627, 2017b.

Ivor Cribben, Ragnheidur Haraldsdottir, Lauren Y Atlas, Tor D Wager, and Martin A Lindquist. Dynamic connectivity regression: determining state-related changes in brain connectivity. *Neuroimage*, 61(4):907–920, 2012.

Ivor Cribben, Tor Wager, and Martin Lindquist. Detecting functional connectivity change points for single-subject fmri data. *Frontiers in computational neuroscience*, 7:143, 2013.

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience*, 18(11):1664, 2015.

Amir Ghasemian, Pan Zhang, Aaron Clauset, Cristopher Moore, and Leto Peel. Detectability thresholds and optimal algorithms for community structure in dynamic networks. *Phys Rev X*, 6 (3), 2016.

P. W. Holland, K. B. Laskey, and S Leinhardt. Stochastic blockmodels:first steps. *Social Networks*, 5(MR0718088):109–137, 1983.

Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 038788145X, 9780387881454.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Ann Stat*, 43(1):215–237, 2015.

Martin A Lindquist, Yuting Xu, Mary Beth Nebel, and Brain S Caffo. Evaluating dynamic bivariate correlations in resting-state fmri: a comparison study and a new approach. *NeuroImage*, 101: 531–546, 2014.

Fuchen Liu, David Choi, Lu Xie, and Kathryn Roeder. Global spectral clustering in dynamic networks. *Proceedings of the National Academy of Sciences*, 2018. URL `http://www.pnas.org/content/early/2018/01/10/1718449115`.

Scott Marek, Kai Hwang, William Foran, Michael N. Hallquist, and Beatriz Luna. The contribution of network organization and integration to the development of cognitive control. *PLOS Biology*, 13(12):1–25, 12 2016. doi: 10.1371/journal.pbio.1002328. URL `https://doi.org/10.1371/journal.pbio.1002328`.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *J R Stat Soc Series B Stat Methodol*, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12200. URL `http://dx.doi.org/10.1111/rssb.12200`.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345, 2014.

Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010. ISBN 0199206651, 9780199206650.

Nam P. Nguyen, Thang N. Dinh, Yilin Shen, and My T. Thai. Dynamic social community detection and its applications. *PLoS One*, 9(4):e91431, 2014. doi: 10.1371/journal.pone.0091431.

Dane Taylor, Sean A Myers, Aaron Clauset, Mason A Porter, and Peter J Mucha. Eigenvector-based centrality measures for temporal networks. *Multiscale Modeling & Simulation*, 15(1):537–574, 2017.

Kevin S Xu and Alfred O Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562, 2014.

Appendices

A Simulation results

A.1 Change point detection

We use SBM to generate different simulation cases. For all of the cases, there are $n = 200$, $K = 4$ communities, and the membership of $g_i^{(1)} \sim Multinolmial(1/K, 1/K, \cdots, 1/K)$. We further let $z \in \mathbb{R}^T$ denote the percent of change of membership during time. For instance, $z(2) = 0.5$ means the 50% of the nodes re-generate there membership at time $t = 2$, i.e.,

$$g_i^{(t+1)} = \begin{cases} g_i^{(t)} & \text{with prob. } 1 - z(t) \\ Multinolmial(1/K, \cdots, 1/K) & \text{otherwise} \end{cases}$$

In all of our simulations, the original connectivity matrix B is fixed as the following.

$$B^{(1)} = \begin{Bmatrix} 0.25 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.25 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.25 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.25 \end{Bmatrix}$$

When we only consider the change of membership, we keep B consistent over time.

We considered 3 cases in this simulation:

- Single change point of membership with growing T. For specific T, $z(\lceil T/2 \rceil) = 0.1, 0.2$ or 0.5, otherwise, $z(t) = 0$. We considered $T = 8, 12, 16, 20$ in this case. For each T and label change probability, we applied 100 repetitions.
- Single change point of connectivity with growing T. $z(t) = 0$ for all t. But when $t > \lceil T/2 \rceil$, $B^{(t)} = 2B^{(1)}$. We considered $T = 8, 12, 16, 20$ in this case. For each T and label change probability, we applied 100 repetitions.
- Double change points of membership with growing T. For specific T, $z(\lceil T/3 \rceil) = z(\lceil 2T/3 \rceil) = 0.1, 0.2$ or 0.5 , otherwise, $z(t) = 0$. We considered $T = 12, 18, 24, 30$ in this case. For each T and label change probability, we applied 100 repetitions.

We apply two measurement to compare the results: (1) exact accuracy which count the number trials when $\hat{\tau} = \tilde{\tau}$; (2) approximate accuracy which calculate the number of trials when $|\hat{\tau} - \tilde{\tau}| < T/10$. LRT perform perfectly in every cases (Figure 6, 7, 8). While ET is less powerful when the probability of label change is small.

B Proofs

Proof of Theorem 3.1. **Part 1**: Let $P = \Theta B \Theta$ is the probability matrix where $P_{ij} = B_{g_i, g_j}$. P_1 is the probability matrix for the first model and P_2 is for the second model. We first prove that under Assumption 1, the true or mixtured probability matrix P can be estimated.

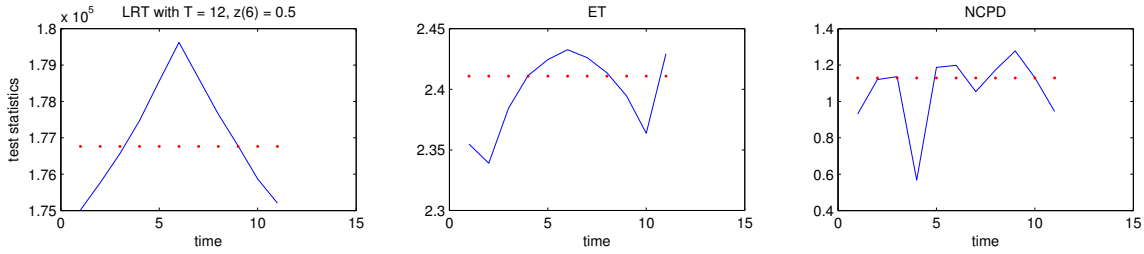

Figure 6: One example of how LRT, ET and NCPD work.

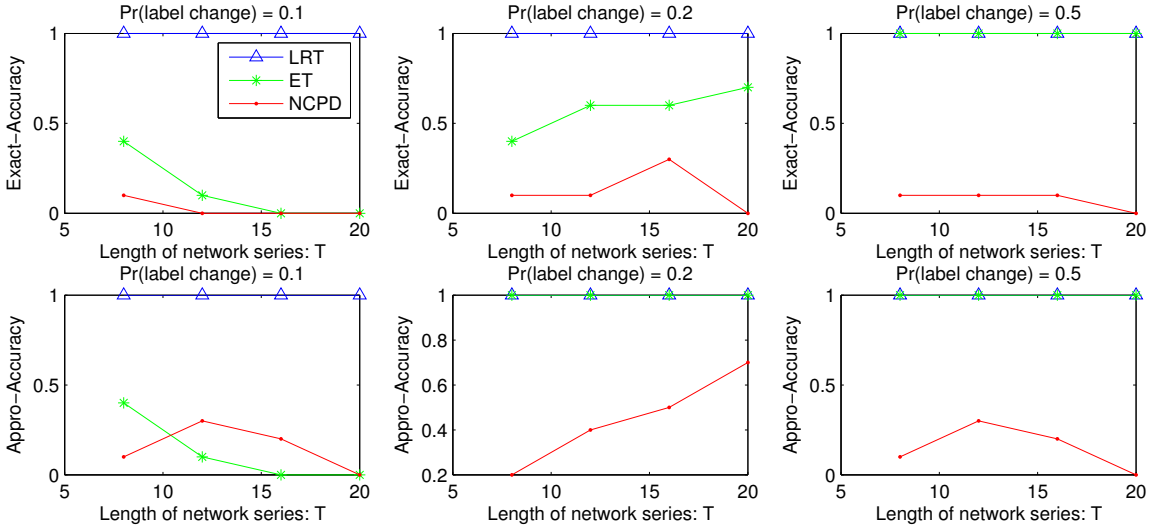

Figure 7: Comparison of LRT, ET and NCPD when there is single change point of membership.

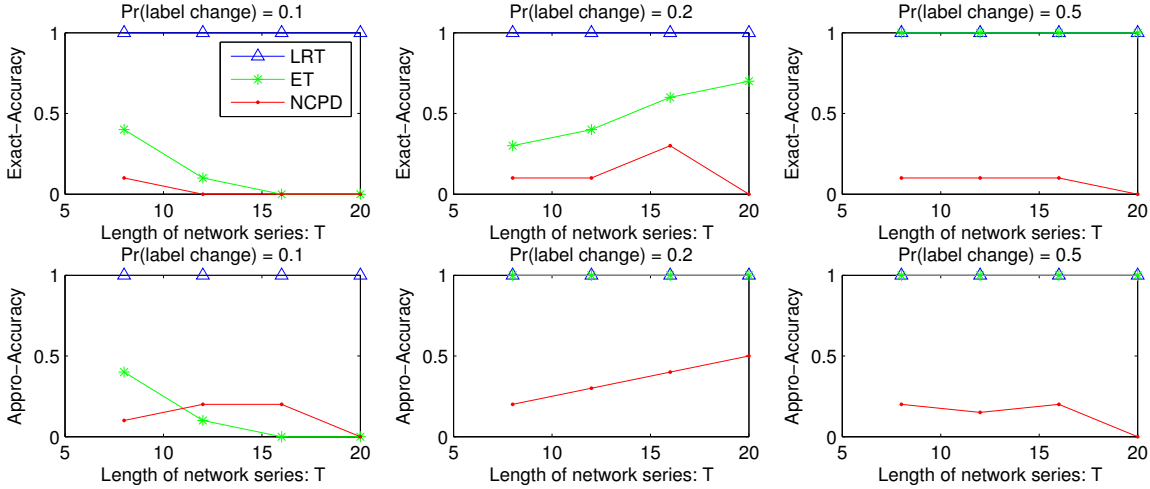

Figure 8: Comparison of LRT, ET and NCPD when there is double change points of membership.

Use $A_{left} = \sum_{t=1}^{\tau} A_t/\tau$ as an example, if $\tau < \tilde{\tau}$, $\lim_{\tau \to \infty} A_{left} \to P_1$, then based on lemma 2.1, we can estimate the true membership g_1 and B_1 as well as P_1 by spectral clustering. If $\tau > \tilde{\tau}$, then $\lim_{(\tau-\tilde{\tau})\to\infty} A_{left} \to \frac{\tilde{\tau}P_1+(\tau-\tilde{\tau})P_2}{\tau} := P_{left}$, this P_{left} can also be well estimated by spectral clustering.

Part 2: Under assumption 1, the log likelihood function L_τ is maximized around the true change point $\tau = \tilde{\tau}$.

$\forall \epsilon > 0$, $P(\lim_{T\to\infty} |\hat{\tau}_T/T - \tilde{\tau}/T| > \epsilon) > 0$ is equivalent with $\forall |\epsilon| > 0$, $P(\lim_{T\to\infty} \Delta_{\tilde{\tau}+\epsilon T} > \Delta_{\tilde{\tau}}) = 0$. Since L_{null} is the same for every time point, we only compare $L_{\tilde{\tau}}$ and $L_{\tilde{\tau}+\epsilon T}$. Notice that the log likelihood of one adjacency matrix A is the summation of log likelihood at every edge of A.

$$
\begin{aligned}
L_\tau &= \sum_{t=1}^{\tau} L(A_t, g_{left}, P_{left}) + \sum_{t=\tau+1}^{T} L(A_t, g_{right}, P_{right}) \\
&= \sum_{t=1}^{\tau} \sum_{ij} [A_t]_{ij} \log(\hat{B_{left}}_{g_{l\hat{e}ft_i}, g_{l\hat{e}ft_j}}) + (1 - [A_t]_{ij}) \log(1 - \hat{B_{left}}_{g_{l\hat{e}ft_i}, g_{l\hat{e}ft_j}}) \\
&\quad + \sum_{t=\tau+1}^{T} \sum_{ij} [A_t]_{ij} \log(\hat{B_{right}}_{g_{r\hat{i}ght_i}, g_{r\hat{i}ght_j}}) + (1 - [A_t]_{ij}) \log(1 - \hat{B_{right}}_{g_{r\hat{i}ght_i}, g_{r\hat{i}ght_j}}) \\
&= \sum_{ij} (\sum_{t=1}^{\tau} [A_t]_{ij} \log(\hat{p}_1)) + (1 - [A_t]_{ij}) \log(1 - \hat{p}_1) \\
&\quad + \sum_{t=\tau+1}^{T} [A_t]_{ij} \log(\hat{p}_2)) + (1 - [A_t]_{ij}) \log(1 - \hat{p}_2)) \\
&= \sum_{ij} l_\tau(ij)
\end{aligned}
$$

where $\hat{p}_1 = \hat{B_{left}}_{g_{l\hat{e}ft_i}, g_{l\hat{e}ft_j}}$ and $\hat{p}_2 = \hat{B_{right}}_{g_{r\hat{i}ght_i}, g_{r\hat{i}ght_j}}$. So we only need to prove that $\forall \epsilon > 0$, $\forall i, j$, $P(\lim_{T\to\infty} l_{\tilde{\tau}+\epsilon T}(ij) > l_{\tilde{\tau}}(ij)) = 0$. It actually turns the problem as the log likelihood in Bernoulli sequence. Based on our results in Part 1, given $\tau = \tilde{\tau} + \epsilon T$,

$$\lim_{T\to\infty, \epsilon T\to\infty} l_\tau(ij) = (\tilde{\tau}+\epsilon T)(\hat{p}_1 log(\hat{p}_1)+(1-\hat{p}_1)(1-log(\hat{p}_1))+(T-\tilde{\tau}-\epsilon T)(p_2 log(p_2)+(1-p_2)(1-log(p_2))$$

where $\hat{p}_1 = \frac{\tilde{\tau}p_1+(\epsilon T)p_2}{\tilde{\tau}+\epsilon T}$, $p_1 = [P_1]_{ij}, p_w = [P_2]_{ij}$ is the probability parameter from the original model in Assumption 1. It is easy to show that this function maximized at $\epsilon = 0$. □