Understanding developing networks with stochastic block models*

Zitong Zhang[†,‡,??], **Shizhe Chen**[§,??]

Address of the First and Second authors
Usually a few lines long
e-mail: zztzhang@ucdavis.edu

Abstract: In neuroscience, it is not well understood how newborn neurons form a mature nervous system, due to the lack of observations. A recent study (Wan et al. 2019) made available a dataset of this functional maturation process on zebrafish. This novel data, however, introduce inherent challenges for the analysis of the formation of the nervous system. First, the formation process is transient by nature. The non-stationarity of the process makes the amount of data pale in comparison to the size of the neural network. Moreover, combining observations on multiple subjects are not straightforward since the neural circuits are not identical across subjects. In this talk, we propose a model for describing the emergence of a coordinated network from isolated nodes. To this end, we adapt and generalize the stochastic block model for random graphs. The proposed method learns the transferable features across subjects, while allowing for individual variabilities. Briefly, the proposed method classifies nodes into different functional groups by identifying typical connecting behavior. We further employ the shape invariant models to handle the nodal delays due to neuron-specific delays. We establish the consistency and minimax optimality of the proposed estimator. We demonstrate the performance of our algorithm on simulation experiments, and on the real zebrafish dataset.

MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35.
Keywords and phrases: sample, LaTeX 2_ε.

1. Introduction

Dynamic networks emerge in many area, such as the neuronal network in the brain during disease[] or learning tasks[], the social network in a time period[], etc. A lot of work has been done in order to study the dynamic networks, but most of these research focus on analyzing the dynamics in a well-developed network. In this paper, we concentrate on the growing networks where isolated nodes develop into a mature functioning network. To the best of our knowledge, there is no existing study about the growing networks, partly because of the lack of data. Fortunately, the data collected by Wan et al. [7] provides us the possibility to pursue this study.

*Footnote to the title with the “thankstext” command.
†Some comment
‡First supporter of the project
§Second supporter of the project

In this paper, we focus on the neural data provided by [7]. Since the neuronal network is complicated, we try to break it down and propose to model it by identifying the typical roles of individual neurons. As supported by [7], neurons in a growing network play different roles in terms of active time and connecting patterns. The connecting pattern of a neuron can be described as the occurrence time of edges that include this neuron. Neurons with the same roles perform similar activities and thus have similar connecting patterns. Learning these roles can help us understand the development procedure of growing networks.

However, being able to identify the typical roles in a single growing network is infeasible because the growing network is transient — there is only one measurement of a growing network due to its non-stationarity. Due to this limit, we combine networks of other subjects as additional samples. This is not trivial, because the neurons in different networks are not one-to-one mapped, and the roles identified from different subjects are not transferable . This constrains our ability to study the common features across subjects. For this reason, we propose to use the stochastic block model as it allows to combine multiple networks and hence resolve the above problems. [SBM is.... It has been applied to ...]

There are, nevertheless, some unique properties of the data that are beyond the scope of the stochastic block model. First, the connection between two neurons is measured over time. Second, the connecting pattern between two neurons are determined not only by their roles but also by their active time, which varies from node to node. Third, the connection is also effected by the spatial distance between neurons — connection cannot occur if two neurons are too away from each other. To incorporate such uniqueness, we propose a generalized stochastic block model in this paper. [Note that these properties also appear in other problems, e.g. venmo...]

Related work

The stochastic block model is first proposed by Holland, Blackmond and Leinhardt [3]. It has many dynamic extensions, Yang et al. [10], Xu and Hero [9], Matias and Miele [4], Xu [8] use the Markov chain to model the time-varying connecting probabilities and/or the clustering matrix. EM algorithm or iterative optimization algorithm is commonly used for inference.

Matias, Rebafka and Villers [5] adapt the stochastic block model to the context of recurrent interaction events in continuous time, where the recurrent events are modeled by Poisson processes with intensities determined by the nodes' group memberships. The maximum likelihood estimator is proposed, but no theoretical analysis is available in the paper.

Optimal rate of convergence is also studied. Gao, Lu and Zhou [2] provides an optimal rate under the mean squared error for the stochastic block model. Pensky [6] derives a penalized least square estimator in a dynamic network setting, and shows that the estimator satisfies an oracle inequality and attains the minimax lower bound for the risk.

Contribution

In this paper, we propose a method for analyzing the growing networks. Our method is able to identify the roles of individual nodes and the connecting patterns. We adapt the stochastic block model to the growing networks context by generalizing the connecting probabilities to intensities of point processes. In addition, we incorporate the time delay of each node so that our model is able to handle the network where nodes become active over time. We derive a least square estimator and show that the estimator converges [in a certain rate]. Finally, an algorithm combining the k-means method and the shape invariant method is proposed for estimation.

Future work

Future working directions include (but not limited to) (i) identifying clusters with similar connecting pattern but different active time phase or different vertex degree, (ii) incorporating the movement of nodes, (iii) seeking for a convex relaxation method that convexify over both clustering matrix and time lags (convex relaxation can also be adapted to solve the penalized least square problem in Pensky [6]), (iv) try other clustering methods.

Organization

The rest of this paper is organized as follows. In Section 2, we review the stochastic block model and introduce the proposed dynamic generalization of the stochastic block model. We introduce the least square estimator and the estimation algorithm in Section 3. Theoretical results are provided in Section 4. Section 5 shows the numerical experiments.

2. Model

2.1. Stochastic block model

A set of n nodes $\Gamma = \{1, \cdots, n\}$ is partitioned into k clusters $\Gamma_1, \cdots, \Gamma_k$. The cluster of node i is represented by $z_i \in \{1, \cdots, k\}$, and the vector of clusters is $\mathbf{z} = (z_i)_{i=1}^n$. Define the adjacency matrix $\mathbf{A} \in \{0,1\}^{n\times n}$ where $A_{i,j} = 1$ if an edge is observed between node i and node j and $A_{i,j} = 0$ otherwise. We set $A_{i,i} \equiv 0$ for any $i = 1, \cdots, n$, and assume that $A_{i,j}$'s are conditionally independent given the cluster vector $\mathbf{z}$:

$$A_{i,j} | z_i = q, z_j = l \overset{ind}{\sim} \text{Bernoulli}(C_{q,l}), \qquad i \neq j,$$

where $\mathbf{C} \in [0,1]^{k\times k}$ denotes the connecting probability matrix.

2.2. Dynamic stochastic block model for point processes

We consider a developing network where pairwise connections between n nodes are observed during time interval $[0,T]$. Each node belongs to one of k groups denoted by $\Gamma_1,\cdots,\Gamma_k$, and the membership for node i is denoted by z_i. We assume that the connection between nodes are completely determined by their group memberships. Specifically, the connection between node i and j is modeled by a point process $N_{i,j}(\cdot)$ with intensity function

$$\lambda_{i,j}(t)=\lambda_{z_i,z_j}(t), \qquad t\in[0,T], \quad i,j=1,\cdots,n,$$

Similar to the stochastic block model, we set $\lambda_{i,i}(\cdot)\equiv 0$ for $i=1,\cdots,n$.

Objective function

We propose the following least-square loss function that measures the overall distance between the observed connection distributions and the true ones for each pair of clusters

$$\min_{\{\Gamma_q\}_{q\in[k]},\{f_{q,l}\}_{q,l\in[k]}} \sum_{q,l\in[k]} w_{q,l}\cdot d^2(N_{\Gamma_q,\Gamma_l},f_{q,l}), \tag{2.1}$$

where $w_{q,l}$ is the weight associated with connection between cluster Γ_q and Γ_l, $N_{\Gamma_q,\Gamma_l}(\cdot)=\sum_{i\in\Gamma_q,j\in\Gamma_l}N_{i,j}(\cdot)$ is the aggregated point process representing connection between Γ_q and Γ_l. The distance $d(N,f)$ is defined as

$$d(N,f)=d(F_N,F)=\left(\int|F_N(t)-F(t)|^2\mathrm{d}t\right)^{1/2}.$$

where F_N is the empirical cumulative distribution function of the point process $N(\cdot)$, and F is the cumulative distribution function corresponding to f.

Incorporating time delay and spatial information

There are additional challenges arose in real problems. In the neuronal network, neurons have different active time, and only connect with neurons near them. To incorporate such restrictions, we assume the intensity functions take the following form

$$\lambda_{i,j}(t)=\lambda_{z_i,z_j}(t-\tau_{i,j})\cdot\mathbf{1}_{\{d_{i,j}\le d^*\}}, \quad t\in[0,T], \quad i,j=1,\cdots,n,$$

here λ_{z_i,z_j} is extend by zero outside $[0,T]$.

With this model, we can adapt our objective function (2.1) as

$$\min_{\substack{\{\Gamma_q\}_{q\in[k]},\\ \{f_{q,l}\}_{q,l\in[k]},\\ \{\tau_{i,j}\}_{i,j\in[n]}}} \sum_{q,l\in[k]} w_{q,l}\cdot\tilde{d}^2(\tilde{N}_{\Gamma_q,\Gamma_l},f_{q,l}), \tag{2.2}$$

where $\tilde{N}_{\Gamma_q,\Gamma_l}(\cdot) = \sum_{i\in\Gamma_q, j\in\Gamma_l} N_{i,j}(\cdot + \tau_{i,j})$ is the point process between cluster Γ_q and Γ_l after aligning all events by their time delays, and the distance function $\tilde{d}$ is the shift-invariant version of d defined as

$$\tilde{d}(N, f) = \tilde{d}(F_N, F) = \inf_{\tau} \left(\int |F_N(t-\tau) - F(t)|^2 \mathrm{d}t \right)^{1/2},$$

For notation convenience, we will use d to represent this distance function in the rest of this paper.

3. Method

3.1. Evaluating the distance function

In practice, the distance between a point process N and a probability distribution function f is evaluated by computing distance between their (smoothed) probability distribution functions, because the distance between cumulative distribution functions can be misleading in some certain cases. We illustrate this through figure 1. Although the black dash pdf is closer to the green pdf than the red pdf, it is closer to the red curve in terms of cdf.

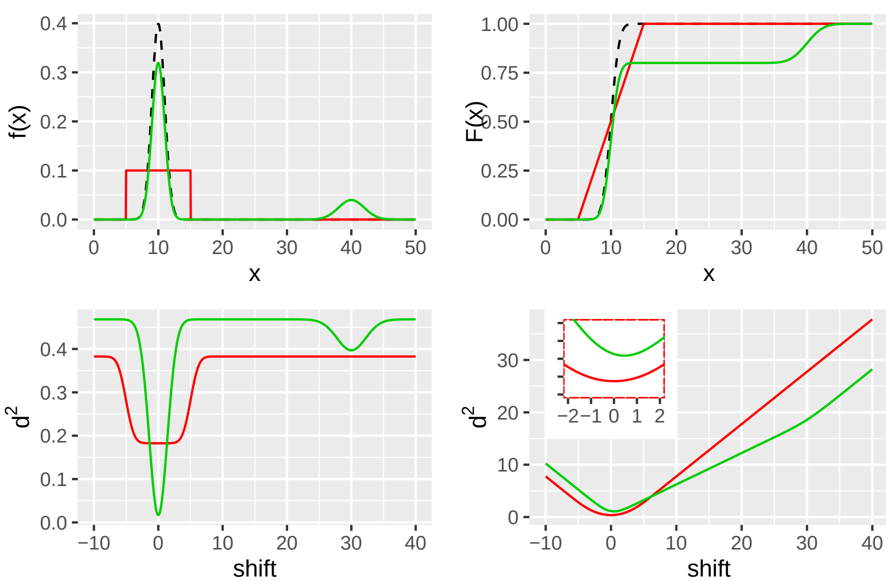

Fig 1: Distance between cdfs is misleading. Top left: pdf to be shifted (black dash line) and target pdf (green and red solid line). Top right: corresponding cdf. Bottom left: distance between pdfs as a function of shift. Bottom right: distance between cdfs as a function of shift.

However, as shown in figure 2, aligning cdf can help us to locate the minimizer of distance function, especially when their is a flat area between two pdfs.

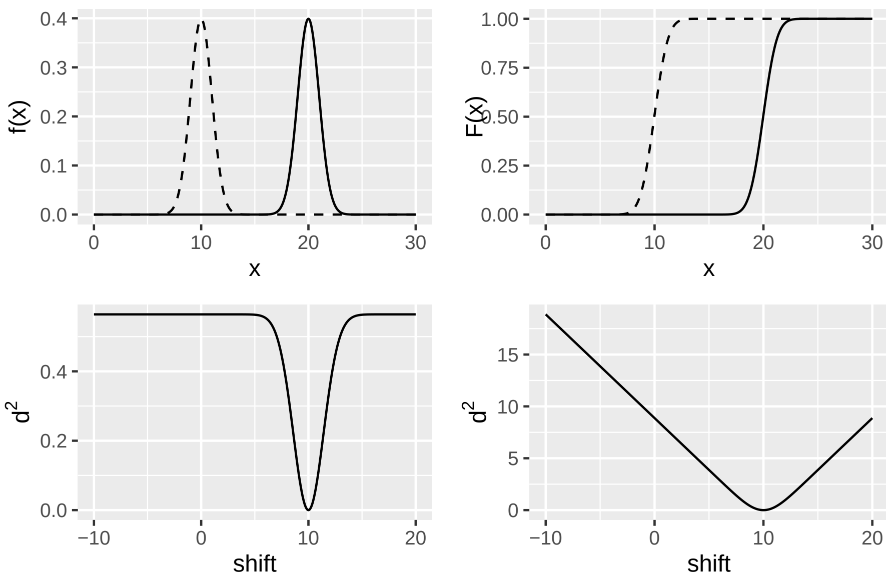

Fig 2: cdf can help to locate the minimizer. Top left: pdf to be shifted (dash line) and target pdf (solid line). Top right: corresponding cdfs. Bottom left: distance between pdfs as a function of shift. Bottom right: distance between cdfs as a function of shift.

Based on these properties of cdf and pdf, we align pdfs using gradient descent algorithm with initialization obtained by aligning their corresponding cdfs. We use the shape-invariant method proposed by Bigot and Gendre [1]. The idea is to find the optimal shift that aligns f and g best in the Fourier domain. Let θ_j and $\gamma_j, j = -(N-1)/2, \cdots, (N-1)/2$ be the discrete Fourier coefficients of f and g, where N is the length of discretized f and g. The time shift parameter τ can be estimated by solving the following problem by gradient descent

$$\hat{n} = \underset{|n|\le(N-1)/2}{\arg\min} \sum_{|j|\le(N-1)/2} \left|\theta_j e^{\mathrm{i}2\pi jn/N} - \gamma_j\right|^2, \tag{3.1}$$

where $n = N \cdot \tau/2T$. Initialization is given by the alignment result of their cdfs. For details see A.
[add details for obtaining smooth pdf.]

3.2. Algorithm

[Jump from estimating $\tau_{i,j}$'s to τ_i's.]

We iterate between two main steps: re-center step and re-cluster step. In the re-center step, we update connecting patterns based on current clustering and time shifts; while in the re-cluster step, we update the clusters and time shifts based on updated connecting patterns.

3.2.1. Re-center step

Based on current estimation of clusters $\{\Gamma_q^{(c)}\}_{q\in[k]}$, we can obtain an (unaligned) point process for each pair of clusters by aggregating connecting times over all pair of nodes in the corresponding two clusters

$$N_{\Gamma_q^{(c)},\Gamma_l^{(c)}}(\cdot) = \sum_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}} N_{i,j}(\cdot).$$

Incorporating estimated time shifts $\{\tau_i^{(c)}\}_{i\in[n]}$ yields two possible alignment: (a) $\tilde{N}_{\Gamma_q^{(c)},\Gamma_l^{(c)}}(\cdot) = \sum_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}} N_{i,j}(\cdot+\tau_i^{(c)})$ and (b) $\tilde{N}_{\Gamma_q^{(c)},\Gamma_l^{(c)}}(\cdot) = \sum_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}} N_{i,j}(\cdot+\tau_j^{(c)})$. We use the variance of the shifted events as a measure of goodness-of-alignment — the one producing the smaller variance is adopted. To be more specific, let

$$\begin{aligned}\mathrm{var}_1 &= \mathrm{var}\left[\{T_{i,j} - \tau_i^{(c)}\}_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}}\right],\\ \mathrm{var}_2 &= \mathrm{var}\left[\{T_{i,j} - \tau_j^{(c)}\}_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}}\right],\end{aligned}$$

where $T_{i,j}$ denotes the edging time between node i and j ($T_{i,j} = \infty$ if there is no edge). Then

$$f_{q,l}^{(u)} = \begin{cases}\text{density}\left(\{T_{i,j} - \tau_i^{(c)}\}_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}}\right), & \text{if } \mathrm{var}_1 < \mathrm{var}_2,\\ \text{density}\left(\{T_{i,j} - \tau_j^{(c)}\}_{i\in\Gamma_q^{(c)},j\in\Gamma_l^{(c)}}\right), & \text{otherwise.}\end{cases} \tag{3.2}$$

[Add how to deal with negative event time after alignment.]

3.2.2. Re-cluster step

Given $\{f_{q,l}^{(u)}\}_{q,l\in[k]}$ and $\{\Gamma_q^{(c)}\}_{q\in[k]}$, we propose to solve a simplified version of the objective function in (2.2)

$$\min_{\substack{\{\Gamma_q\}_{q\in[k]},\\ \{\tau_{i,j}\}_{i,j\in[n]}}} \sum_{q,l\in[k]} w_{q,l}\cdot d^2(\tilde{N}_{\Gamma_q,\Gamma_l^{(c)}}, f_{q,l}^{(u)}). \tag{3.3}$$

Setting the weights to be $w_{q,l} \propto \deg(\Gamma_q, \Gamma_l^{(c)})$, we derive an upper bound of (3.3)

$$\begin{aligned}
&\sum_{q,l\in[k]} \deg(\Gamma_q, \Gamma_l^{(c)}) \cdot d^2(\tilde{N}_{\Gamma_q,\Gamma_l^{(c)}}, f_{q,l}^{(u)}) \\
&\quad \le \sum_{q,l\in[k]} \deg(\Gamma_q, \Gamma_l^{(c)}) \cdot \left[\sum_{i\in\Gamma_q} \frac{\deg(i, \Gamma_l^{(c)})}{\deg(\Gamma_q, \Gamma_l^{(c)})} \cdot d^2\left(\tilde{N}_{i,\Gamma_l^{(c)}}, f_{q,l}^{(u)}\right) \right] \\
&\quad = \sum_q \sum_{i\in\Gamma_q} \left[\sum_l \deg(i, \Gamma_l^{(c)}) \cdot d^2\left(\tilde{N}_{i,\Gamma_l^{(c)}}, f_{q,l}^{(u)}\right) \right] \\
&\quad \le \sum_i \left[\sum_l \sqrt{\deg(i, \Gamma_l^{(c)})} \cdot d\left(\tilde{N}_{i,\Gamma_l^{(c)}}, f_{z_i,l}^{(u)}\right) \right]^2 .
\end{aligned}$$

The first inequality is obtained by applying Jensen's inequality.
This upper bound leads to our re-cluster step

$$z_i^{(u)} = \arg\min_{q\in[k],} \sum_l \sqrt{\deg(i, \Gamma_l^{(c)})} \cdot d\left(\tilde{N}_{i,\Gamma_l^{(c)}}, f_{q,l}^{(u)}\right). \tag{3.4}$$

Update time shifts

Based on current clusters, we are able to get the unaligned connecting point process $N_{i,\Gamma_q^{(u)}}$ between each node i and each cluster $\Gamma_q^{(u)}$. We will align nodes from the same cluster and set the minimum time shifts in each cluster as zero for identifiability reason.

For each cluster $\Gamma_q^{(u)}$, randomly select a node $i^* \in \Gamma_q^{(u)}$ as a point of reference. The time lag between node $i \in \Gamma_q^{(u)}$ and i^* is then determined by

$$\tau_{(i,i^*),l} = \arg\min_\tau d(N_{i,\Gamma_l^{(u)}}(\cdot + \tau), N_{i^*,\Gamma_l^{(u)}}(\cdot)), \quad l \in [k], \tag{3.5}$$

$$\tau_{i,i^*} = \max\left(\big| \max_{l\in[k], l\neq q} \tau_{(i,i^*),l} \big|, \big| \min_{l\in[k], l\neq q} \tau_{(i,i^*),l} \big| \right). \tag{3.6}$$

[Add explanation.]
Finally, the time shifts are updated as

$$\tau_i^{(u)} = \tau_{i,i^*} - \min_{i\in\Gamma_q^{(u)}} \tau_{i,i^*}, \quad i \in \Gamma_q^{(u)}. \tag{3.7}$$

3.3. Initialization

Naive k-means algorithm is applied to obtain initialization of clusterings and time shifts. Time shifts are obtained by aligning each aggregated point process with the first point process, followed by shifting all nodes together so that the minimum time shift is zero. Clustering $\{\Gamma_q^{(c)}\}_{q\in[k]}$ is then initialized by applying k-means++ algorithm to

$$\begin{bmatrix} T_{1,\cdot} - \tau_1^{(c)} \\ T_{2,\cdot} - \tau_2^{(c)} \\ \vdots \\ T_{n,\cdot} - \tau_n^{(c)} \end{bmatrix}.$$

3.4. Summary of algorithm

The algorithm is summarized as following.

Algorithm 1: [name of algorithm]

Initialize $\{\Gamma_q^{(c)}\}_{q\in[k]}$ and $\{\tau_i^{(c)}\}_{i\in[n]}$;

while $\{\Gamma_q^{(c)}\}_{q\in[k]} \neq \{\Gamma_q^{(u)}\}_{q\in[k]}$ **do**

  Update $\{f_{q,l}^{(u)}\}_{q,l\in[k]}$ via (3.2) with $\{\Gamma_q^{(c)}\}_{q\in[k]}$ and $\{\tau_i^{(c)}\}_{i\in[n]}$;

  Update $\{\Gamma_q^{(u)}\}_{q\in[k]}$ via (3.4) with $\{f_{q,l}^{(u)}\}_{q,l\in[k]}$, $\{\Gamma_q^{(c)}\}_{q\in[k]}$ and $\{\tau_i^{(c)}\}_{i\in[n]}$;

  Update $\{\tau_i^{(u)}\}_{i\in[n]}$ via (3.7) with $\{\Gamma_q^{(u)}\}_{q\in[k]}$;

  Evaluate the stopping criterion ;

  $\{f_{q,l}^{(c)}\}_{q,l\in[k]} \leftarrow \{f_{q,l}^{(u)}\}_{q,l\in[k]}$;

  $\{\Gamma_q^{(c)}\}_{q\in[k]} \leftarrow \{\Gamma_q^{(u)}\}_{q\in[k]}$;

  $\{\tau_i^{(c)}\}_{i\in[n]} \leftarrow \{\tau_i^{(u)}\}_{i\in[n]}$;

end

Output: $\{f_{q,l}^{(c)}\}_{q,l\in[k]}$, $\{\Gamma_q^{(c)}\}_{q\in[k]}$, $\{\tau_i^{(c)}\}_{i\in[n]}$.

4. Theory

Assume $\mathbf{F} = (F_i)_{i=1}^n$ is from the parameter space

$$\mathcal{F}_k = XXX.$$

Theorem 4.1. *For any constant $C' > 0$, there is a constant $C > 0$ only depending on C', such that*

$$\frac{1}{n}\sum_{i=1}^{n} \left\| \hat{F}_i - F_i \right\|^2 \leq C\,(XXX)\,,$$

with probability at least $1 - \exp(-C'XXX)$, uniformly over $\mathbf{F} \in \mathcal{F}_k$.

Proof. This is a sketch of proof and is based on the proof in [2].

We denote the true value by $\theta_i^* = F^*_{z_i^*}(\cdot - \tau_i^*)$. For the estimated $\hat{z}$, define $\tilde{\theta} = \arg\min_{\theta \in \Theta_k(\hat{z})} \|\theta^* - \theta\|^2$. By the definition of the estimator, we have

$$L(\hat{F}, \hat{Z}, \hat{\tau}) \leq L(F^*, Z^*, \tau^*),$$

which can be rewritten as

$$\|\hat{\theta} - F^{obs}\|^2 \leq \|\theta^* - F^{obs}\|^2. \tag{4.1}$$

The left-hand side of (4.1) can be decomposed as

$$\|\hat{\theta} - \theta^*\|^2 + 2\langle \hat{\theta} - \theta^*, \theta^* - F^{obs} \rangle + \|\theta^* - F^{obs}\|^2. \tag{4.2}$$

Combining (4.1) and (4.2), we have

$$\|\hat{\theta} - \theta^*\|^2 \leq 2\langle \hat{\theta} - \theta^*, F^{obs} - \theta^* \rangle. \tag{4.3}$$

The right-hand side of (4.3) can be bounded as

$$\begin{aligned}
\langle \hat{\theta} - \theta^*, F^{obs} - \theta^* \rangle &= \langle \hat{\theta} - \tilde{\theta}, F^{obs} - \theta^* \rangle + \langle \tilde{\theta} - \theta^*, F^{obs} - \theta^* \rangle \\
&\leq \|\hat{\theta} - \tilde{\theta}\| \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{\|\hat{\theta} - \tilde{\theta}\|}, F^{obs} - \theta^* \right\rangle \right| && (4.4) \\
&\quad + \Big(\|\tilde{\theta} - \hat{\theta}\| + \|\hat{\theta} - \theta^*\| \Big) \left| \left\langle \frac{\tilde{\theta} - \theta^*}{\|\tilde{\theta} - \theta^*\|}, F^{obs} - \theta^* \right\rangle \right|. && (4.5)
\end{aligned}$$

Using Lemmas XXX, the following three terms:

$$\|\hat{\theta} - \tilde{\theta}\|, \quad \left| \left\langle \frac{\hat{\theta} - \tilde{\theta}}{\|\hat{\theta} - \tilde{\theta}\|}, F^{obs} - \theta^* \right\rangle \right|, \quad \left| \left\langle \frac{\tilde{\theta} - \theta^*}{\|\tilde{\theta} - \theta^*\|}, F^{obs} - \theta^* \right\rangle \right| \tag{4.6}$$

can all be bounded by XXX with probability at least

$$XXX.$$

Combining these bounds with (4.4), (4.5) and (4.3), we get

$$\|\hat{\theta} - \theta^*\|^2 \leq XXX$$

with probability at least XXX. □

Now we present the lemmas, which bound the three terms in (4.6), respectively.

Lemma 1. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on C', such that*

$$\|\hat{\theta} - \tilde{\theta}\| \leq CXXX,$$

with probability at least XXX.

Proof of Lemma 1.

Step 1: Control $\mathbb{E}\|\hat\tau - \tilde\tau\|^2$.

Fix a group $\hat\Gamma_k$, denote the group size by $J = |\Gamma_k|$. For notation simplicity, relabel the nodes in group $\hat\Gamma_k$ with $\{1, \cdots, J\}$. Let

$$M(\tau_1, \cdots, \tau_J) = \frac{1}{J}\sum_{j=1}^{J}\sum_{|k|\le k_0}\left|c_{j,k}e^{\mathrm{i}2\pi k\tau_j} - \frac{1}{J}\sum_{j'=1}^{J} c_{j',k}e^{\mathrm{i}2\pi k\tau_{j'}}\right|^2,$$

where $c_{j,k}, k \in \mathbb{Z}$, are the Fourier coefficients of the distribution function $F^*_{z^*_j}(\cdot - \tau^*_j)$. Let

$$\hat M(\tau_1, \cdots, \tau_J) = \frac{1}{J}\sum_{j=1}^{J}\sum_{|k|\le k_0}\left|\hat c_{j,k}e^{\mathrm{i}2\pi k\tau_j} - \frac{1}{J}\sum_{j'=1}^{J} \hat c_{j',k}e^{\mathrm{i}2\pi k\tau_{j'}}\right|^2,$$

where $\hat c_{j,k}, k \in \mathbb{Z}$, are the Fourier coefficients of the empirical distribution function F^{obs}_j. Then $\hat\tau$ is the minimizer of $\hat M$, and $\tilde\tau$ is the minimizer of M.
By Proposition 3.1 in Bigot and Gendre [1], under proper assumptions of F and distribution of τ,

$$\frac{1}{J}\|\hat\tau - \tilde\tau\|^2 \le C^{-1}\cdot(M(\hat\tau_1, \cdots, \hat\tau_J) - M(\tilde\tau_1, \cdots, \tilde\tau_J)).$$

Note that $M(\hat\tau) - M(\tilde\tau) = M(\hat\tau) - \hat M(\hat\tau) + \hat M(\hat\tau) - M(\tilde\tau) \le 2\ \sup_\tau |M(\tau) - \hat M(\tau)|$, so it suffices to control $\mathbb{E}\sup_\tau |M(\tau) - \hat M(\tau)|$.

$\mathbb{E}\sup_\tau |M(\tau) - \hat M(\tau)|$ is controlled by $F^{obs}_j - F^*_j$ or $\hat c_{j,k} - c_{j,k}$?

Step 2: Bound $\|\hat\theta - \tilde\theta\|$.

□

Lemma 2. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on C', such that*

$$\left|\left\langle \frac{\tilde\theta - \theta^*}{\|\tilde\theta - \theta^*\|}, F^{obs} - \theta^*\right\rangle\right| \le CXXX,$$

with probability at least XXX.

Proof. Note that

$$\tilde\theta_i - \theta^*_i = \tilde F_{\hat z_i}(\cdot - \tilde\tau_i) - F^*_{z^*_i}(\cdot - \tau^*_i)$$

is a function of the partition $\hat{\mathbf{z}}$, then we have

$$\left|\sum_i \left\langle \frac{\tilde\theta_i - \theta^*_i}{\sqrt{\sum_i \|\tilde\theta_i - \theta^*_i\|^2}}, F^{obs}_i - F^*_{z^*_i}(\cdot - \tau^*_i)\right\rangle\right| \le \max_{\mathbf{z}\in\mathcal{Z}_{n,k}}\left|\sum_i \left\langle \gamma_i(\mathbf{z}), F^{obs}_i - F^*_{z^*_i}(\cdot - \tau^*_i)\right\rangle\right|$$

where

$$\gamma_i(\mathbf{z}) \propto \tilde{F}_{z_i}(\cdot - \tilde{\tau}_i) - F^*_{z_i^*}(\cdot - \tau_i^*)$$

satisfies $\sum_i \|\gamma_i(\mathbf{z})\|^2 = 1$. By [some inequality similar to Hoeffding's inequality] and union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{\mathbf{z}\in\mathcal{Z}_{n,k}}\left|\sum_i \left\langle \gamma_i(\mathbf{z}), F_i^{obs} - F^*_{z_i^*}(\cdot - \tau_i^*)\right\rangle\right| > t\right)\\
\leq &\sum_{z\in\mathcal{Z}_{n,k}} \mathbb{P}\left(\left|\sum_i \left\langle \gamma_i(\mathbf{z}), F_i^{obs} - F^*_{z_i^*}(\cdot - \tau_i^*)\right\rangle\right| > t\right)\\
\leq &|\mathcal{Z}_{n,k}| \exp\left(-C_1 t^2\right),
\end{aligned}$$

for some universal constant $C_1 > 0$. Choosing $t \propto \sqrt{n \log k}$, the proof is complete.

□

Lemma 3. *For any constant $C' > 0$, there exists a constant $C > 0$ only depending on C', such that*

$$\left|\left\langle \frac{\hat{\theta} - \tilde{\theta}}{\|\hat{\theta} - \tilde{\theta}\|}, F^{obs} - \theta^*\right\rangle\right| \leq CXXX,$$

with probability at least XXX.

Proof. Note that

$$\hat{\theta}_i - \tilde{\theta}_i = \hat{F}_{\hat{z}_i}(\cdot - \hat{\tau}_i) - \tilde{F}_{\hat{z}_i}(\cdot - \tilde{\tau}_i)$$

is a function of both the partition $\hat{\mathbf{z}}$ and the observations F^{obs}. For each $\mathbf{z} \in \mathcal{Z}_{n,k}$, define the set $\mathcal{B}_{\mathbf{z}}$ by

$$XXX$$

Thus, we have the bound

$$\left|\sum_i \left\langle \frac{\tilde{\theta}_i - \theta_i^*}{\sqrt{\sum_i \|\tilde{\theta}_i - \theta_i^*\|^2}}, F_i^{obs} - F^*_{z_i^*}(\cdot - \tau_i^*)\right\rangle\right| \leq \max_{\mathbf{z}\in\mathcal{Z}_{n,k}} \sup_{c\in\mathcal{B}_{\mathbf{z}}} \left|\sum_i \left\langle c_i, F_i^{obs} - F^*_{z_i^*}(\cdot - \tau_i^*)\right\rangle\right|$$

If set $\mathcal{B}_{\mathbf{z}}$ is not too large, applying union bound (and Hoeffding-like inequality) completes the proof.

□

5. Simulation

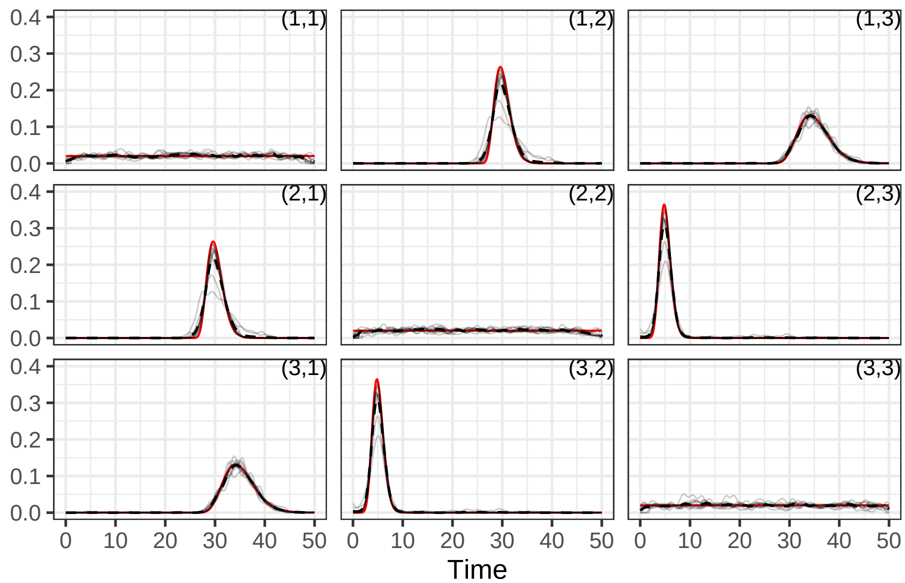

Fig 3: Estimated connecting patterns (ten trials). Dash line: average curves of estimated connecting patterns. Red line: true connecting patterns.

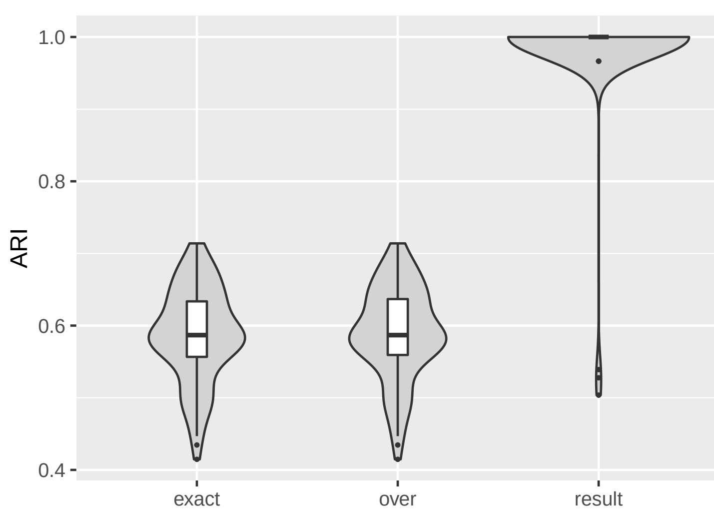

Fig 4: Clustering result.

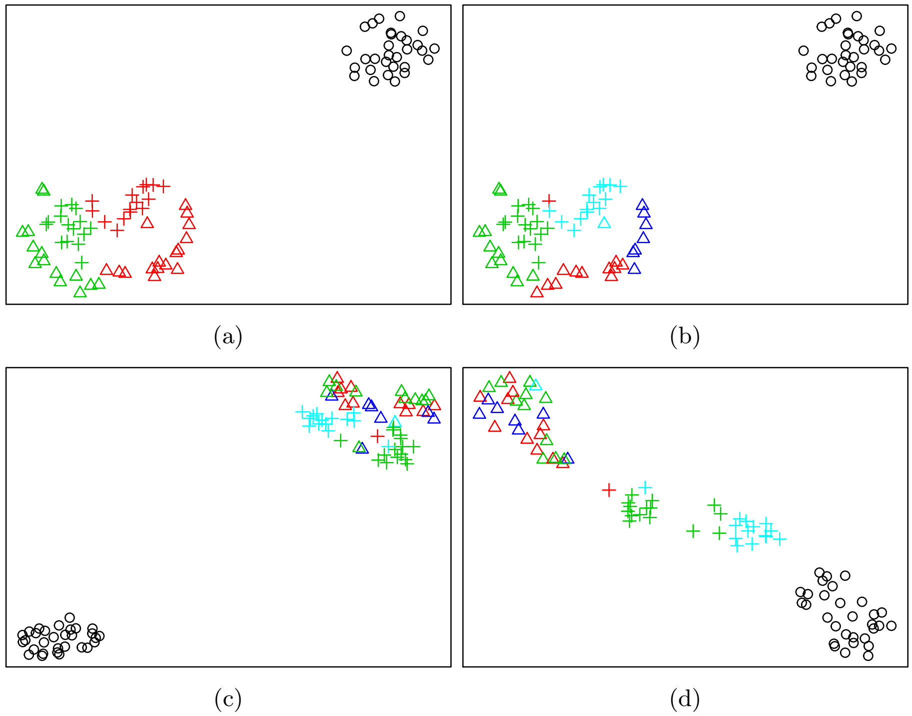

Fig 5: 2D representation by t-SNE. (a) Aggregated point processes with estimated three clusters. (b) Aggregated point processes with estimated five clusters. (c) Point processes incorporated clustering results from (a). (d) Point processes incorporated clustering results from (b).

6. Conclusion

Appendix A: Appendix section

A.1. Derivation of distance between pdfs and cdfs

Suppose a pdf f is defined on $[0, T]$. Extend f by zero to $[-T, 0]$, and define the shifted pdf $S_\tau \circ f(\cdot)$ by

$$S_\tau \circ f(t) = \begin{cases} 0, & t \in [-T, -\tau) \\ f(t+\tau), & t \in [-\tau, T-\tau) \\ 0, & t \in [T-\tau, T] \end{cases} . \tag{A.1}$$

Suppose the extended f (defined on $[-T,T]$) is represented by $X_{-(N-1)/2},\cdots,X_{(N-1)/2}$, the Fourier coefficients are

$$\theta_j = \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} X_k \cdot \exp\{-\mathrm{i}2\pi(k/N)j\}, \quad j = -\frac{(N-1)}{2},\cdots,\frac{(N-1)}{2}.$$

Inverse transformation

$$X_k = \frac{1}{N}\sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} \theta_j \cdot \exp\{\mathrm{i}2\pi(j/N)k\}, \quad k = -\frac{(N-1)}{2},\cdots,\frac{(N-1)}{2}.$$

The Fourier coefficients of $S_\tau \circ f$ is

$$\begin{aligned}
\theta'_j &= \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}-n_0} X_{k+n_0} \cdot \exp\{-\mathrm{i}2\pi(k/N)j\} + \sum_{k=\frac{(N-1)}{2}-n_0+1}^{\frac{N-1}{2}} 0 \\
&= \sum_{k=\frac{-(N-1)}{2}+n_0}^{\frac{N-1}{2}} X_k \cdot \exp\{-\mathrm{i}2\pi((k-n_0)/N)j\} + \sum_{k=\frac{-(N-1)}{2}}^{\frac{-(N-1)}{2}+n_0-1} 0 \\
&= \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} X_k \cdot \exp\{-\mathrm{i}2\pi((k-n_0)/N)j\} \\
&= \theta_j \cdot \exp\{\mathrm{i}2\pi(n_0/N)j\}.
\end{aligned}$$

Suppose another pdf g is also defined on $[0,T]$ and extended to $[-T,T]$ with Fourier coefficients γ_j, $j = -\frac{(N-1)}{2},\cdots,\frac{(N-1)}{2}$. The shift-invariant distance between f and g is then

$$
\begin{aligned}
d(f,g) = \min\Bigg\{ &\inf_{\tau\in[0,T]} \left(\int_{-T}^{T} \left|S_\tau \circ f(t) - g(t)\right|^2 \mathrm{d}t \right)^{1/2}, \\
&\inf_{\tau\in[0,T]} \left(\int_{-T}^{T} \left|f(t) - S_\tau \circ g(t)\right|^2 \mathrm{d}t \right)^{1/2} \Bigg\} \\
= \min &\left\{ \inf_{\tau\in[0,T]} \left(\frac{2T}{N} \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |X'_k - Y_k|^2 \right)^{1/2}, \inf_{\tau\in[0,T]} \left(\frac{2T}{N} \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |X_k - Y'_k|^2 \right)^{1/2} \right\} \\
= \min &\left\{ \min_{0\le n_0 \le (N-1)/2} \left(\frac{2T}{N} \frac{N}{N^2} \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta'_j - \gamma_j|^2 \right)^{1/2}, \min_{0\le n_0 \le (N-1)/2} \left(\frac{2T}{N} \frac{N}{N^2} \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta_j - \gamma'_j|^2 \right)^{1/2} \right\} \\
= \min \Bigg\{ &\min_{0\le n_0 \le (N-1)/2} \left(\frac{2T}{N^2} \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta_j \cdot \exp\{\mathrm{i}2\pi(n_0/N)j\} - \gamma_j|^2 \right)^{1/2}, \\
&\min_{0\le n_0 \le (N-1)/2} \left(\frac{2T}{N^2} \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta_j - \gamma_j \cdot \exp\{\mathrm{i}2\pi(n_0/N)j\}|^2 \right)^{1/2} \Bigg\}.
\end{aligned}
$$

Here $n_0 = N \cdot \tau / 2T$.

For loss function

$$
L(n_0) = \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta_j \cdot \exp\{\mathrm{i}2\pi(n_0/N)j\} - \gamma_j|^2,
$$

the gradient is

$$
\nabla_{n_0} = \frac{4\pi}{N} \cdot \sum_{|j|\le (N-1)/2} j \cdot \mathrm{Im}\left(\theta_j \overline{\gamma_j} e^{\mathrm{i}2\pi n_0 j/N} \right).
$$

For cdf F and G, define the shifted (towards left) cdf as

$$
S_\tau \circ F(t) = \begin{cases} 0, & t \in [-T, -\tau) \\ F(t+\tau), & t \in [-\tau, T-\tau) \\ 1, & t \in [T-\tau, T] \end{cases}. \tag{A.2}
$$

The Fourier coefficients of $S_\tau \circ F$ is

$$
\begin{aligned}
\theta'_j &= \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}-n_0} X_{k+n_0} \cdot \exp\{-\mathrm{i}2\pi(k/N)j\} + \sum_{k=\frac{(N-1)}{2}-n_0+1}^{\frac{N-1}{2}} 1 \cdot \exp\{-\mathrm{i}2\pi(k/N)j\} \\
&= \sum_{k=\frac{-(N-1)}{2}+n_0}^{\frac{N-1}{2}} X_k \cdot \exp\{-\mathrm{i}2\pi((k-n_0)/N)j\} + \sum_{k=\frac{-(N-1)}{2}}^{\frac{-(N-1)}{2}+n_0-1} 0 + \sum_{k=\frac{(N-1)}{2}-n_0+1}^{\frac{N-1}{2}} 1 \cdot \exp\{-\mathrm{i}2\pi(k/N)j\} \\
&= \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} X_k \cdot \exp\{-\mathrm{i}2\pi((k-n_0)/N)j\} + \sum_{k=\frac{(N-1)}{2}-n_0+1}^{\frac{N-1}{2}} 1 \cdot \exp\{-\mathrm{i}2\pi(k/N)j\} \\
&= \theta_j \cdot \exp\{\mathrm{i}2\pi(n_0/N)j\} + \frac{e^{\mathrm{i}2\pi j(n_0/N)}-1}{1-e^{\mathrm{i}2\pi j/N}} \cdot \mathbf{1}_{\{j\neq 0\}} + n_0 \cdot \mathbf{1}_{\{j=0\}}.
\end{aligned}
$$

The shift-invariant distance between F and G is then

$$
\begin{aligned}
d(F,G) &= \min\Big\{ \inf_{\tau\in[0,T]} \left(\int_{-T}^{T} \left| S_\tau \circ F(t) - G(t) \right|^2 \mathrm{d}t \right)^{1/2}, \\
&\qquad \inf_{\tau\in[0,T]} \left(\int_{-T}^{T} \left| F(t) - S_\tau \circ G(t) \right|^2 \mathrm{d}t \right)^{1/2} \Big\} \\
&= \min\left\{ \inf_{\tau\in[0,T]} \left(\frac{2T}{N} \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |X'_k - Y_k|^2 \right)^{1/2}, \inf_{\tau\in[0,T]} \left(\frac{2T}{N} \sum_{k=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |X_k - Y'_k|^2 \right)^{1/2} \right\} \\
&= \min\left\{ \min_{0\le n_0\le (N-1)/2} \left(\frac{2T}{N}\frac{N}{N^2} \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta'_j - \gamma_j|^2 \right)^{1/2}, \min_{0\le n_0\le (N-1)/2} \left(\frac{2T}{N}\frac{N}{N^2} \sum_{j=\frac{-(N-1)}{2}}^{\frac{N-1}{2}} |\theta_j - \gamma'_j|^2 \right)^{1/2} \right. \\
&= \min\Big\{ \min_{0\le n_0\le (N-1)/2} \Big(\frac{2T}{N^2} \Big[\sum_{0<|j|\le (N-1)/2} \left| \theta_{i,j} e^{\mathrm{i}2\pi jn/N} + \frac{e^{\mathrm{i}2\pi jn/N}-1}{1-e^{\mathrm{i}2\pi j/N}} - \gamma_{l,j} \right|^2 + |\theta_0 + n - \gamma_0|^2 \Big] \Big)^{1/2}, \\
&\qquad \min_{0\le n_0\le (N-1)/2} \Big(\frac{2T}{N^2} [\cdots] \Big)^{1/2} \Big\}.
\end{aligned}
$$

Loss function

$$
\begin{aligned}
L(n_0) &= \sum_{0<|j|\leq(N-1)/2} \left| \theta_j e^{\mathrm{i}2\pi j(n_0/N)} + \frac{e^{\mathrm{i}2\pi j(n_0/N)}-1}{1-e^{\mathrm{i}2\pi j/N}} - \gamma_j \right|^2 + |\theta_0 + n_0 - \gamma_0|^2 \\
&= \sum_{0<|j|\leq(N-1)/2} \left| \left(\theta_j + \frac{1}{1-e^{\mathrm{i}2\pi j/N}}\right) e^{\mathrm{i}2\pi j(n_0/N)} - \left(\gamma_j + \frac{1}{1-e^{\mathrm{i}2\pi j/N}}\right) \right|^2 + \\
&\qquad |\theta_0 + n_0 - \gamma_0|^2 \\
&\triangleq \sum_{0<|j|\leq(N-1)/2} \left| \theta_j'' e^{\mathrm{i}2\pi j(n_0/N)} - \gamma_j'' \right|^2 + |\theta_0 + n_0 - \gamma_0|^2 .
\end{aligned}
$$

Gradient:

$$
\nabla_{n_0} = \frac{4\pi}{N} \cdot \sum_{0<|j|\leq(N-1)/2} j \cdot \mathrm{Im}\left(\theta_j'' \overline{\gamma_j''} e^{\mathrm{i}2\pi n_0 j/N}\right) + 2n_0 + 2\mathrm{Re}\,(\theta_0 - \gamma_0) . \quad \text{(A.3)}
$$

TO DO:

1. Complete Method section. More visualization. [by 4/21]
2. Reading: basic knowledge about point process, and shift operation. [by 4/21]
3. Literature review: SBM, best permutation algorithm, SIM, etc. [by 4/21]
4. Proposal. [draft, by 04/28]
5. Slides. [draft, by 04/28]
6. Theory. [May, if time permits]
7. Apply algorithm of Matias' paper.
8. Real data.

Acknowledgements

See Supplement A for the supplementary material example.

Supplementary Material

Supplement A: Title of the Supplement A
(http://www.e-publications.org/ims/support/dowload/imsart-ims.zip). Dum esset rex in accubitu suo, nardus mea dedit odorem suavitatis. Quoniam confortavit seras portarum tuarum, benedixit filiis tuis in te. Qui posuit fines tuos

References

[1] Bigot, J. and Gendre, X. (2013). Minimax properties of Fréchet means of discretely sampled curves. *Ann. Stat.* **41** 923–956.
[2] Gao, C., Lu, Y. and Zhou, H. H. (2015). Rate-optimal graphon estimation. *Ann. Stat.* **43** 2624–2652.

[3] Holland, P. W., Blackmond, K. and Leinhardt, S. (1983). STOCHASTIC BLOCKMODELS: FIRST STEPS * Educational Testing Seroice ** Technical Report.

[4] Matias, C. and Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **79** 1119–1141.

[5] Matias, C., Rebafka, T. and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika* **105** 665–680.

[6] Pensky, M. (2019). Dynamic network models and graphon estimation. *Ann. Stat.* **47** 2378–2403.

[7] Wan, Y., Wei, Z., Looger, L. L., Koyama, M., Druckmann, S., Keller Correspondence, P. J. and Keller, P. J. (2019). Single-Cell Reconstruction of Emerging Population Activity in an Entire Developing Circuit. *Cell* **179**.

[8] Xu, K. S. (2015). Stochastic block transition models for dynamic networks. *J. Mach. Learn. Res.* **38** 1079–1087.

[9] Xu, K. S. and Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.* **8** 552–562.

[10] Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks - A Bayesian approach. *Mach. Learn.* **82** 157–189.