Title

Simultaneous Clustering and Estimation for Recurrent Event Data with Time Shifts

Author

Zhang, Zitong

Publication Date

2024

Peer reviewed|Thesis/dissertation

Simultaneous Clustering and Estimation for Recurrent Event Data with Time Shifts

By

ZITONG ZHANG

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Statistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Shizhe Chen, Chair

Thomas C.M. Lee

Can M. Le

Committee in Charge

2024

Abstract

Recurrent event data, characterized by repeated occurrences of events associated with one or more subjects, is prevalent in various fields, including neuroscience, healthcare, and social science. Recent technological advancements in neuroscience have significantly increased the availability of such data, presenting both new opportunities and challenges. This dissertation focuses on the statistical analysis of two common types of recurrent event data in neuroscience: neural firing activity and functional connections between neurons. For each type of data, we discuss the unique challenges posed by the data and propose efficient statistical methods to address the challenges. The proposed methods aim to identify groups of neurons with similar activity patterns while accommodating temporal disparities among neurons. We establish conditions for the identifiability of model parameters. We conduct extensive numerical experiments to evaluate the empirical performance of the proposed methods. Additionally, we apply the proposed methods to real-world neural data to reveal distinct roles of neurons and identify representative neural activity patterns.

To my parents and grandparents.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof. Shizhe Chen, for his invaluable guidance, unwavering support, and insightful feedback throughout my doctoral study. His expertise in statistics has been instrumental in shaping my work. Additionally, I have benefited significantly from his mentorship in presenting research both orally and verbally. His unwavering commitment to excellence has been a source of inspiration throughout my PhD journey. I would also like to acknowledge Prof. Jie Peng for her guidance during the first two years of my study and for her assistance in preparing for my qualifying oral exam.

I am thankful to the members of my dissertation committee, Prof. Hao Chen, Prof. Thomas Lee, Prof. Can Le, Prof. Jane-Ling Wang for their assistance, encouragement and valuable comments. I extend my appreciation to the staff in the Department of Statistics for their prompt responses to my inquiries and guidance on many processes.

I am grateful for the support and understanding of my friends, whose encouragement and patience have been a constant source of strength during this challenging endeavor. In particular, I would like to thank my peers, Jinhua Cao, Emily Chang, Rui Hu, Yiran Li, Lingyou Pang, Tesi Xiao, Junwen Yao, Yuxuan Zhang, for making my PhD journey less intimidating through their companionship. I extend my gratitude to Jue Wang for our numerous sweet conversations. Additionally, special thanks are owed to Yidi Zhang for providing me with tremendous emotional support.

Lastly, I would like to express my heartfelt gratitude to my family for their unconditional love and support. I am truly grateful for their encouragement and belief in me, which have played a crucial role in my achievements. Their support has not only shaped my academic endeavors but also enriched my personal growth, and for that, I am forever thankful.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Recurrent event data is prevalent across various fields, such as neuroscience, healthcare, and social science. This type of data contains repeated occurrences of events associated with one or more subjects, where the event's timing contains specific information about the subjects. In neuroscience, the recurrent events may represent a series of firings of a neuron [see, e.g. Perkel et al., 1967]. Analyzing neural firings can enhance our understanding of how neurons encode information and perform computations [Knill and Pouget, 2004, Rolls and Treves, 2011, Steinmetz et al., 2019]. In healthcare, the recurrent events may represent the repeated hospitalizations of a patient. Studying patterns of hospitalizations can help identify high-risk patients that may benefit from targeted interventions [Mudge et al., 2011, Rogers et al., 2014, Dini et al., 2020]. In social science, the recurrent events may describe the recurring interactions such as tweets, phone calls, and text messages among individuals. Analyzing these interactions can aid in identifying anomalies in social behavior [McIntire et al., 2010, Chae et al., 2012, Mazza et al., 2019].

For a concrete discussion of the recurrent event data, we employ the framework of point processes [Cox and Isham, 1980, Daley and Vere-Jones, 2003, Karr, 2017]. Consider a set of *recurrent events* from n subjects $\mathcal{O} \equiv \left\{ \{t_{i,j}\}_{j=1,\ldots,N_i(T)} : 0 < t_{i,1} < \cdots < t_{i,N_{i,r}(T)} < T, i = 1, \ldots, n \right\}$, where $t_{i,j}$ denotes the occurrence time of the jth event associated with the ith subject, T denotes the duration of the observation, $N_i(T)$ denotes the total number of events associated with the ith subject. For $i \in \{1, \ldots, n\}$, we model the recurrent events by adopting the definition of *counting process* [see, e.g. Daley and Vere-Jones, 2003] as $N_i(t) \equiv \sum_{j=1}^{N_i(T)} \mathbf{1}(t_{i,j} \leq t)$ for $t \in [0, T]$. The counting process $N_i(t)$ can be characterized by the *intensity* function, which is defined as $\lambda_i(t) \equiv \mathbb{E}\{\mathrm{d}N_i(t)/\mathrm{d}t\}$. As a concrete example, in the context of neural firing data, $t_{i,j}$ represents the time of the jth firing of the ith neuron, $N_i(t)$ represents the number of firings of the ith neuron up to time t, and $\lambda_i(t)$ represents the firing rate of the ith neuron at time t.

In this dissertation, we focus on neuroscience data arising from recent technological advancements. These advancements have opened up new opportunities and raised interesting questions.

For instance, Neuropixels [Jun et al., 2017, Steinmetz et al., 2018, 2021] allow researchers to record the precise timing of recurrent firings of a large volume of neurons from specific brain regions or circuits. This technology enables the exploration of how information is encoded and processed in the brain [Steinmetz et al., 2019]. Additionally, calcium imaging [Grienberger and Konnerth, 2012, Dana et al., 2019, Zong et al., 2022] enables the tracking of individual neurons' activities over extended periods, offering opportunities to explore how neurons synchronize their activities to establish functional circuits [Wan et al., 2019].

However, alongside the technological advancements, the increasing complexity and volume of the neural data pose challenges in data analysis. One key issue is the heterogeneity of neurons across subjects, where the identity of individual neurons cannot be directly mapped between subjects [see, e.g., Zeng and Sanes, 2017]. This lack of a one-to-one mapping complicates the task of identifying common firing patterns or functions of neurons across subjects [Connors and Gutnick, 1990, Markram et al., 2004, Komendantov et al., 2019, Steinmetz et al., 2019]. Moreover, the high volume of data makes it impractical to analyze each neuron individually. In light of these challenges, we aim to develop statistical methods to identify groups of neurons that exhibit similar activity patterns.

Various clustering methods can be applied on the recurrent event data. For instance, when recurrent events are associated with individual subjects, such as neural firings, *functional clustering* methods [see, e.g., Jacques and Preda, 2014a, Wang et al., 2016] can be employed through the modeling of intensity functions. In the realm of functional clustering, researchers have extended the k-means clustering method to functional data via functional basis expansion [Abraham et al., 2003, Garcia-Escudero and Gordaliza, 2005, Kayano et al., 2010]. Additionally, the k-means clustering and mixture models have been integrated with the functional principal component analysis [Chiou and Li, 2007, 2008, Bouveyron and Jacques, 2011, Jacques and Preda, 2013, 2014b, Yin et al., 2021], where clusters are allowed to have different mean and covariance functions. When recurrent events are associated with pairs of subjects, such as synchronization of activity between neurons, the *dynamic stochastic block models* can be applied [see, e.g., Xing et al., 2010, Yang et al., 2011, Xu and Hero, 2014, Xu, 2015, Matias and Miele, 2017, Matias et al., 2018, Kim et al., 2018, Lee and Wilkinson, 2019].

However, the aforementioned clustering methods are inadequate for the neural data analysis due to the presence of unobserved time shifts in activities of individual neurons. These time shifts are well known, and might result from various factors, such as different neural response latency [Oram et al., 2002, Levakova et al., 2015, Lee et al., 2020] or varying functional maturation progress [Overstreet-Wadiche and Westbrook, 2006, Nicholas et al., 2013, Wan et al., 2019]. Failing to account for these time shifts may lead to misleading results and limited insights into neural activity patterns.

To address this challenge, we propose to integrate *curve alignment* methods with clustering methods. Curve alignment, or curve registration, involves finding an optimal transformation (such as translation or scaling) to align a collection of curves as closely as possible. This task has been explored in numerous studies [Kneip and Gasser, 1992, Ramsay and Li, 1998, James, 2007, Telesca and Inoue, 2008, Cheng et al., 2016]. In cases where curves differ only by unknown time shifts, the task of estimating these time shifts has been studied in the context of *shape invariant models* [Beath, 2007, Bigot and Gadat, 2010, Vimond, 2010b, Bigot et al., 2013, Bigot and Gendre, 2013, Bontemps and Gadat, 2014]. These models aim to capture the common shape of a collection of curves while accommodating their unique time shifts.

In this dissertation, we develop statistics methods to analyze recurrent event data in neuroscience. In particular, we focus on analyzing the roles of neurons based on neural activities. In Chapter 2, we propose an additive shape invariant mixture model capable of simultaneously decompose, align, and cluster the neural spike trains. In Chapter 3, we propose a dynamic stochastic block model that allows for unknown neuron-specific time shifts for analyzing neural circuit development. The theoretical proofs and supplementary results for Chapter 2 and 3 are presented in Appendix A and B, respectively.

CHAPTER 2

Simultaneous Clustering and Estimation of Additive Shape Invariant Models for Recurrent Event Data

2.1. Introduction

Recent technological advancements have greatly improved our ability to record neural activities with high spatiotemporal resolution and over large volumes. Recording device such as Neuropixels [Jun et al., 2017, Steinmetz et al., 2018, 2021] enable scientists to record neurons across multiple brain areas while subjects engage in behavioral tasks. The growing amount of neural data facilitates the exploration of how neural firing patterns encode information, but also presents challenges to existing statistical methods. In this paper, we consider the problem of identifying subgroups of neurons that share similar responses to stimuli. In a typical experiment, a subject will be exposed to a series of stimuli (e.g., visual cue, auditory cue, reward) over the course of a trial, which means that the recorded spike train is a superposition of neural responses to these stimuli [Benucci et al., 2009, Capilla et al., 2011, Orhan and Ma, 2015]. It is challenging to disentangle the response to each stimulus given the stochastic nature of neural activities. Moreover, even if two neurons share the same responses to a stimulus, they might respond in different time due to the difference in their response latency [Oram et al., 2002, Levakova et al., 2015, Lee et al., 2020], making the estimation of shared neural responses more challenging. Lastly, even within the same area of the brain, we cannot assume that neurons share the same responses [Molyneaux et al., 2007, Lake et al., 2016, Cembrowski and Spruston, 2019]. In light of these intricacies, the primary objective of this study is to develop statistical methods for recurrent events to simultaneously address three key tasks: (i) *decompose* the neural firing activities into their constituent components, (ii) *align* neural firing patterns across neurons, and (iii) *cluster* neurons based on their similar firing patterns. As a concrete example, we consider the representative experiment by Steinmetz et al. [2019], where mice were exposed to a series of visual and auditory stimuli, and thousands of neurons were recorded

throughout the experiment. We display four neurons selected by our method in Figure 2.1 with distinct firing patterns to the stimuli.

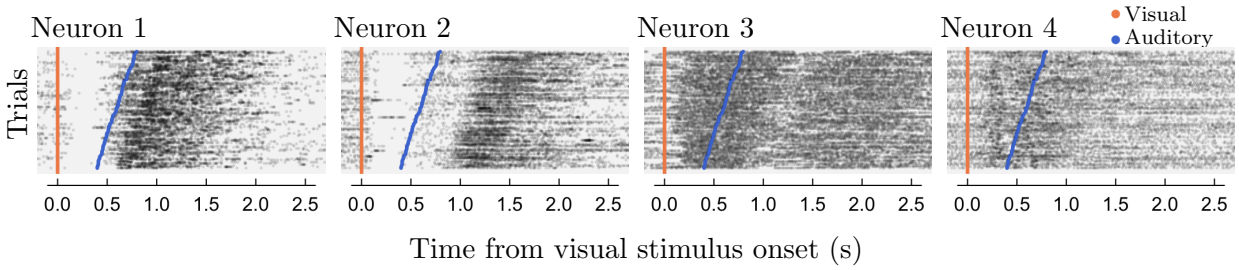

FIGURE 2.1. Activities of four example neurons from Steinmetz et al. [2019]. The four neurons are all from the midbrain region of the mouse brain, where their firings are shown as black dots. Each panel corresponds to a single neuron, where the x-axis represents time since visual stimulus onset, and the y-axis represents experiment trials. Trials are aligned by the visual stimulus onset time shown as orange dots , and ordered by the auditory stimulus onset time shown as blue dots.

However, existing methods are inadequate in addressing these challenges at hand. Shape invariant models [Beath, 2007, Bigot and Gadat, 2010, Vimond, 2010b, Bigot et al., 2013, Bigot and Gendre, 2013, Bontemps and Gadat, 2014] have been studied extensively to align a collection of curves with identical means up to unknown time shifts. The combination of shape invariant models and mixture models has led to the development of techniques for simultaneously aligning and clustering curves [Chudova et al., 2003, Gaffney and Smyth, 2004, Liu and Yang, 2009, Lu and Lou, 2019]. In a related vein, Sangalli et al. [2010] propose an approach to simultaneously align and cluster curves that share a common shape but are subject to non-linear distortions in temporal alignment. Nonetheless, all of these methods are limited in their capacity to decompose curves comprising multiple components.

A class of problems related to the decomposing superimposed curves has been studied in the context of functional principal component analysis (FPCA) [Yao et al., 2005, Morris and Carroll, 2006, Di et al., 2009, Crainiceanu et al., 2009]. FPCA aims to deconstruct functional data into orthogonal components that effectively account for the variation within the dataset. Researchers have applied the FPCA to analyze point processes through the modeling of intensity functions [Wu et al., 2013, Xu et al., 2020]. Furthermore, the FPCA models have been combined with mixture models to simultaneously cluster and decompose curves [Chiou and Li, 2007, Bouveyron and Jacques, 2011, Jacques and Preda, 2013, Yin et al., 2021]. However, the orthogonal components derived from

these methods may not match the components of our primary interest. Indeed, the congruence between the outcomes of these methods and our proposed approach is limited to specific scenarios (see Section 2.2.3). Besides, unknown time shifts have not been studied in FPCA.

In this paper, we propose an additive shape invariant mixture model (ASIMM) capable of simultaneously decompose, align, and cluster recurrent event data, such as neural firings. In Section 2.2, we introduce an additive shape invariant model for simultaneous decomposition and alignment. Section 2.3 expands upon the additive shape invariant model by integrating simultaneous clustering. In Section 2.4, we present a computationally efficient algorithm. The performance of ASIMM is assessed through simulation experiments in Section 2.5. In Section 2.6, we use ASIMM to study neural firing patterns in mice. Finally, we discuss potential future works in Section 2.7.

2.2. Additive Shape Invariant Model

2.2.1. Model. Consider a set of *recurrent events* from repeated measurements of n subjects $\mathcal{O} \equiv \left\{ \{t_{i,r,j}\}_{j=1,\dots,N_{i,r}(T)} : 0 < t_{i,r,1} < \cdots < t_{i,r,N_{i,r}(T)} < T, i = 1,\dots,n, r = 1,\dots,R \right\}$, where $t_{i,r,j}$ denotes the time of the j-th event of the i-th subject in the r-th observation, T denotes the duration of each observation, $N_{i,r}(T)$ denotes the total number of events associated with the i-th subject and the r-th observation. For $i \in [n]$ and $r \in [R]$, we adopt the definition of *counting process* [see, e.g. Daley and Vere-Jones, 2003] as $N_{i,r}(t) \equiv \sum_{j=1}^{N_{i,r}(T)} \mathbf{1}(t_{i,r,j} \le t)$ for $t \in [0,T]$. We use the *intensity* function, i.e., $\lambda_{i,r}(t) \equiv \mathbb{E}\{\mathrm{d}N_{i,r}(t)/\mathrm{d}t\}$, to characterize each counting process.

We start with a model that tackles the first two of the aforementioned tasks: decomposition and alignment. To be specific, we assume that the true underlying intensities are formed through the superposition of multiple components, and these intensity components exhibit uniformity across subjects, differing only in temporal shifts. We arrive at the following model

$$\lambda_{i,r}(t) = a + \sum_{m \in [M]} S^{v_{i,m} + w^*_{r,m}} f_m(t), \tag{2.1}$$

where $a \in [0,\infty)$ represents the *baseline intensity*, $M \in \mathbb{N}^+$ is the number of components, $v_{i,m} \in [0,V]$ represents the *subject-specific time shift* associated with the i-th subject and the m-th component, $w^*_{r,m} \in [0,W]$ represents the known *observation-specific time shift* of the m-th component in the r-th observation, S^v is the shift operator defined as $S^v x(t) \equiv x(t-v)$, and $f_m(t) \in \mathcal{F}$ represents the m-th *intensity component*. Here $V, W \in (0,T)$, and $\mathcal{F} \equiv \{f \in L^2(\mathbb{R}) : f(t) = 0 \text{ for } t \in$

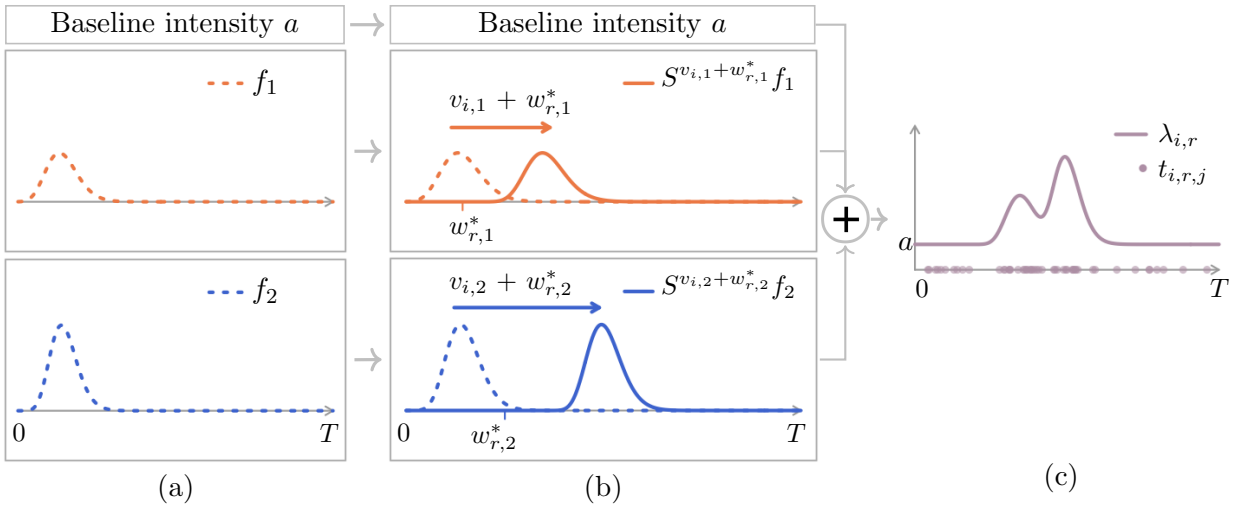

FIGURE 2.2. Graphical representation of the additive shape invariant model with two components. Panel (a) shows the intensity components shared across subjects and observations. Panel (b) shows the intensity components associated with subject i and observation r where the intensity is shifted by $v_{i,m} + w^*_{r,m}$. Panel (c) shows the expected intensity of subject i in observation r, and one realization of the point process $\{t_{i,r,j} : j = 1, \cdots, N_{i,r}(T)\}$.

$\mathbb{R} \setminus (0, T_0)\}$ with $T_0 \in (0, T)$. For further clarity, a graphical representation of the model in (2.1) is provided in Figure 2.2.

To provide a concrete example for Model 2.1, consider the experiment in Steinmetz et al. [2019] where $N_{i,r}(t)$ represents the recorded spike train of the i-th neuron in the r-th trial, and $\lambda_{i,r}(t)$ denotes the corresponding firing rate (see for instance Figure 2.1). The terms a and $\{f_m(\cdot) : m = 1, 2\}$ can be interpreted as the spontaneous firing rate and the neural response elicited by the visual gratings and auditory cue, respectively. The subject-specific time shift $v_{i,m}$ corresponds to the response latency of the i-th neuron in response to the m-th stimulus. The observation-specific time shift $w^*_{r,m}$ corresponds to the occurrence time of the m-th stimulus in the r-th trial, which is randomly generated by the experimenters.

2.2.2. Identifiability. We denote the collection of model parameters as $\boldsymbol{\theta}_0 \equiv (a, \mathbf{f}, \mathbf{v}) \in \Theta_0$, where $\mathbf{f} \equiv (f_m)_{m\in[M]}$, $\mathbf{v} \equiv (v_{i,m})_{i\in[n],m\in[M]}$, and $\Theta_0 \equiv \{(a, \mathbf{f}, \mathbf{v}) : a \in [0, \infty), \mathbf{f} \in \mathcal{F}^M, \mathbf{v} \in [0, V]^{n\times M}\}$. In addition, for any $N_{i,r}(t)$, we denote its conditional intensity given the observation-specific time shifts as $\lambda_{\boldsymbol{\theta}_0,i}(t, \mathbf{w}^*) \equiv \mathbb{E}_{\boldsymbol{\theta}_0}\{\mathrm{d}N_{i,r}(t)/\mathrm{d}t \mid \mathbf{w}^*_r = \mathbf{w}^*\}$, where $\mathbf{w}^*_r \equiv (w^*_{r,m})_{m\in[M]}$. We further denote $\boldsymbol{\lambda}_{\boldsymbol{\theta}_0}(t, \mathbf{w}^*) \equiv (\lambda_{\boldsymbol{\theta}_0,i}(t, \mathbf{w}^*))_{i\in[n]}$. In this context, *identifiability* pertains to the *injectivity* of the mapping $\boldsymbol{\theta}_0 \mapsto \boldsymbol{\lambda}_{\boldsymbol{\theta}_0}(\cdot, \cdot)$ for $\boldsymbol{\theta}_0 \in \Theta_0$.

With these notations, we can formally present the identifiability results in Proposition 2.2.1.

PROPOSITION 2.2.1. *Suppose that the following assumptions hold.*

(C2.1) $T \geq T_0 + V + W$.

(C2.2) $\mathbb{E}[\overline{\boldsymbol{\eta}^*(\xi)}\boldsymbol{\eta}^*(\xi)^\top]$ *is invertible for* $\xi \in \mathbb{R}\backslash\{0\}$, *where* $\boldsymbol{\eta}^*(\xi) \equiv (\eta_m)_{m\in[M]}$, $\eta_m \equiv \exp\{-\mathrm{j}\, 2\pi\xi w_m^*\}$, j *denotes the imaginary unit, and* $\overline{\boldsymbol{\eta}^*(\xi)}$ *denotes the complex conjugate of* $\boldsymbol{\eta}^*(\xi)$.

Let $(a^*, \mathbf{f}^*, \mathbf{v}^*) \in \Theta_0$ *denote the true parameters. Then, we can verify the following statements hold.*

(P2.1) For $m \in [M]$, *the intensity component* f_m^* *is identifiable up to a time shift.*

(P2.2) For $m \in [M]$ *such that* f_m^* *is non-zero on a set of positive measure, the subject-specific time shifts* $\{v_{i,m}^* : i \in [n]\}$ *are identifiable up to an additive constant.*

(P2.3) The baseline intensity a^* *is identifiable.*

Assumption C2.1 posits that the observation duration sufficiently extends to avoid censorship. Given C2.1, the model described in (2.1) can be formulated in the form of a linear regression model in the frequency domain, where the response variables are determined by $N_{i,r}$'s, the explanatory variables depend on $w_{r,m}^*$'s, and the regression coefficients are functions of f_m's and $v_{i,m}$'s. Within the context of linear regression, the significance of Assumption C2.2 becomes straightforward. Specifically, Assumption C2.2 effectively assumes no collinearity among explanatory variables. We note that Assumption C2.2 is satisfied if either (i) the observation-specific time shifts are randomized, or (ii) the gaps between the observation-specific time shifts are randomized, i.e., $w_m^* = w_{m-1}^* + \delta_{m-1}$ for $m = 2, \ldots, M$, where $\{\delta_m \in \mathbb{C} : m \in [M-1]\}$ are independent random variables with non-zero variance. Detailed proof of Proposition 2.2.1 is presented in Section A.1 of Appendix A.

2.2.3. Connections with existing models. In special scenarios, the model presented in (2.1) has connections with three distinct branches of research. Firstly, considering a special case where $M = 1$ and $R = 1$, the model in (2.1) can be simplified as follows:

$$\lambda_{i,1}(t) = a + S^{v_{i,1}+w_{1,1}^*} f_1(t) \equiv S^{v_{i,1}} g(t), \tag{2.2}$$

where $g(t) \equiv a + S^{w_{1,1}^*} f_1(t)$. This model has been investigated in the realm of *shape invariant models* [see, e.g., Beath, 2007, Bigot and Gadat, 2010, Vimond, 2010b, Bigot et al., 2013, Bigot and Gendre, 2013, Bontemps and Gadat, 2014]. Consequently, the proposed model can be seen as a generalization of shape invariant models to incorporate additive components. Various assumptions

have been proposed to ensure the identifiability of shape invariant models, including availability of repeated observations [Bigot and Gendre, 2013], partial knowledge of time shifts [Bigot and Gadat, 2010, Bigot et al., 2013], or the knowledge of the noise distribution [Bontemps and Gadat, 2014]. For the proposed additive shape invariant model, where M is allowed to exceed 2, a combination of repeated observations and partial knowledge of time shifts suffices to ensure the model identifiability, as elucidated in Proposition 2.2.1.

Secondly, considering the scenario where the intensity components $\{S^{v_{i,m}+w^*_{r,m}} f_m(t) : m \in [M]\}$ have non-overlapping supports for all $i \in [n]$ and $r \in [R]$, model (2.1) can be expressed as:

$$\lambda_{i,r}(t) = a + \sum_{m\in[M]} S^{v_{i,m}+w^*_{r,m}} f_m(t) \equiv g\{h_{i,r}(t)\}, \tag{2.3}$$

where $g(s) \equiv a + \sum_{m\in[M]} f_m(s)$ and $h_{i,r}(t) \equiv \sum_{m\in[M]} \{t - (v_{i,m} + w^*_{r,m})\} \times \mathbf{1}[f_m\{t - (v_{i,m} + w^*_{r,m})\} \neq 0]$. The reformulated model in (2.3) means that $\lambda_{i,r}$'s are identical subject to unknown *time warping functions* (i.e., $h_{i,r}(t)$'s). The task of estimating time warping functions has been explored in the domain of *curve registration* [Kneip and Gasser, 1992, Ramsay and Li, 1998, James, 2007, Telesca and Inoue, 2008, Cheng et al., 2016].

Lastly, considering the scenario where the variances of $v_{i,m}$'s and $w^*_{r,m}$'s are both close to zero, model (2.1) can be approximated using Taylor expansion as:

$$\lambda_{i,r}(t) \approx \mu(t) + \sum_{m\in[M]} \zeta_{i,r,m}\ \psi_m(t) \tag{2.4}$$

where $\mu(t) \equiv a + \sum_{m\in[M]} f_m(t - \mathbb{E}u_{i,r,m})$, $u_{i,r,m} \equiv v_{i,m} + w^*_{r,m}$, $\zeta_{i,r,m} \equiv -(u_{i,r,m} - \mathbb{E}u_{i,r,m})\|Df_m(t - \mathbb{E}u_{i,r,m})\|_t$, Df_m denotes the first order derivative of f_m, and $\psi_m(t) \equiv Df_m(t - \mathbb{E}u_{i,r,m})\|Df_m(t - \mathbb{E}u_{i,r,m})\|_t^{-1}$. We defer the derivation of this approximation to Section A.3 of Appendix A. When functions $\{S^{\mathbb{E}u_{i,r,m}} f_m(t) : m \in [M]\}$ exhibit non-overlapping supports, the approximate model in (2.4) corresponds to the models of *functional principal component analysis* (FPCA) [Yao et al., 2005, Morris and Carroll, 2006, Di et al., 2009, Crainiceanu et al., 2009, Xu et al., 2020].

2.2.4. Estimation. We consider the case that $N_{i,r}(t)$'s are Poisson processes. To effectively estimate the parameters in (2.1), we exploit the two sources of stochasticity of the Poisson processes. Firstly, the number of events over $[0, T]$ for any Poisson process is a random variable that follows a Poisson distribution, in other words, $N_{i,r}(T) \sim \text{Poisson}(\Lambda_{i,r})$, where $\Lambda_{i,r} \equiv \int_0^T \lambda_{i,r}(t)\mathrm{d}t$ represents

the *expected event count*. When Assumption C2.1 holds, we can derive from the model in (2.1) that $\Lambda_{i,r} = aT + \sum_{m\in[M]} \int_0^T f_m(t)\mathrm{d}t$, denoted as Λ for simplicity. Secondly, conditioning $N_{i,r}(T)$, the event times $\{t_{i,r,j} : j = 1, \ldots, N_{i,r}(T)\}$ can be regarded as independent and identically distributed random variables. The probability density function of these event times, or *event time distribution*, is characterized by $\lambda_{i,r}(t)\Lambda^{-1} = a\Lambda^{-1} + \sum_{m\in[M]} S^{v_{i,m}+w^*_{r,m}} f_m(t)\Lambda^{-1}$.

Accordingly, we estimate $a, \mathbf{f}, \mathbf{v}$ through the following reparameterization. Letting $a' \equiv a\Lambda^{-1}$and $\mathbf{f}' \equiv (f_m\Lambda^{-1})_{m\in[M]}$, we have the following optimization problem:

$$\begin{aligned}\hat{a}', \hat{\mathbf{f}}', \hat{\mathbf{v}} &\equiv \underset{a',\mathbf{f}',\mathbf{v}}{\arg\min}\ L_1(a', \mathbf{f}', \mathbf{v}) \\ &\equiv \underset{a',\mathbf{f}',\mathbf{v}}{\arg\min} \sum_{i\in[n],r\in[R]} \beta_{i,r}\ \frac{1}{T}\left\| \frac{y_{i,r}(t)}{N_{i,r}(T)} - \left\{a' + \sum_{m\in[M]} S^{v_{i,m}+w^*_{r,m}} f'_m(t)\right\}\right\|_t^2,\end{aligned} \tag{2.5}$$

where $\|\cdot\|_t$ denotes the L^2-norm with respect to t, $\beta_{i,r} \equiv N_{i,r}(T)$, $y_{i,r}(t) \equiv \{N_{i,r}(t + \Delta t) - N_{i,r}(t)\}\Delta t^{-1}$, and Δt represents a infinitesimally small value. In (2.5), the objective function $L_1(a', \mathbf{f}', \mathbf{v})$ measures the discrepancy between the empirical and estimated distributions of event timings, where $y_{i,r}(t)N_{i,r}(T)^{-1}$ and $\{a' + \sum_{m\in[M]} S^{v_{i,m}+w^*_{r,m}} f'_m(t)\}$ serve as the *empirical distribution* and the *estimated distribution* of $\{t_{i,r,j} : j = 1, \ldots, N_{i,r}(T)\}$, respectively. The term $\beta_{i,r}$ serves as the weight of the counting process $N_{i,r}(t)$. For instance, setting $\beta_{i,r} = N_{i,r}(T)$ means equal weights for all events. Finally, we estimate Λ using the empirical mean

$$\hat{\Lambda} \equiv (nR)^{-1} \sum_{i\in[n],r\in[R]} N_{i,r}(T). \tag{2.6}$$

The parameters $a, \mathbf{f}$ can be estimated from $\hat{\Lambda}, \hat{a}', \hat{\mathbf{f}}'$ by $\hat{a} \equiv \hat{a}'\hat{\Lambda}$ and $\hat{\mathbf{f}} \equiv \hat{\mathbf{f}}'\hat{\Lambda}$.

2.3. Additive Shape Invariant Mixture Model

2.3.1. Model. We extend model (2.1) to simultaneously perform decomposition, alignment, and clustering. Assume that the n subjects can be classified into K distinct clusters, that is, $[n] = \cup_{k=1}^K \mathcal{C}_k$, where $\mathcal{C}_1, \ldots, \mathcal{C}_K$ represent mutually exclusive subsets of $[n]$. These clusters are delineated based on the similarity of intensity components across subjects. Specifically, we introduce the following model:

$$\lambda_{i,r}(t) = a_{z_i} + \sum_{m\in[M]} S^{v_{i,m}+w^*_{r,m}} f_{z_i,m}(t), \tag{2.7}$$

where $z_i \in [K]$ represents the *cluster membership* of subject i such that $z_i = k$ if $i \in \mathcal{C}_k$. We refer to the model in (2.7) as the *additive shape invariant mixture model*, or ASIMM for short. Conditioning on each cluster, the additive shape invariant mixture model in (2.7) simplifies to the additive shape invariant model in (2.1). Similar to the connection between model in (2.1) and FPCA, the additive shape invariant mixture model in (2.7) has a close connection with clustering methods based on FPCA [Chiou and Li, 2007, Bouveyron and Jacques, 2011, Jacques and Preda, 2013, Yin et al., 2021].

In the context of neural data [Steinmetz et al., 2019], $\mathcal{C}_k$'s can be interpreted as functional groups of neurons, wherein neurons exhibit similar firing patterns. The additive shape invariant mixture model in (2.7) enables us to simultaneously identify functional groups of neurons (i.e., $\mathcal{C}_k$'s), discern representative neural firing patterns (i.e., $f_{k,m}$'s), and estimate individual neural response latencies (i.e., $v_{i,m}$'s). It is worthwhile to emphasize that the applicability of the proposed method extends beyond neural data analysis. For instance, the additive shape invariant mixture model in (2.7) can be employed in analyzing recurrent consumer activity in response to advertising [Xu et al., 2014, Zadeh and Sharda, 2014, Tanaka et al., 2016, Bues et al., 2017], or studying hospital admission rates following the implementation of societal disease prevention policies [Barone-Adesi et al., 2006, Sims et al., 2010, Klevens et al., 2016, Evans et al., 2021].

2.3.2. Identifiability. We denote the collection of unknown parameters in (2.7) as $\mathbf{z} \equiv (z_i)_{i\in[n]}$, $\mathbf{a} \equiv (a_k)_{k\in[K]}$, $\mathbf{f} \equiv (f_{k,m})_{k\in[K],m\in[M]}$, $\mathbf{v} \equiv (v_{i,m})_{i\in[n],m\in[M]}$. Let $(\mathbf{z}^*, \boldsymbol{a}^*, \mathbf{f}^*, \mathbf{v}^*) \in \Theta_1$ denote the true parameters, where $\Theta_1 \equiv \{(\mathbf{z}, \boldsymbol{a}, \mathbf{f}, \mathbf{v}) : \mathbf{z} \in [K]^n, \boldsymbol{a} \in [0,\infty)^K, \mathbf{f} \in \mathcal{F}^{K\times M}, \mathbf{v} \in [0, V]^{n\times M}\}$. When conditioning on $\mathbf{z}^*$, Proposition 2.2.1 establishes the identifiability of $\boldsymbol{a}^*, \mathbf{f}^*, \mathbf{v}^*$. However, to ensure the identifiability of $\mathbf{z}^*$, an additional assumption regarding the separability of clusters is required. The formal presentation of model identifiability is provided in Proposition 2.3.1.

PROPOSITION 2.3.1. *Suppose that Assumptions C2.1 and C2.2 hold, and furture assume that*

(C2.3) For any $k, k' \in [K]$ that $k \neq k'$, there exists $m_0 \in [M]$ such that for any $x \in \mathbb{R}$, $\{t \in \mathbb{R} : S^x f^*_{k,m_0}(t) \neq f^*_{k',m_0}(t)\}$ has a positive measure.

Then, we can verify the following statements hold.

(P2.4) The cluster memberships $\mathbf{z}^*$ are identifiable up to a permutation of cluster labels.

(P2.5) The baseline values $\boldsymbol{a}^*$ are identifiable up to a permutation of cluster labels.

(P2.6) *The response components* $\mathbf{f}^*$ *are identifiable up to a permutation of cluster labels and time shifts.*

(P2.7) *For* $k \in [K], m \in [M]$ *such that the set* $\{t : f^*_{k,m}(t) \neq 0\}$ *is of positive measure, the set* $(v^*_{i,m})_{i \in \mathcal{C}^*_k}$ *is identifiable up to a constant independent of* i*.*

Assumption C2.3 mandates that each cluster exhibits at least one signature intensity component that is unique to this cluster. Statement P2.4 directly stems from Assumption C2.3. Statements P2.5, P2.6, and P2.7 can be derived by applying Proposition 2.2.1 to each individual cluster. Detailed proof of Proposition 2.3.1 is presented in Section A.2 of Appendix A.

2.3.3. Estimation. We estimate the parameters in (2.7) by generalizing the optimization approach in Section 2.2.4 to incorporate the cluster structure. For any $k \in [K]$ and $m \in [M]$, let $\Lambda_k \equiv a_k T + \sum_{m \in [M]} \int_0^T f_{k,m}(t)\mathrm{d}t$, $a'_k \equiv a_k \Lambda_k^{-1}$, and $f'_{k,m} \equiv f_{k,m}\Lambda_k^{-1}$. Denoting $\mathbf{a}' \equiv (a'_k)_{k \in [K]}$, $\mathbf{f}' \equiv (f'_{k,m})_{k \in [K], m \in [M]}$, and $\mathbf{\Lambda} \equiv (\Lambda_k)_{k \in [K]}$, we propose the following optimization problem:

$$\hat{\mathbf{z}}, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \hat{\mathbf{v}}, \hat{\mathbf{\Lambda}} \equiv \underset{\mathbf{z}, \mathbf{a}', \mathbf{f}', \mathbf{v}, \mathbf{\Lambda}}{\arg\min} \left\{ L_1(\mathbf{z}, \mathbf{a}', \mathbf{f}', \mathbf{v}) + \gamma \, L_2(\mathbf{z}, \mathbf{\Lambda}) \right\}, \tag{2.8}$$

where

$$L_1(\mathbf{z}, \mathbf{a}', \mathbf{f}', \mathbf{v}) \equiv \sum_{i \in [n], r \in [R]} \beta_{i,r} \, \frac{1}{T} \left\| \frac{y_{i,r}(t)}{N_{i,r}(T)} - \left\{ a'_{z_i} + \sum_{m \in [M]} S^{v_{i,m} + w^*_{r,m}} f'_{z_i,m}(t) \right\} \right\|_t^2, \tag{2.9}$$

$$L_2(\mathbf{z}, \mathbf{\Lambda}) \equiv \sum_{i \in [n], r \in [R]} \left| N_{i,r}(T) - \Lambda_{z_i} \right|^2, \tag{2.10}$$

and $\gamma \in (0, \infty)$ is a tuning parameter. In essence, L_1 and L_2 assess the within-cluster variance of event time distributions and event counts, respectively. When the number of clusters is reduced to one (i.e., $K = 1$), the definitions of L_1 in (2.9) is identical to the definition in (2.5). The tuning parameter γ modulates the relative importance of L_2 compared to L_1 in the optimization with respect to $\mathbf{z}$. When γ is sufficiently small, the estimator $\hat{\mathbf{z}}$ defined in (2.8) is predominantly determined by L_1, resulting in a potentially suboptimal value of L_2. Conversely, when γ is sufficiently large, the dominance shifts towards L_2, resulting in a $\hat{\mathbf{z}}$ that achieves the minimum value of L_2, while L_1 may be relegated to suboptimal values. Subsequent to addressing the optimization problem in (2.8), the estimations of $\mathbf{a}$ and $\mathbf{f}$ can be established via $\hat{a}_k \equiv \hat{a}'_k \hat{\Lambda}_k$ and $\hat{\mathbf{f}}_{k,m} \equiv \hat{\mathbf{f}}'_{k,m} \hat{\Lambda}_k$ for $k \in [K]$, $m \in [M]$.

The optimization problem in (2.8) involves two tuning parameters γ and K. To determine these tuning parameters, we employ a heuristic method. We first establish a preliminary estimate of K using simple methods, such as applying the k-means algorithm on $N_{i,r}(T)$'s and selecting K using the elbow method [Thorndike, 1953]. Given this preliminary estimation of K, we choose the largest γ before observing a significant increase in L_1. We provide simulation experiments to justify this heuristic method and demonstrate the robustness of selected γ to changes of the preliminary selection of K in Section A.6.2. Finally, conditioning on the selected γ, we refine the value of K by identifying the elbow point on the curve of the objective function in (2.8) against K.

2.4. Algorithm

We now present an algorithm for solving the optimization problem (2.8). This optimization problem aims to minimize the within-cluster variances pertaining to event time distributions and event counts. To this end, we propose an algorithm that resembles the k-means algorithm that alternates between a *centering step* and a *clustering step* presented as follows.

$$\text{(centering step)} \quad \hat{\mathbf{a}}', \hat{\mathbf{f}}' = \arg\min_{\mathbf{a}', \mathbf{f}'} L_1(\hat{\mathbf{z}}, \mathbf{a}', \mathbf{f}', \hat{\mathbf{v}}), \quad \hat{\mathbf{\Lambda}} = \arg\min_{\mathbf{\Lambda}} L_2(\hat{\mathbf{z}}, \mathbf{\Lambda}), \tag{2.11}$$

$$\text{(clustering step)} \quad \hat{\mathbf{z}}, \hat{\mathbf{v}} = \arg\min_{\mathbf{z}, \mathbf{v}} \{L_1(\mathbf{z}, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \mathbf{v}) + \gamma\, L_2(\mathbf{z}, \hat{\mathbf{\Lambda}})\}. \tag{2.12}$$

In the centering step (2.11), the estimators of $\mathbf{a}'$, $\mathbf{f}'$ and $\mathbf{\Lambda}$ are updated conditioned on the values of $\hat{\mathbf{z}}$ and $\hat{\mathbf{v}}$. In the clustering step (2.12), the estimators of $\mathbf{z}$ and $\mathbf{v}$ are updated conditioned on the values of $\hat{\mathbf{a}}'$, $\hat{\mathbf{f}}'$ and $\hat{\mathbf{\Lambda}}$. This alternating scheme facilitates a closed-form solution in the centering step and an effective optimization process in the clustering step.

2.4.1. The centering step. The centering step (2.11) involves two optimization problems. To solve the first optimization problem, we formulate it in the frequency domain as follows:

$$\hat{\mathbf{a}}', \hat{\boldsymbol{\phi}}' = \arg\min_{\mathbf{a}', \boldsymbol{\phi}'} \sum_{i\in[n], r\in[R]} \beta_{i,r} \sum_{l\in\mathbb{Z}} \left| \frac{\eta_{i,r,l}}{N_{i,r}(T)} - \left\{ a'_{\hat{z}_i} \mathbf{1}(l=0) + \sum_{m\in[M]} \exp\left\{ -\mathrm{j}\, 2\pi l (\hat{v}_{i,m} + w^*_{r,m}) T^{-1} \right\} \phi'_{\hat{z}_i,m,l} \right\} \right|^2, \tag{2.13}$$

where $\boldsymbol{\phi}' \equiv (\phi'_{k,m,l})_{k\in[K],m\in[M],l\in\mathbb{Z}}$, $\{\phi'_{k,m,l} : l \in \mathbb{Z}\}$ denotes the Fourier coefficients of $f'_{k,m}(t)$, $\{\eta_{i,r,l} : l \in \mathbb{Z}\}$ denotes the Fourier coefficients of $y_{i,r}(t)$, and j denotes the imaginary unit. Notably,

the multiplicative term $\exp(-\mathrm{j}\,2\pi l[\hat{v}_{i,m}+w^*_{r,m}]T^{-1})$ serves as the frequency domain counterpart of the shift operator $S^{\hat{v}_{i,m}+w^*_{r,m}}$. In essence, the Fourier transformation converts the *shift operators* into *multiplication*. As a result, the objective function becomes more tractable compared to its counterpart in the original domain. Indeed, an analytical solution for $\hat{\boldsymbol{\phi}}'$ can be derived. Let $\boldsymbol{\phi}'_{k,*,l} \equiv (\phi'_{k,m,l})_{m\in[M]}$ for any $k \in [K]$ and $l \in \mathbb{Z}$. For $l \neq 0$, the solution to (2.13) with respect to $\boldsymbol{\phi}'_{k,*,l}$ can be expressed as follows:

$$\hat{\boldsymbol{\phi}}'_{k,*,l} = \left(\overline{\mathbf{E}_{k,l}}^{\top}\mathbf{B}_k\ \mathbf{E}_{k,l}\right)^{-1}\left(\overline{\mathbf{E}_{k,l}}^{\top}\mathbf{B}_k\ \mathbf{h}_{k,l}\right), \qquad \text{for } l \neq 0, \tag{2.14}$$

$$\hat{\boldsymbol{\phi}}'_{k,*,0} = -\sum_{|l|\leq \ell_0, l\neq 0} \hat{\boldsymbol{\phi}}'_{k,*,l}, \qquad \text{for } l = 0. \tag{2.15}$$

Here, $\mathbf{E}_{k,l}$ is defined as

$$\mathbf{E}_{k,l} \equiv [\exp\{-\mathrm{j}\,2\pi l(\hat{v}_{i,m}+w^*_{r,m})T^{-1}\}]_{(i,r)\in\hat{\mathcal{C}}_k\times[R],m\in[M]}, \tag{2.16}$$

where $\hat{\mathcal{C}}_k \equiv \{i \in [n] : \hat{z}_i = k\}$. $\overline{\mathbf{E}_{k,l}}$ denotes the complex conjugate of $\mathbf{E}_{k,l}$, $\mathbf{B}_k$ is a diagonal matrix of $(\beta_{i,r})_{(i,r)\in\hat{\mathcal{C}}_k\times[R]}$, $\mathbf{h}_{k,l} \equiv (\eta_{i,r,l}N_{i,r}(T)^{-1})_{(i,r)\in\hat{\mathcal{C}}_k\times[R]}$, and $\ell_0 \in \mathbb{N}$ is a truncation parameter introduced to ensure the numerical feasibility of computing $\hat{\boldsymbol{\phi}}'_{k,*,0}$. For $l \neq 0$, the objective function concerning $\boldsymbol{\phi}'_{k,*,l}$ in (2.13) is a weighted sum of squares, hence the estimate in (2.14) can be derived using the well-known least squares estimator. For $l = 0$, the estimate in (2.15) can be obtained by exploiting the definition of $\mathcal{F}$. The detailed derivations of (2.14) and (2.15) can be found in Section A.4.

Upon obtaining $\hat{\boldsymbol{\phi}}'$, the solution to the first optimization problem in (2.11) can be derived as follows. For $k \in [K]$ and $m \in [M]$,

$$\hat{a}'_k = T^{-1} - \sum_{m\in[M]} \hat{\phi}'_{k,m,0}, \tag{2.17}$$

$$\hat{f}'_{k,m}(t) = \sum_{|l|\leq\ell_0} \hat{\phi}_{k,m,l}\exp(\mathrm{j}\,2\pi l t T^{-1}), \tag{2.18}$$

where (2.17) is obtained by substituting $\hat{\boldsymbol{\phi}}'$ into (2.13), and (2.18) follows from the inverse Fourier transformation.

The second optimization problem in (2.11) aims to minimize the within-cluster variances of event counts given cluster memberships. The solution to this optimization problem is straightforward:

for any $k \in [K]$,

$$\hat{\Lambda}_k = \arg\min_{\Lambda_k} \sum_{i \in \hat{\mathcal{C}}_k, r \in [R]} \left| N_{i,r}(T) - \Lambda_k \right|^2 = \left(|\hat{\mathcal{C}}_k| \; R \right)^{-1} \sum_{i \in \hat{\mathcal{C}}_k, r \in [R]} N_{i,r}(T), \tag{2.19}$$

where $|\hat{\mathcal{C}}_k|$ denotes the cardinality of the set $\hat{\mathcal{C}}_k$. In summary, the solution to the centering step is encapsulated by equations (2.17), (2.18), and (2.19).

2.4.2. The clustering step. The optimization problem in (2.12) can be scaled down to the subject level, allowing for the independent estimation of parameters associated with each subject. For any subject $i \in [n]$, let $\mathbf{v}_i \equiv (v_{i,m})_{m \in [M]}$ denote its associated time shifts. In (2.12), the parameters z_i and $\mathbf{v}_i$ are estimated through the following sub-problem:

$$\hat{z}_i, \hat{\mathbf{v}}_i = \arg\min_{z_i, \mathbf{v}_i} \{ L_{1,i}(z_i, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \mathbf{v}_i) + \gamma \; L_{2,i}(z_i, \hat{\mathbf{\Lambda}}) \}, \tag{2.20}$$

where $L_{1,i}$ and $L_{2,i}$ are defined as

$$L_{1,i}(z_i, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \mathbf{v}_i) \equiv \sum_{r \in [R]} \beta_{i,r} \; \frac{1}{T} \left\| \frac{y_{i,r}(t)}{N_{i,r}(T)} - \left\{ \hat{a}'_{z_i} + \sum_{m \in [M]} S^{v_{i,m} + w^*_{r,m}} \hat{f}'_{z_i,m}(t) \right\} \right\|_t^2, \tag{2.21}$$

$$L_{2,i}(z_i, \hat{\mathbf{\Lambda}}) \equiv \sum_{r \in [R]} \left| N_{i,r}(T) - \hat{\Lambda}_{z_i} \right|^2. \tag{2.22}$$

The parameter dimension for the problem in (2.20) is significantly reduced compared to the original optimization problem in (2.12). Consequently, solving the problem in (2.20) is computationally more efficient than addressing the original problem stated in (2.12).

To solve the optimization problem in (2.20), we employ the following procedure:

$$\tilde{\mathbf{v}}_{i|k} = \arg\min_{\mathbf{v}_i} \; L_{1,i}(k, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \mathbf{v}_i), \quad \text{for } k \in [K], \tag{2.23}$$

$$\hat{z}_i = \arg\min_{z_i \in [K]} \{ L_{1,i}(z_i, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \tilde{\mathbf{v}}_{i|z_i}) + \gamma \; L_{2,i}(z_i, \hat{\mathbf{\Lambda}}) \}, \tag{2.24}$$

$$\hat{\mathbf{v}}_i = \tilde{\mathbf{v}}_{i|\hat{z}_i}. \tag{2.25}$$

In the first step (2.23), we determine the optimal time shift for each potential cluster membership $k \in [K]$. The optimization problem in this step can be solved in the frequency domain using the Newton's method (see Section A.5). In the second step (2.24), we evaluate the objective function for each possible cluster membership, leveraging the optimal time shift corresponding to that particular

cluster. Subsequently, we designate the cluster associated with the minimal objective function value as the estimated cluster membership. In the last step (2.25), we choose the optimal time shift for the estimated cluster membership. In summary, the solution to the clustering step is encapsulated by (2.23), (2.24), and (2.25).

2.4.3. Overall estimation procedure. The objective function in (2.8) exhibits a non-convex nature. This characteristic poses a crucial need for an appropriate initialization scheme and convergence criterion. Our proposed initialization scheme is described in Remark 2.4.1. The overall estimation procedure is summarized in Algorithm 2.1.

Algorithm 2.1: Iterative algorithm for ASIMM

Input: $\{N_{i,r}(t) : i \in [n], r \in [R]\}, K, \gamma, \ell_0$

Initialize $\hat{\mathbf{v}}^{(0)}, \hat{\mathbf{z}}^{(0)}$ via (2.26), (2.27), let $s = 0$, and $L^{(0)} = \infty$;

while *not stop* **do**

Update $\hat{\mathbf{a}}'^{(s+1)}$, $\hat{\mathbf{f}}'^{(s+1)}$, $\hat{\mathbf{\Lambda}}^{(s+1)}$ via (2.17) - (2.19) given $(\hat{\mathbf{z}}^{(s)}, \hat{\mathbf{v}}^{(s)})$;

Update $\hat{\mathbf{z}}^{(s+1)}$, $\hat{\mathbf{v}}^{(s+1)}$ via (2.23) - (2.25), given $(\hat{\mathbf{a}}'^{(s+1)}, \hat{\mathbf{f}}'^{(s+1)}, \hat{\mathbf{\Lambda}}^{(s+1)})$;

Evaluate the loss function: $L^{(s+1)} \equiv L_1(\hat{\mathbf{z}}^{(s+1)}, \hat{\mathbf{a}}'^{(s+1)}, \hat{\mathbf{f}}'^{(s+1)}, \mathbf{v}^{(s+1)}) + \gamma\, L_2(\hat{\mathbf{z}}^{(s+1)}, \hat{\mathbf{\Lambda}}^{(s+1)})$;

Evaluate the stopping criterion: $\{L^{(s)} - L^{(s+1)}\}/L^{(s+1)} \leq \epsilon$;

$s = s + 1$;

end

Output: $\hat{\mathbf{z}}^{(s)}, \hat{\mathbf{a}}'^{(s)}, \hat{\mathbf{f}}'^{(s)}, \hat{\mathbf{v}}^{(s)}, \hat{\mathbf{\Lambda}}^{(s)}$.

REMARK 2.4.1. ***Initialization.*** *The subject-specific time shifts are initialized based on the earliest event occurrence following each stimulus. Specifically, the value of* $\hat{\mathbf{v}}^{(0)}$ *is defined as follows. For* $i \in [n]$, $m \in [M]$,

$$\hat{v}_{i,m}^{(0)} \equiv \min\{t_{i,r,j} - w_{r,m}^* : t_{i,r,j} > w_{r,m}^*, r \in [R], j \in [N_{i,r}(T)]\}. \tag{2.26}$$

The cluster memberships are initialized based on the adjusted *event times that roughly aligned the point processes using the initial subject-specific time shifts in* (2.26)*. These adjusted event times are calculated by shifting the event times associated with each stimulus to an anchor point for that stimulus. To be specific, we shift the event times as* $\tilde{t}_{i,r,j} \equiv t_{i,r,j} - \hat{u}_{i,r,m}^{(0)} + \min_{r' \in [R]} w_{r',m}^*$ *for event* j *where* $t_{i,r,j} \in [\hat{u}_{i,r,m}^{(0)}, \hat{u}_{i,r,m+1}^{(0)}]$*. Here,* $\hat{u}_{i,r,m}^{(0)} \equiv \hat{v}_{i,m}^{(0)} + w_{r,m}^*$ *denotes the total time shift associated with stimulus* m *for* $m \in [M]$*,* $\hat{u}_{i,r,M+1}^{(0)} \equiv T$*, and* $\min_{r' \in [R]} w_{r',m}^*$ *represents the anchor point of stimulus*

m. Subsequently, cluster memberships are initialized by applying the k-means algorithm on adjusted event times:

$$\hat{\mathbf{z}}^{(0)} \equiv \arg\min_{\mathbf{z}} \sum_{k\in[K]} \sum_{i,j:z_i=z_j=k} \left\| y_i'(t) - y_j'(t) \right\|^2, \tag{2.27}$$

where $y_i'(t)$ denotes the empirical distribution of $\{\tilde{t}_{i,r,j} : r \in [R], j \in [N_{i,r}(T)]\}$. The efficacy of the proposed initialization approach is illustrated in Figure A.5 of Appendix A, where it is shown to outperform the random initialization with multiple restarts.

2.5. Simulation

2.5.1. Simulation experiment design. We assess the performance of the proposed method in three synthetic experiments. In the first experiment, we aim to explore the intensity decomposition performance. To this end, we generate Poisson processes $N_{i,r}$'s whose intensities follow the additive shape invariant model in (2.1). Specifically, we set $T = 2.5$, $M = 2$, $v_{i,1} \sim \text{Unif}(0, 1/64)$, $v_{i,2} \sim \text{Unif}(0, 1/16)$, $w_{r,1}^* \sim \text{Unif}(0, \tau)$, $w_{r,2}^* \sim \text{Unif}(0.8, 0.8 + \tau)$. These parameter values remain unchanged across all subsequent experiments. In addition, we set $a = 20$, and set $\mathbf{f}$ as

$$\begin{aligned} f_1(t) &= 70 \times \left[\{2 - 2\cos(4\pi[t - 0.4])\} \times \mathbf{1}(t \in [0.4, 0.9])\right] \equiv 70 \times q_1(t), \\ f_2(t) &= 70 \times \left[\{2 - 2\cos(2\pi|2t|^{1/2})\} \times \mathbf{1}(t \in [0, 0.5])\right] \equiv 70 \times q_2(t). \end{aligned} \tag{2.28}$$

Notably, $q_1(t)$ and $q_2(t)$ capture the event time distribution of the two components, and both $q_1(t)$ and $q_2(t)$ integrate to unity. We tune the signal strength in synthetic data by altering the values of R, n and τ. Intuitively, R and n serve as the sample size, hence are positively associated with signal strength. And τ is associated with the variance of $w_{r,m}^*$'s and thus the identifiability of the intensity components.

In the second experiment, we compare the clustering performance of the proposed ASIMM (2.7) and two relevant methods: kCFC, introduced in Chiou and Li [2007], and k-mean alignment, introduced in Sangalli et al. [2010]. Here we do not apply the method by Yin et al. [2021] since it considers a multiplicative model whereas our model assumes additive components. We generate Poisson processes using the model in (2.7). In particular, we set the true cluster memberships by sequentially assigning $n = 40$ subjects into $K = 4$ clusters of equal size, that is, $z_i = \lceil (i/n)K \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. In addition, we set $a_k = 20$, and set $f_{k,m}$'s as shown in

TABLE 2.1. True values of $\{f_{k,m}(t) : k \in [K], m \in [M]\}$ in Scenario 2. The parameter ρ controls the distinctiveness across clusters, whose value is altered in the experiment. The functions $q_1(t)$ and $q_2(t)$ are defined in (2.28), $h_1(x) \equiv |\max(x, 0)|^{1/2}$, and $h_2(x) \equiv 1 + \min(x, 0)$.

	$m = 1$	$m = 2$
$k = 1$	$52.5 \times q_1(t)$	$52.5 \times q_2(t)$
$k = 2$	$60 \times [1 - h_1(2\rho - 1)] \times q_1(t)$ $+48 \times h_2(2\rho - 1) \times q_2(2[t - 0.8])$	$60 \times [1 + h_1(2\rho - 1)] \times q_2(t)$ $-48 \times h_2(2\rho - 1) \times q_2(2t)$
$k = 3$	$67.5 \times (1 + 0.5\rho) \times q_1(t)$	$67.5 \times (1 - 0.5\rho) \times q_2(t)$
$k = 4$	$75 \times (1 + \rho) \times q_1(t)$	$75 \times (1 - \rho) \times q_2(t)$

Table 2.1. In Table 2.1, we introduce $\rho \in (0, 1)$ to control the distinctiveness of clusters. Consider the mean intensity associated with the k-th cluster. For $i \in \mathcal{C}_k$ and $r \in [R]$, we can derive from (2.7) that $\mathbb{E}\lambda_{i,r}(t) = a_k + \sum_{m\in[M]}(p_m \star f_{k,m})(t)$, where p_m denotes the probability density function of $v_{i,m} + w^*_{r,m}$, and $\star$ denotes the convolution operator that $(p_m \star f_{k,m})(t) = \int_0^{T_0} p_m(t - x)f_{k,m}(x)\mathrm{d}x$. When $\rho = 0$, the shapes of $\sum_{m\in[M]}(p_m \star f_{k,m})(t)$'s are identical across clusters, whereas when $\rho = 1$, the shapes of $\sum_{m\in[M]}(p_m \star f_{k,m})(t)$'s exhibit substantial distinctions across clusters. In essence, as ρ increases, the clusters become more separable.

The third experiment is a continuation of the second experiment, where our focus is directed towards evaluating the clustering performance of the proposed method. We manipulate signal strength by varying the values of variables R, n, and τ. Through this manipulation, we aim to investigate the impact of R, n, and τ on the performance of clustering estimation.

2.5.2. Intensity estimation performance. In the first experiment, we investigate the effect of varying R, n and τ on the intensity estimation performance of the proposed method. When $K = 1$, the value of γ does not affect the estimation result, hence can be set to zero. We let $\ell_0 = 10$, and $\epsilon = 0.005$ in all experiments. A sensitivity analysis concerning ℓ_0 demonstrates the robustness of estimation results to changes in ℓ_0, as detailed in Section A.6.3 of Appendix A.

We evaluate the intensity estimation performance via the mean integrated squared error (MISE). The MISE is defined as:

$$\text{MISE} \equiv \frac{1}{M}\sum_{m\in[M]} d\left\{\hat{f}'_m(t), \frac{f^*_m(t)}{\Lambda^*}\right\}. \tag{2.29}$$

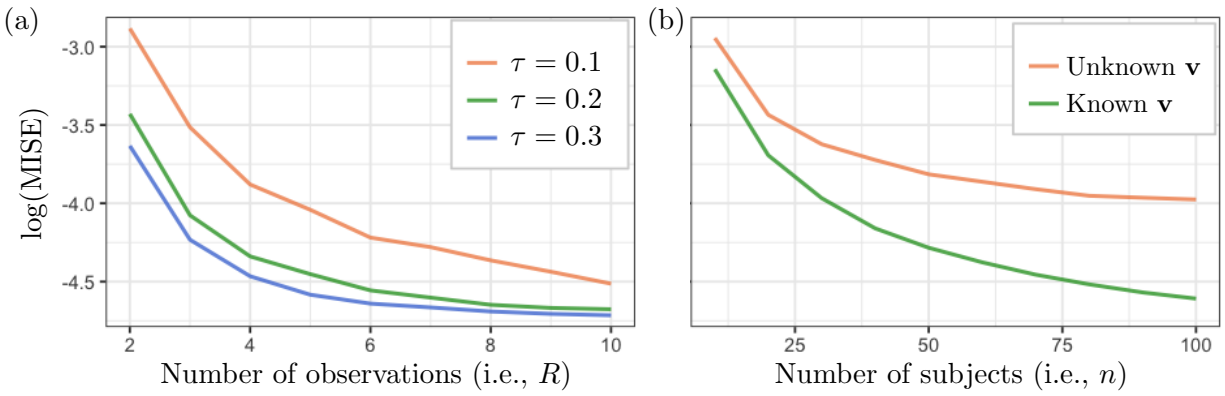

FIGURE 2.3. Intensity estimation performance in Experiment 1 with 5000 replicates. Synthetic data is generated with varying R, n, and τ. MISE is shown in log scale for better visualization, where smaller values indicate better performances. Panel (a) shows the performance of intensity estimation with varying values of R and τ. Panel (b) demonstrates the performance of intensity estimation with varying n. The curve labeled "Unknown $\mathbf{v}$" shows results when the algorithm is not provided with the true value of $\mathbf{v}$, while the curve labeled "Known $\mathbf{v}$" depicts results when the algorithm is provided with the true value of $\mathbf{v}$.

where $d\{f_1, f_2\} \equiv \min_{v\in[-T,T]} \|S^v f_1 - f_2\|^2$. It is worth noting that the definition of MISE considers $\hat{\mathbf{f}}'$ rather than $\hat{\mathbf{f}}$. We excludes $\hat{\Lambda}$ from the evaluation criterion since its performance, as a sample mean, is well-studied.

The intensity estimation performance is shown in Figure 2.3. Firstly, the MISE rapidly improves as R increases, because when R increases, each subject is associated with more samples, while the number of unknown parameters remain constant. Secondly, a decrease in MISE is observed as τ increases. This is because when τ is small, there is a potential non-identifiability issue due to limited sample size. As τ increases, the variance of $w^*_{r,m}$'s increases, thereby alleviating the non-identifiability issue. Thirdly, the MISE exhibits a decreasing trend with an increase in n, since n serves as the sample size for the estimation of intensity components. However, it is noteworthy that the MISE decreases slower in response to an increase in n compared to an increase in R. This is because an increment in n leads to a proportional increase in the quantity of unknown subject-specific time shifts (i.e., $v_{i,m}$'s). Consequently, when the algorithm is provided with the true values of $v_{i,m}$'s, the MISE shows a significant reduction.

2.5.3. Comparison with relevant methods. In the second experiment, we compare the clustering performance of the proposed ASIMM (2.7) with the kCFC [Chiou and Li, 2007] and

the k-mean alignment [Sangalli et al., 2010]. We apply proposed method with $K = 4$, $\gamma = 0.01$, where the selection of γ follows the procedure outlined in Section 2.3.3. We apply the kCFC by employing the implementation provided in the `R` package *fdapace* [Gajardo et al., 2021], specifically the function named "kCFC". The parameters for this implementation are set as follows: the desired number of clusters is specified as 4, and the maximum number of principal components is set to 2. Additionally, we specify the type of design as "dense", and set the maximum number of iterations to 30. We employ the k-mean alignment by utilizing the implementation available in the `R` package *fdasrvf* [Tucker, 2023], specifically the function named "kmeans_align". In configuring the algorithm parameters, we specify the desired number of clusters specified to 4, the maximum number of iterations specified to 50, and the minimum number of curves per cluster to 2. Both the kCFC and the k-mean alignment are applied on the empirical intensities aggregated across observations, which can be expressed as $y_i(t) \equiv \{N_i(t + \Delta t) - N_i(t)\}\Delta t^{-1}$, where $N_i(t) \equiv R^{-1} \sum_{r\in[R]} N_{i,r}(t)$.

We evaluate the cluster estimation performance via the Adjusted Rand Index (ARI) [Hubert and Arabie, 1985]. Let $\boldsymbol{\mathcal{C}}^* \equiv \{\mathcal{C}_k^* : k \in [K]\}$, $\hat{\boldsymbol{\mathcal{C}}} \equiv \{\hat{\mathcal{C}}_{k'} : k' \in [K']\}$ denote the set of true clusters and the set of estimated clusters, where K and K' are the true number of clusters and specified number of clusters. The ARI is formally defined as

$$\text{ARI} \equiv \frac{\sum_{k,k'} \binom{d_{k,k'}}{2} - \left[\sum_k \binom{b_k}{2} \sum_{k'} \binom{c_{k'}}{2}\right] \binom{n}{2}^{-1}}{\frac{1}{2}\left[\sum_k \binom{b_k}{2} + \sum_{k'} \binom{c_{k'}}{2}\right] - \left[\sum_k \binom{b_k}{2} \sum_{k'} \binom{c_{k'}}{2}\right] \binom{n}{2}^{-1}}, \tag{2.30}$$

where $b_k \equiv |\mathcal{C}_k^*|$, $c_{k'} \equiv |\hat{\mathcal{C}}_{k'}|$, and $d_{k,k'} \equiv |\mathcal{C}_k^* \cap \hat{\mathcal{C}}_{k'}|$ for $k \in [K]$ and $k' \in [K']$. The ARI quantifies the similarity between $\boldsymbol{\mathcal{C}}^*$ and $\hat{\boldsymbol{\mathcal{C}}}$. For instance, when $\boldsymbol{\mathcal{C}}^*$ and $\hat{\boldsymbol{\mathcal{C}}}$ are equivalent up to a permutation of cluster labels, the ARI is equal to one. Conversely, when $\hat{\boldsymbol{\mathcal{C}}}$ is entirely random, the ARI has a mean value of zero.

Figure 2.4 shows the clustering performance of the three considered methods. Across all depicted scenarios in Figure 2.4, it is evident that the proposed method consistently outperforms both kCFC and k-mean alignment. The superiority of proposed method becomes especially clear when the values of R and ρ are large. This is expected since the proposed method excels in handling data with additive intensity components and time shifts. In contrast, kCFC and k-mean alignment were not devised to handle the setting in this experiment. In essence, the proposed method serves as a valuable complement to existing approaches for functional clustering.

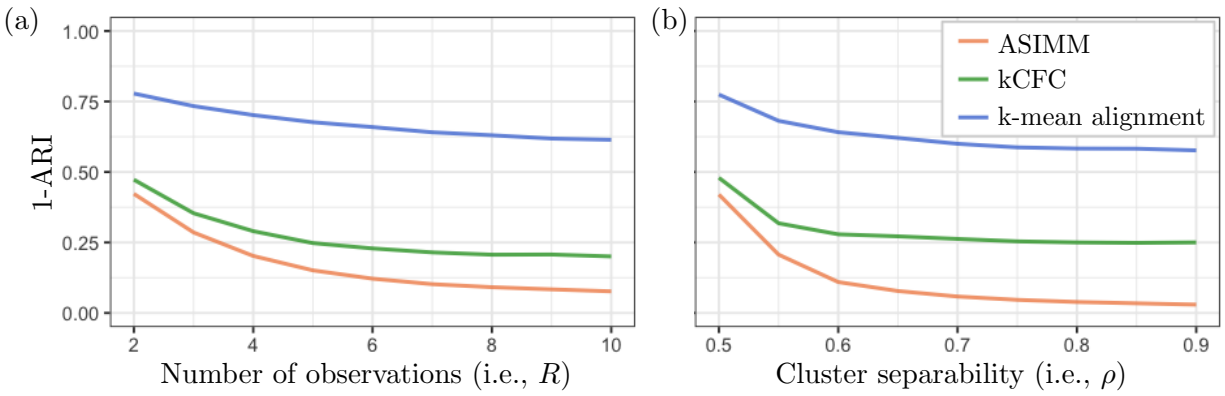

FIGURE 2.4. Clustering performance in Experiment 2 with 5000 replicates of our proposal in orange, kCFC by Chiou and Li [2007] in green, and k-mean alignment by Sangalli et al. [2010] in blue. Synthetic data is generated with $n = 40$, $\tau = 0.1$, varying R and ρ. In panel (a), the value of ρ is fixed as $\rho = 0.5$. In panel (b), the value of R is fixed as $R = 2$.

2.5.4. Cluster estimation performance. In the third experiment, we investigate the effect of R, n and τ on the cluster estimation performance of the proposed method. We apply proposed method with the same set of tuning parameters as in the second experiment.

The clustering performance is displayed in Figure 2.5. It is evident that the clustering performance improves as R, τ, and n increases. This is because R, τ, and n help in estimating the intensity components, as demonstrated in the first experiment, which in turn improves the cluster estimation. Furthermore, R serves as sample size for cluster memberships, thereby contributing to improved clustering performance. However, the impact of increasing n on clustering performance is only marginal. This is because as n increases, the number of unknown cluster memberships also increases, meaning that n does not serve as the sample size for cluster memberships.

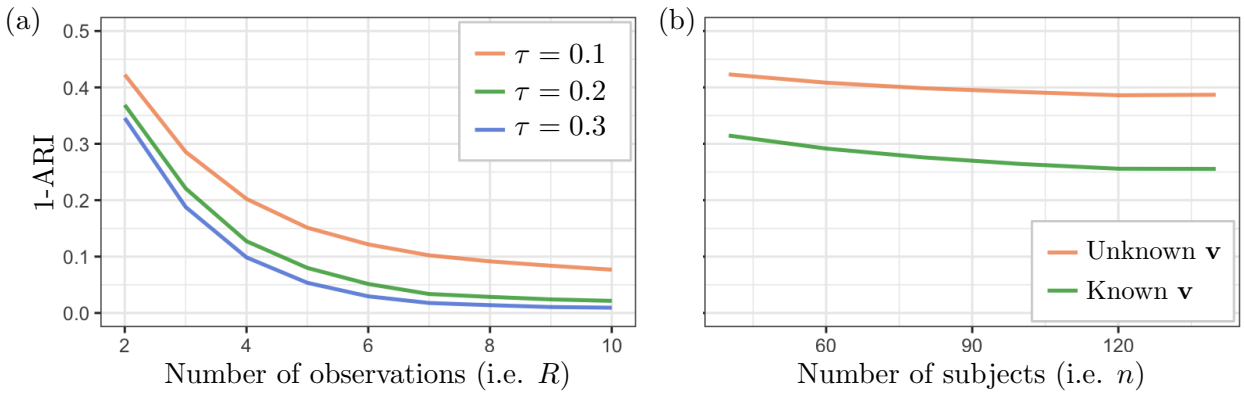

FIGURE 2.5. Clustering performance in Experiment 3 with 5000 replicates. Synthetic data is generated under the setting with varying R, n, and τ. Panel (a) shows the clustering performance as R and τ increases, where the value of n is fixed as $n = 40$. Panel (b) shows the clustering performance as n increases, where the values of R and τ are fixed as $R = 2$ and $\tau = 0.1$. The curve labeled "Unknown $\mathbf{v}$" shows results when the algorithm is not provided with the true value of $\mathbf{v}$, while the curve labeled "Known $\mathbf{v}$" depicts results when the algorithm is provided with the true value of $\mathbf{v}$.

2.6. Real Data Application

We consider the neural data during visual discrimination tasks from Steinmetz et al. [2019]. In each trial, the mouse encountered a sequential presentation of two stimuli. The first stimulus comprised visual gratings of varying contrasts displayed on two screens, one to the left and one to the right of the mouse. The second stimulus was an auditory tone cue which was set off after a randomized delay between 0.4 to 0.8 seconds after the onset of the first stimulus. The mouse could rotate a wheel which, after the auditory cue, would move the visual gratings. When one constrast is higher than the other, the mouse succeeded and gained rewards if the visual grating of higher contrasts was moved to the center screen. The complete criteria for success are detailed in Table A.2 of Appendix A. Throughout the experiment, researchers simultaneously recorded firing activities of hundreds of neurons in the *left* hemisphere of the mouse's brain using Neuropixels [Jun et al., 2017, Steinmetz et al., 2018, 2021]. We aim to identifying groups of neurons with distinct responses to the two stimuli using our proposed method.

Following the notation in (2.7), we index trials using $r \in \{1, \ldots, R\}$ and neurons using $i \in \{1, \ldots, n\}$. Consequently, the firing activities of neuron r in trial i is $N_{i,r}(t), t \in [0, T]$ where $T = 3.5$. We set $m = 1$ for the visual grating and $m = 2$ for the auditory tone cue. We further

denote the onset time of two stimuli in trial r as $w^*_{r,1}$ and $w^*_{r,2}$. Neurons might exhibit firing latencies in response to each stimulus denoted as $v_{i,m}$ for $m \in \{1, 2\}$ and $i \in \{1, \ldots, n\}$. To demonstrate the usage of proposed method, we focus on $R = 102$ experimental trials where the left visual grating was of higher contrast and the mouse successfully gained rewards. We study 264 neurons in the midbrain region. We exclude 39 neurons with fewer than one spike per trial on average, resulting in a total of $n = 225$ neurons for analysis. We apply the proposed algorithm with $K = 3$, $\gamma = 10^{-4}$, $\ell_0 = 10$ and $\epsilon = 0.005$. Using the proposed method, we identify three clusters of neurons with distinct responses to the two stimuli shown in Fig 2.6. The first column of Figure 2.6 contains the refined intensity components. To better understand the roles of each clusters, we illustrate the average firing rates from the training trials and the trials that are not used to fit the model in the second and third columns of Figure 2.6. More details of this analysis can be found in Section A.7 in Appendix A.

Recalling that the first stimulus is the visual gratings and the second stimulus is the auditory cue, we have the following observations.

Cluster 1 (63 neurons). There seems to be only one non-zero component in Cluster 1. It is immediately clear from Figure 2.6(1c) that the firing rates of neurons in Cluster 1 are highly in sync with the wheel velocity. When aligning the firing rates of neurons in Cluster 1 by movement onset in Figure 2.6(1b), we can see that the firing rates share almost the same trajectory across conditions, but their peaks depend on the choice the mouse made. In particular, the firing rate has a higher peak when the mouse chose the left visual grating. This preference, also known as laterality, is likely due to the fact that these neurons are from the left hemisphere. We hypothesize that neurons in Cluster 1 are responsible for executing the turning of the wheel. The crucial role of midbrain in coordinating movement has been identified in prior studies [see, e.g., Boecker et al., 2008, Coddington and Dudman, 2019, Inagaki et al., 2022].

Cluster 2 (23 neurons). The first intensity component exhibits a decrease-then-increase pattern post stimuli onset, and the second component shows a sharp increase after the auditory cue, as shown in Figure 2.6(2a). Figure 2.6(2b) shows that the decrease-then-increase pattern seems common in trials when the mouse chose the left visual grating. There is, however, no period of suppression when the mouse chose the right visual grating. When aligning the firing rates by feedback delivery time, it is clear that the firing rates peaked almost immediately after rewards delivery

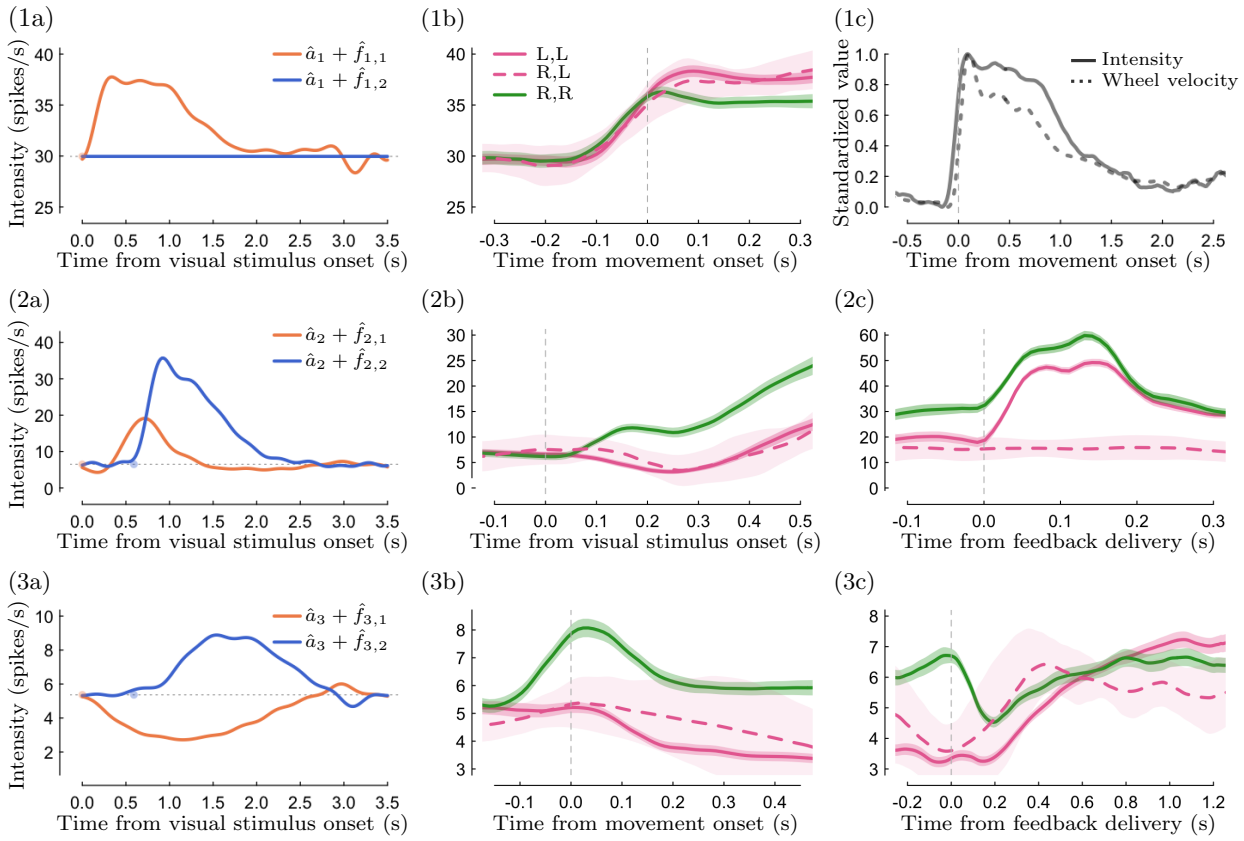

FIGURE 2.6. Estimation of intensity components and average firing intensities in various conditions. Each row corresponds to one estimated cluster. The first column presents the estimated intensity components. The second and third columns display the average neural firing intensities from both the training trials and other trials that are not used to fit the model. The shaded area represents the mean firing intensities plus or minus two standard errors of the mean. The legend in panel (1b) represents "scenario, choice", for instance, "R,L" represents the trials where the right grating was of a higher contrast, and the mouse chose to move the left grating to the center screen resulting in a failure in that trial. Panel (1c) illustrates the average firing intensity and wheel velocity, where both the firing intensity and the wheel velocity are standardized to range from 0 to 1 for alignment. The condition "L,R" is omitted in this figure since there are only three trials, but its firing intensity can be found in Figure A.10 of Appendix A.

regardless of the choice, and, in the absence of rewards, the firing rates remain stationary. We hypothesize that neurons in Cluster 2 might respond to perceptions of stimuli (e.g., visual gratings, rewards), and their activities are further regulated by laterality. Similar firing patterns of neurons in midbrain have been identified in prior studies [Coddington and Dudman, 2018, Steinmetz et al., 2019].

Cluster 3 (139 neurons). The estimated first component exhibits suppressed activities throughout the trial, and the second component shows increased activities after the auditory cue in Figure 2.6(3a). Neural responses in Cluster 3 bear a resemblance to those in Cluster 2 that they both peaked after reward delivery, and they both show preference to contralateral choices (i.e., choosing the right visual grating), as shown in Figures 2.6(3b) and 2.6(3c). Unlike Cluster 2 where rewards trigger immediate response, there is a one-second delay from reward delivery till the firing rates peaked in Cluster 3 in Figure 2.6(3c). Additionally, the activity levels of neurons in Cluster 3 are lower than those in the other two clusters, and the number of neurons in Cluster 3 exceeds that of the other two clusters. It might be possible that neurons in Cluster 3 are not directly involved in the completion of the task. Instead, neurons in Cluster 3 might be responsible for other aspects of the mouse behavior, such as sensing the surrounding environment, or planning the initiation of the next trials. Furthermore, Figure 2.6(3b) shows that the suppression of activities started after the movement onset in trials with ipsilateral choices, but activities increased before the movement onset in trials with contralateral choices.

2.7. Discussion

In this article, we tackle the problem of simultaneously decomposing, aligning, and clustering recurrent event data. We note that there are a few directions that can be explored in future works. First, the proposed method assumes that all subjects exhibit consistent intensity across observations, which may not hold true in cases where the subjects might respond randomly. For instance, neurons may exhibit periods of reduced responsiveness during tasks known as *local sleep* [see, e.g., Krueger et al., 2008, Vyazovskiy et al., 2011, Vyazovskiy and Harris, 2013]. To address this limitation, we can generalize the proposed model by incorporating a hidden Markov model for the responsive status [Tokdar et al., 2010, Escola et al., 2011, Mews et al., 2023]. Second, the proposed method only considers observations from the same experimental condition. A natural extension is to allow observations from different experimental conditions. For instance, it would be beneficial to learn neural functions based on their firing patterns in various visual conditions and feedback types. To achieve this, we can incorporate a cluster structure of observations by employing biclustering models [Madeira and Oliveira, 2004, Slimen et al., 2018, Galvani et al., 2021]. Third, theoretical properties of the proposed estimators have not been investigated. Previous studies have

established the consistency of the k-means clustering [Pollard, 1981, Sun et al., 2012] and clustering based on FPCA [Yin et al., 2021]. In addition, theoretical properties of simultaneous registration and clustering models have been established in Tang et al. [2023]. Building on previously established framework [Pollard, 1981, Yin et al., 2021, Tang et al., 2023], it may be possible to incorporate the theory of shape-invariant models [Bigot et al., 2013, Bigot and Gendre, 2013] to analyze the theoretical properties of the proposed method.

CHAPTER 3

Semiparametric Estimation for Dynamic Networks with Shifted Connecting Intensities

3.1. Introduction

Neural circuits are fundamental components of the nervous system that play a crucial role in instructing animal behavior. Understanding the neural circuit development is beneficial for understanding how the nervous system works, identifying the causes of nervous system disorders, and developing new treatments for nervous system disorders. However, the study of neural circuit development has been limited by difficulty in data acquisition until recently. Modern technologies have enabled the recording of individual neural activities over extended durations, thereby facilitated the exploration of neural circuit development.

Using novel technologies, Wan et al. [2019] recorded the development of a neural circuit in the spine of zebrafish during embryogenesis. The neural circuit development process involves individual neurons forming functional connections, resulting in a highly interconnected network. The process is illustrated in Figure 3.1(a), where nodes represent neurons and edges represent functional connections between neurons. The dataset on neural circuit development exhibits three inherent properties. Firstly, the neural circuit development process is transient, making it impossible to obtain repeated observations of the same system. Secondly, neural circuits from different subjects possess distinct sets of neurons, despite potentially sharing similar underlying developmental mechanisms. Thirdly, neurons have different functional maturation progress, resulting in different *activation times* [Wan et al., 2019]. The neural activation times can obscure the similarity in neural connecting behavior. Figure 3.1(b) depicts two neurons exhibiting seemingly distinct connecting behavior. However, upon adjusting the connection times based on the estimated neural activation times using our proposed method, Figure 3.1(c) shows that the two neurons actually share similar connecting behaviors. Additional examples of neuron pairs with similar connecting behavior can be found in Figure B.14 in Appendix B. In light of these inherent properties, our primary objective

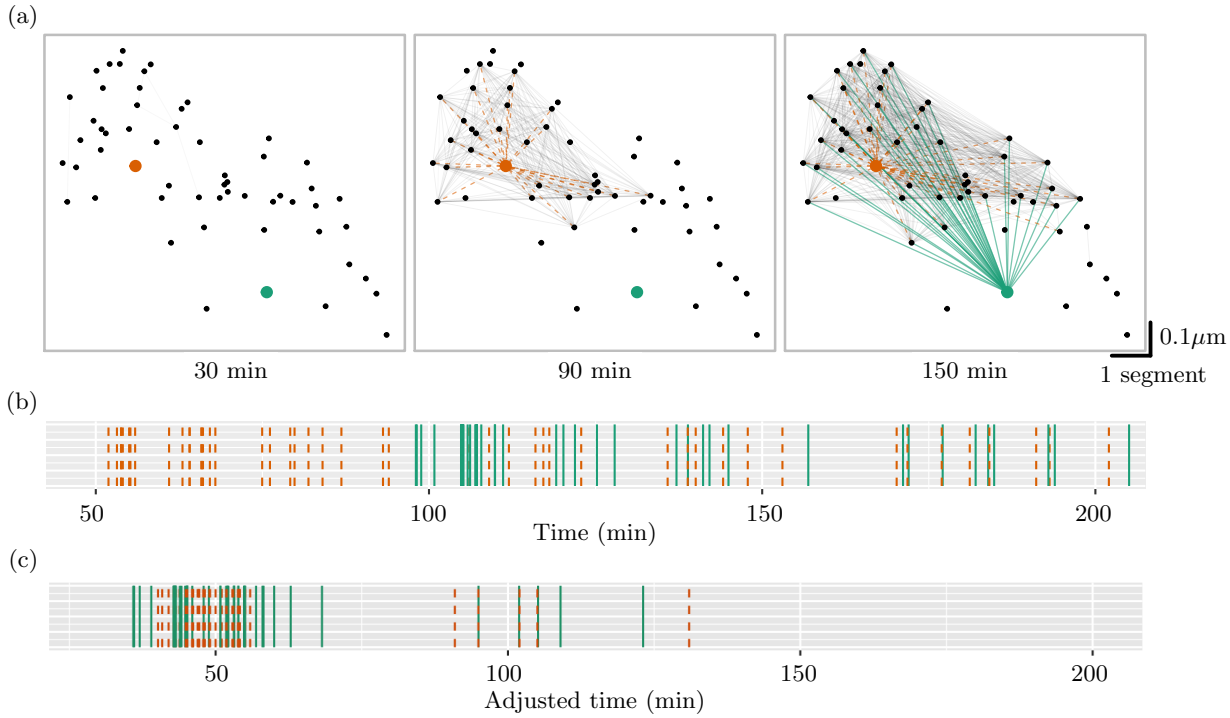

FIGURE 3.1. Isolated neurons form a densely connected network during embryogenesis of zebrafish. All time stamps indicate the time since the experiment starts. Panel (a) displays neural network snapshots at 30, 90 and 150 minutes, where edges appear progressively over time. In the networks, nodes represent neurons, and edges represent functional connections between neurons. The horizontal axis is the anterior-posterior (AP) axis and the vertical axis is the left-right (LR) axis of zebrafish (see Wan et al. 2019 for details). Preprocessing steps for obtaining these networks are available in Section B.5 in Appendix B. The colored dots are two neurons with similar connecting behaviour. Edges associated with the two nodes are displayed as dashed and solid line segments, respectively. The rest of edges are shown in gray. Panel (b) displays connecting behaviour of the two highlighted neurons. Vertical bars represent the time points when edges associated with the two neurons are formed. The color coding and line types are consistent with Panel (a). Panel (c) displays the adjusted connecting behaviour of the two highlighted neurons. Each vertical bar has been shifted according to the estimated time shift of the associated edge, as determined by the proposed method. Upon this adjustment, it becomes evident that the connecting behavior of the two neurons exhibits a high degree of similarity.

in this study is to cluster neurons based on their connecting behaviors after accounting for diverse activation times. The resulting clusters are expected to comprise neurons that play similar roles during neural circuit development.

Nonetheless, existing methods in the field of network analysis have limitations in identifying clusters of neurons when varying activation times are present. A canonical approach for clustering

nodes in networks is the *dynamic stochastic block models* [see, e.g., Xing et al., 2010, Yang et al., 2011, Xu and Hero, 2014, Xu, 2015, Matias and Miele, 2017, Matias et al., 2018, Longepierre and Matias, 2019, Pensky, 2019]. In essence, the dynamic stochastic block model assumes that nodes in a network fall into distinct clusters, and the probability of an edge forming between a pair of nodes over time, i.e., the *connecting intensity*, depends solely on the the cluster memberships of the two nodes. For networks observed at a discrete time grid, the connecting intensities are typically parameterized using Markov chains [see, e.g., Xing et al., 2010, Yang et al., 2011, Xu and Hero, 2014, Xu, 2015, Matias and Miele, 2017]. For networks observed in continuous time, Matias et al. [2018] propose to model recurrent instantaneous edges (e.g., email exchanges) as point processes, where connecting intensities are estimated nonparametrically. However, a crucial assumption in these models is that nodes within the same cluster share *identical* connecting intensities, which is essential for identifiability and estimation. Consequently, these models are unable to identify neural activation times and, as a result, may overlook the similarity in connecting behavior among neurons.

A class of problems relevant to identifying neural activation times has been studied in the context of *shape invariant models* [see e.g., Vimond, 2010a, Bigot and Gadat, 2010, Bigot et al., 2013, Bigot and Gendre, 2013, Bontemps and Gadat, 2014]. These models aim to align a set of curves that are identical up to unknown time shifts. However, the shape invariant models are incapable of handling multiple clusters of curves, where the curves within each cluster are identical up to unknown time shifts.

In this paper, we propose a dynamic stochastic block model that allows for unknown node-specific time shifts. We establish the identifiability of the time shifts, the cluster memberships, and the connecting intensities under suitable assumptions in Section 3.2. We propose semiparametric estimation procedures that simultaneously estimate model parameters including the time shifts in Section 3.3, along with computationally efficient algorithms in Section 3.4. We investigate the performance of the proposed estimation procedures through simulation experiments in Section 3.5. Finally, we use the proposed model to study neural behaviour during embryogenesis of zebrafish in Section 3.6.

3.2. Dynamic Stochastic Block Model

We model the development of neural circuits as a dynamic network. Specifically, a dynamic network is defined as $\{\mathcal{G}(t) : \mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t)), t \in [0, \infty)\}$, where $\mathcal{V} \equiv \{1, 2, \ldots, p\}$ denotes a fixed set of nodes across time, and $\mathcal{E}(t)$ represents the set of edges at time t, defined as

$$\mathcal{E}(t) \equiv \{(i, j) : \ i, j \in \mathcal{V}, \text{ nodes } \ i, j \text{ are connected at time } t\}.$$

In the context of neural circuits, $\mathcal{V}$ corresponds to the set of neurons, and $\mathcal{E}(t)$ denotes the set of functional connections between neurons at time t. The functional connections among neurons are identified by the highly correlated activities of neurons (see Section B.5.1). These functional connections have the following properties: (i) no self-connection, i.e., $(i, i) \notin \mathcal{E}(t)$ for any $i \in \mathcal{V}$ and $t \geq 0$, (ii) undirected connections, i.e., $(i, j) \in \mathcal{E}(t)$ if $(j, i) \in \mathcal{E}(t)$, (iii) no connection at the onset of observation, i.e., $\mathcal{E}(0) = \emptyset$, and (iv) persistence, i.e., $(i, j) \in \mathcal{E}(t')$ for any $t' \geq t$ if $(i, j) \in \mathcal{E}(t)$. Further exploration of the persistence of connections can be found in Figure B.10. For simplicity, we assume that the network has stabilized by time T, that is, $\mathcal{G}(t) = \mathcal{G}(T)$ for $t > T$. Hereafter, unless otherwise specified, we let (i, j) denote two arbitrary nodes that $i \neq j$ and $i, j \in \mathcal{V}$.

For any pair of nodes (i, j), we define the *connecting time* as $t_{i,j} \equiv \inf\{t \in [0, T] : (i, j) \in \mathcal{E}(t)\}$, that is, the time when nodes i and j connect, and we denote $t_{i,j} = \infty$ if $\{t \in [0, T] : (i, j) \in \mathcal{E}(t)\}$ is empty. We define the *connecting process* as $N_{i,j}(t) \equiv \mathbf{I}(t_{i,j} \leq t)$, $t \in [0, T]$, which is known as a point process or a counting process in the literature [see, among others, Daley and Vere-Jones, 2003]. The persistence of edges implies that $N_{i,j}(t)$ remains to be unity once an edge is formed. This type of point process is commonly studied in survival analysis as the outcome [see, among others, Kleinbaum and Klein, 2012]. We introduce the (marginal) intensity $\lambda_{i,j}(t) \equiv \Pr(\mathrm{d}N_{i,j}(t) = 1)/\mathrm{d}t$ that we refer to as *connecting intensity*, whose integral $\Lambda_{i,j}(T) \equiv \int_0^T \lambda_{i,j}(t)\mathrm{d}t$ is the *connecting probability* between nodes i, j throughout $[0, T]$. In survival analysis, the probability $\Lambda_{i,j}(T)$ can be less than one due to censoring. In our setting, we assume that there is no censoring but $\Lambda_{i,j}(T)$ can still be less than one since the final graph is not necessarily complete. Conditioning on that nodes i and j are connected, the density of the connecting time, or *connecting density*, is $\lambda_{i,j}(t)/\Lambda_{i,j}(T)$.

Following the convention in Matias et al. [2018] and Pensky [2019], we assume that $\mathcal{V} = \cup_{k=1}^{K} \mathcal{C}_k$, where $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_K$ are disjoint clusters with distinct connecting intensities to be specified in Assumption 3.2.1. Nodes within each cluster have identical connecting intensities up to unknown

time shifts. To be specific, denote $[K] \equiv \{1, \ldots, K\}$ and let $z_i \in [K]$ be the *cluster membership* of node i for $i \in \mathcal{V}$. Then we assume that the connecting intensity between nodes i and j satisfies that

$$\lambda_{i,j}(t) = S^{w_{i,j}} f_{z_i,z_j}(t), \tag{3.1}$$

where $w_{i,j} \equiv \max(v_i, v_j)$ with $v_i, v_j \in [0, T/2]$ being the *time shifts* of nodes i and j, S^w is the shift operator defined as $S^w x(t) \equiv x(t-w)$, and $f_{z_i,z_j}(\cdot)$ is a non-negative function with support in $[0, T/2]$ and $\|f_{z_i,z_j}\|_1 \leq 1$. For $q, k \in [K]$, the function $f_{q,k}(\cdot)$ is called the *representative connecting intensity* between any pairs of nodes in clusters q and k. Figure 3.2 shows a graphical representation of the proposed dynamic stochastic block model with time shifts. Here the bounded supports for the representative connecting intensities and time shifts are put in place to avoid the discussion of censoring, which can be easily relaxed.

The model presented in (3.1) is tailored to the neural circuit development data. Specifically, $w_{i,j}$ is defined as the time when both neurons i and j are mature enough to establish functional connections. As a result, the proposed model can be applied to identify the clusters of neurons with similar connecting behaviour (i.e., $\mathcal{C}_k$'s), the representative connecting behaviour of neurons (i.e., $f_{q,k}$'s), and activation time of neurons (i.e., v_i's). Nonetheless, the model in (3.1) can be easily adapted to other applications with minor changes. For instance, it can be used to analyze information exchanges such as tweets, emails, or text messages in social networks [see, e.g., Haythornthwaite, 1996, Diesner and Carley, 2005, Osatuyi, 2013], where nodes represent individuals and edges represent the initiation of information exchange between individuals on a particular topic. In this context, $w_{i,j}$ can be redefined to represent the minimal time shifts of individuals i and j, signifying the time when either individual i or j becomes intrigued by the topic.

The proposed model has close connections with two lines of research in the literature. First of all, the proposed model generalizes the dynamic stochastic block model in Matias et al. [2018] by allowing unknown time shifts in connecting intensities between pairs of nodes. In fact, if we average out time shifts to treat $\mathbb{E}_{w_{i,j}}[\lambda_{i,j}(\cdot)]$'s as the connecting intensities, the proposed model reduces to the Poisson process stochastic block model (PPSBM) in Matias et al. [2018]. Secondly, when there is only one cluster, i.e., $K = 1$, the proposed model reduces to a shape invariant model with point processes [Bigot et al., 2013], where the point processes are assumed to share a common intensity

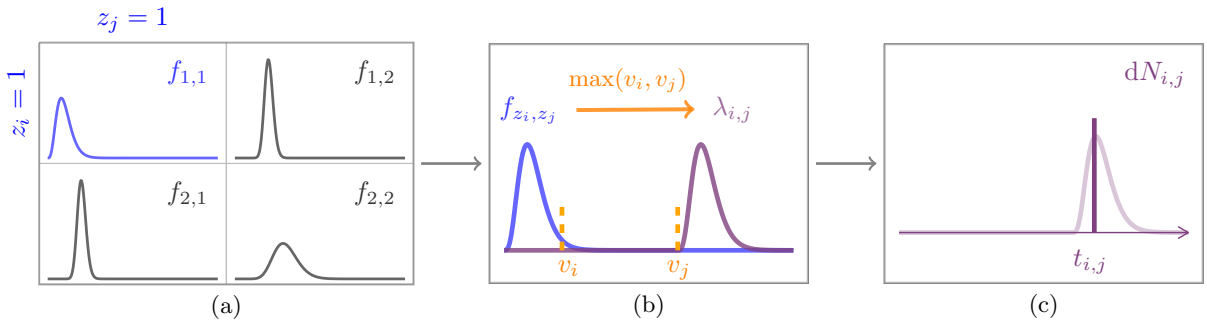

FIGURE 3.2. Graphical representation of our model. Panel (a) shows a matrix of representative connecting intensities $(f_{q,k})_{K\times K}$, where the highlighted curve represents f_{z_i,z_j}. Panel (b) shows the connecting intensity between node i and j after incorporating the time shift, where $\lambda_{i,j} = S^{\max(v_i,v_j)} f_{z_i,z_j}$. Panel (c) shows the observed connecting process where $\mathbb{E}[\mathrm{d}N_{i,j}(t)] = \lambda_{i,j}(t)\mathrm{d}t$. Only the connecting process is observed in the data.

function subject to time shifts. Hence, we call the proposed model as the *shape invariant dynamic stochastic block model*, or SidSBM for short.

Additional conditions are required to ensure the identifiability of parameters in (3.1). In the case without unknown time shifts, Matias et al. [2018] establish the identifiability of representative connecting intensities and cluster memberships, under the assumption that all representative connecting intensities are distinct. In the case with one single cluster, Bigot et al. [2013] ensure the identifiability of the common intensity function by assuming a known distribution of time shifts. However, in our setting, we need to simultaneously identify representative connecting intensities, cluster memberships, and time shifts. Therefore, additional conditions are required for identifiability as summarized in Assumption 3.2.1 and Proposition 3.2.1.

ASSUMPTION 3.2.1. *Denote true parameters in* (3.1) *by* $\{v_i^* : i \in \mathcal{V}\}$, $\{f_{q,k}^* : q, k \in [K]\}$ *and* $\{z_i^* : i \in \mathcal{V}\}$. *Let* $\mathcal{C}_q^* = \{i \in \mathcal{V} : z_i^* = q\}$ *for* $q \in [K]$. *We assume the true parameters satisfy the following conditions.*

(C3.1) For any two distinct clusters $q, q' \in [K], q \neq q'$, *there exists at least one cluster* $k \in [K]$ *such that* $f_{q,k}^* \neq S^w f_{q',k}^*$ *for any* $w \in \mathbb{R}$.

(C3.2) For any $q, k \in [K]$, *if* $f_{q,k}^* \not\equiv 0$, *then* $\max\{v_i^* : i \in \mathcal{C}_q^*\} > \min\{v_j^* : j \in \mathcal{C}_k^*\}$ *and* $\min\{v_i^* : i \in \mathcal{C}_q^*\} < \max\{v_j^* : j \in \mathcal{C}_k^*\}$.

(C3.3) For any $q, k \in [K]$, *there exists a sequence of clusters* $(q_1, q_2, \cdots, q_a) \in [K]^a$ *such that* $q_1 = q$ *and* $q_a = k$ *and* $f_{q_i,q_{i+1}}^* \not\equiv 0$ *for* $i = 1, \cdots, a-1$, *where* $a \in \mathbb{N}$, $a \geq 2$.

(C3.4) There exist $q_0, k_0 \in [K]$ *such that* $\inf\{t \in [0,T] : f^*_{q_0,k_0}(t) > 0\} = 0$.

PROPOSITION 3.2.1. *Suppose that we observe connecting processes* $\{N_{i,j} : i,j \in \mathcal{V}\}$ *and that Assumption 3.2.1 holds. Then, we can verify the following statements.*

(P3.1) The cluster memberships $\{z^*_i : i \in \mathcal{V}\}$ *and representative connecting intensities* $\{f^*_{q,k} : q,k \in [K]\}$ *are identifiable up to a permutation of cluster labels.*

(P3.2) The time shifts $\{v^*_i : i \in \mathcal{V}\}$ *are identifiable except for nodes in the set*

$$\{i \in \mathcal{V} : v^*_i \le v^*_j \text{ for any } j \in \mathcal{C}^*_{z^*_i}, \text{ or } v^*_i \le v^*_j \text{ for any } j \in \mathcal{V} \text{ s.t. } \lambda^*_{i,j} \not\equiv 0\}. \tag{3.2}$$

Assumption 3.2.1 lists the technical conditions that ensure the identifiability of the model. Here we briefly explain how conditions in Assumption 3.2.1 lead to the identifiability in Proposition 3.2.1, and we defer the detailed proof to Section B.1 in Appendix B. Firstly, Condition C3.1 assumes that any two different clusters are distinct in at least one of their connecting intensities with other clusters, which ensures the identifiability of $\{z^*_i : i \in \mathcal{V}\}$, up to a permutation of cluster labels. Secondly, Condition C3.2 requires that the ranges of time shifts overlap for any two clusters with a positive connecting probability. As a result, for any cluster $q \in [K]$, its non-zero connecting intensities $\{f^*_{q,k} : f^*_{q,k} \not\equiv 0, k \in [K]\}$ are identifiable up to a shared time shift. Moreover, Condition C3.3 grants that for any pair of nodes, there is a positive probability that these two nodes are directly or indirectly connected. In other words, the network is not separable. Consequently, we can show that the representative connecting intensities of all clusters are identifiable up to a global time shift. Thirdly, Condition C3.4 pins down the global time shift in a similar way as the zero-sum constraint in Bigot and Gendre [2013]. Lastly, the definition of $w_{i,j}$ in (3.1) implies that there is a set of nodes whose time shifts can not be identified. To see this, consider the case with one cluster. The unidentifiable set (3.2) reduces to the node with the minimum time shift, which is unidentifiable since the minimum time shift is always masked by other time shifts. We verify in Section 3.6 that the conditions in Assumption 3.2.1 are feasible in real data.

3.3. Estimation Procedure

The estimation of stochastic block models can be performed based on the squared error [Gao et al., 2015, Pensky, 2019, Lei et al., 2020] or the likelihood [Matias et al., 2018, Pavlović et al., 2020]. We opt to extend the squared-error-based estimation approach to accommodate unknown

time shifts, as it allows us to leverage techniques employed in shape invariant models [Kneip and Engel, 1995, Bigot and Gendre, 2013] for fast computing.

In what follows, we represent vectors and matrices in boldface, e.g., $\mathbf{f} \equiv (f_{q,k})_{q,k\in[K]}$, $\mathbf{z} \equiv (z_i)_{i\in\mathcal{V}}$, $\mathbf{v} \equiv (v_i)_{i\in\mathcal{V}}$, and let $L^2_+([0,T])$ denote the space of non-negative squared integrable functions on $[0,T]$. Recall that in our model, we only observe the connecting processes $\{N_{i,j} : i,j \in \mathcal{V}\}$. The optimization problem associated with our model takes the following form

$$(\hat{\mathbf{f}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}) = \underset{\Theta_1}{\arg\min} \sum_{i,j\in\mathcal{V}, i\neq j} \left\{ \int_0^T S^{w_{i,j}} f^2_{z_i,z_j}(t)\mathrm{d}t - 2\int_0^T S^{w_{i,j}} f_{z_i,z_j}(t)\mathrm{d}N_{i,j}(t) \right\}, \tag{3.3}$$

where $w_{i,j} = \max(v_i, v_j)$, and $\Theta_1 \equiv \{(\mathbf{f}, \mathbf{z}, \mathbf{v}) : \mathbf{f} \in \mathcal{B}_1^{K\times K}, \mathbf{z} \in [K]^p, \mathbf{v} \in [0,T]^p\}$ with $\mathcal{B}_1 \equiv \{f \in L^2_+([0,T]) : \|f\|_1 \leq 1\}$. The objective function in (3.3) is identical to the sum of squared deviations of the observed connecting processes from their intensities up to a constant. To estimate $\mathbf{f}$ non-parametrically, we employ a truncated basis expansion. In particular, we consider the functional space with truncated Fourier bases $\mathcal{B}_{1,\ell_0} \equiv \{f \in \mathcal{B}_1 : \int_0^T \exp\{-\mathbf{i}2\pi\ell t/T\} f(t)\mathrm{d}t = 0 \text{ for } |\ell| \geq \ell_0\}$. The selection of the tuning parameter ℓ_0 is discussed later in this section.

We notice that replacing the intensities $\{f_{q,k} : q,k \in [K]\}$ with the cumulative intensities $\{F_{q,k} : F_{q,k}(t) \equiv \int_0^t f_{q,k}(s)\mathrm{d}s, t \in [0,T], q,k \in [K]\}$ can avoid the frequency truncation. In particular, the optimization problem based on cumulative intensities takes the following form

$$(\hat{\mathbf{F}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}) = \underset{\Theta_2}{\arg\min} \sum_{i,j\in\mathcal{V}, i\neq j} \int_0^T \left\{ S^{w_{i,j}} F^2_{z_i,z_j}(t) - 2S^{w_{i,j}} F_{z_i,z_j}(t) N_{i,j}(t) \right\} \mathrm{d}t, \tag{3.4}$$

where $w_{i,j} = \max(v_i, v_j)$, and $\Theta_2 \equiv \{(\mathbf{F}, \mathbf{z}, \mathbf{v}) : \mathbf{F} \in \mathcal{B}_2^{K\times K}, \mathbf{z} \in [K]^p, \mathbf{v} \in [0,T]^p\}$ with $\mathcal{B}_2 \equiv \{F \in [0,1]^{[0,T]} : F \text{ is non-decreasing and right-continuous}, F(0) = 0\}$. The objective function in (3.4) is almost identical to the one in (3.3) except for the use of $N_{i,j}$ and F_{z_i,z_j}. In (3.4), a natural estimator of $\mathbf{F}$ is the empirical cumulative intensity which does not require frequency truncation.

However, the objective functions based on the plain vanilla L^2-distance in (3.3) and (3.4) may tend to overlook the information in the *shapes* of intensities. To see this, consider a simple case where we assign one node to one of K clusters with known intensities and no time shifts. In this case, problem (3.4) reduces to

$$\hat{z}_1 = \underset{z_1\in[K]}{\arg\min} \sum_{k\in[K]} |\mathcal{C}^*_k| \int_0^T \left| F^*_{z_1,k}(t) - \overline{N}_{1,k}(t) \right|^2 \mathrm{d}t, \tag{3.5}$$

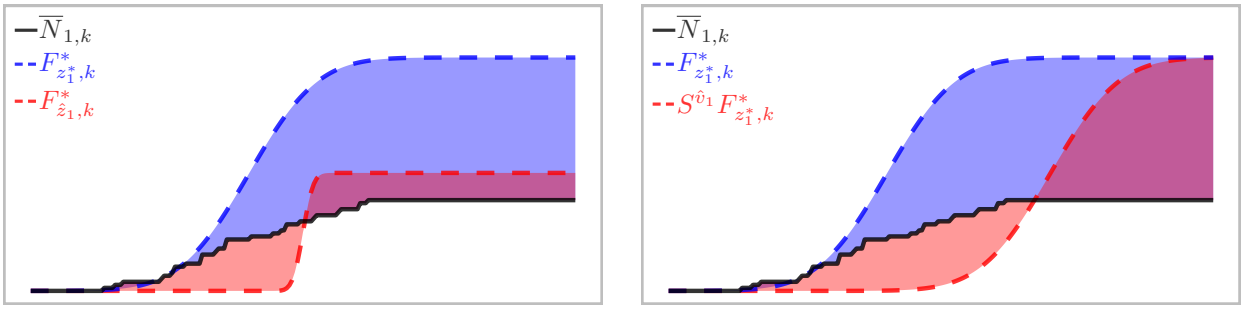

FIGURE 3.3. Deviation of the empirical cumulative intensity from the true cumulative intensity and from the minimizer-induced cumulative intensities. Panel (a) shows that the objective function may select wrong cluster $\hat{z}_1$ due to the similarity in scales, despite the difference in shapes. Panel (b) shows that $\hat{v}_1$ which minimizes the objective function in (3.4) may shift the cumulative intensity away from the truth.

where $\{\mathcal{C}_k^* : k \in [K]\}$ are known clusters for nodes in $\mathcal{V}\backslash\{1\}$, $\{F^*_{z_1,k} : z_1, k \in [K]\}$ are known cumulative intensities across clusters, and $\{\overline{N}_{1,k} : \overline{N}_{1,k} \equiv |\mathcal{C}_k^*|^{-1}\sum_{j\in\mathcal{C}_k^*} N_{1,j}, k \in [K]\}$ are empirical cumulative intensities. The cumulative intensity, e.g., $F^*_{z_1,k}(t)$, can be broken down into two parts: the cumulative distribution $F^*_{z_1,k}(t)/F^*_{z_1,k}(T)$ and the connecting probability $F^*_{z_1,k}(T)$. In a sense, the two parts can be seen as the shape and the scale of the cumulative intensity. As shown in Figure 3.3(a), the minimizer $\hat{z}_1$ of the objective function in (3.5) might be predominantly decided by scales. Moreover, the problem is exacerbated when time shifts are unknown as in (3.4). Figure 3.3(b) shows that the minimum might be achieved by over estimating the time shifts.

To address this issue, we propose a new loss function that separates the shape and scale. We modify (3.4) as

$$
\begin{aligned}
\left(\hat{\mathbf{F}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}\right) &= \arg\min_{\Theta_C} M_C(\mathbf{F}, \mathbf{z}, \mathbf{v}; \gamma) \\
&\equiv \arg\min_{\Theta_C} \Bigg[\sum_{(i,j)\in\mathcal{E}(T)} \frac{1}{T}\int_0^T \left\{ \left| \frac{S^{w_{i,j}} F_{z_i,z_j}(t)}{F_{z_i,z_j}(T)} \right|^2 - 2N_{i,j}(t) \frac{S^{w_{i,j}} F_{z_i,z_j}(t)}{F_{z_i,z_j}(T)} \right\} \mathrm{d}t \\
&\qquad + \gamma \sum_{i,j\in\mathcal{V}, i\neq j} \left| N_{i,j}(T) - F_{z_i,z_j}(T) \right|^2 \Bigg],
\end{aligned} \tag{3.6}
$$

where $\gamma \geq 0$ is a user-specified parameter, $\mathcal{E}(T) = \{(i,j) : i,j \in \mathcal{V}, N_{i,j}(T) = 1\}$, $w_{i,j} = \max(v_i, v_j)$, and $\Theta_C \equiv \{(\mathbf{F}, \mathbf{z}, \mathbf{v}) : \mathbf{F} \in \mathcal{B}_2^{K\times K}, \mathbf{z} \in [K]^p, \mathbf{v} \in [0,T]^p, F_{z_i,z_j}(T) > 0 \text{ if } (i,j) \in \mathcal{E}(T)\}$. The objective

function in (3.6) has two components. The first component is the sum of squared deviations related to cumulative connecting distributions (i.e., shapes), and the second component is the sum of squared deviations related to connecting probabilities (i.e., scales). The relative weight of shapes and scales is governed by γ. The selection of γ is discussed later in this section. The estimator in (3.6) is based on the cumulative connecting intensities, thus we refer to it as *SidSBM-C*.

In a similar manner, we introduce *SidSBM-P*, an intensity-based estimator that solves the following optimization problem

$$
\begin{aligned}
\big(\hat{\mathbf{f}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}\big) &= \arg\min_{\Theta_{\mathrm{P}}} \; M_{\mathrm{P}}(\mathbf{f}, \mathbf{z}, \mathbf{v}; \gamma) \\
&\equiv \arg\min_{\Theta_{\mathrm{P}}} \Bigg[\sum_{(i,j)\in\mathcal{E}(T)} \frac{1}{T} \Bigg\{ \int_0^T \bigg| \frac{S^{w_{i,j}} f_{z_i,z_j}(t)}{F_{z_i,z_j}(T)} \bigg|^2 \mathrm{d}t - 2 \int_0^T \frac{S^{w_{i,j}} f_{z_i,z_j}(t)}{F_{z_i,z_j}(T)} \mathrm{d}N_{i,j}(t) \Bigg\} \\
&\qquad\qquad + \gamma \sum_{i,j\in\mathcal{V}, i\neq j} \big| N_{i,j}(T) - F_{z_i,z_j}(T) \big|^2 \Bigg],
\end{aligned} \tag{3.7}
$$

where γ, $\mathcal{E}(T)$, $w_{i,j}$ are defined the same way as in (3.6), $F_{z_i,z_j}(T) = \int_0^T f_{z_i,z_j}(s)\mathrm{d}s$, and $\Theta_{\mathrm{P}} \equiv \big\{(\mathbf{f}, \mathbf{z}, \mathbf{v}) : \mathbf{f} \in \mathcal{B}_{1,\ell_0}^{K\times K}, \mathbf{z} \in [K]^p, \mathbf{v} \in [0,T]^p, f_{z_i,z_j} \not\equiv 0 \text{ if } (i,j) \in \mathcal{E}(T)\big\}$.

We adapt the *integrated classification likelihood* (ICL) criterion [Biernacki et al., 2000, Daudin et al., 2008] to select tuning parameters including the number of clusters K, the relative importance of intensity scales γ, and the frequency truncation parameter ℓ_0. Integrated classification likelihood is based on a BIC-like approximation that has been adapted to semiparametric stochastic block models by Matias et al. [2018]. For our model, the integrated classification likelihood criterion takes the following form

$$
\begin{aligned}
\mathrm{ICL} = &\sum_{(i,j)\in\mathcal{E}_0} \log\Big\{1 - \hat{F}_{\hat{z}_i,\hat{z}_j}(T)\Big\} + \sum_{(i,j)\in\mathcal{E}_1} \log\Big\{\hat{f}_{\hat{z}_i,\hat{z}_j}(t_{i,j})\Big\} + \sum_{k\in[K]} \big|\hat{\mathcal{C}}_k\big| \log\Big(p^{-1}\big|\hat{\mathcal{C}}_k\big|\Big) \\
&- \frac{1}{2}(K-1)\log p - \frac{1}{4}K(K+1)\ell_0 \log\frac{p(p-1)}{2},
\end{aligned} \tag{3.8}
$$

where $\mathcal{E}_0 \equiv \{(i,j) : i,j \in \mathcal{V}, i < j, N_{i,j}(T) = 0\}$ and $\mathcal{E}_1 \equiv \{(i,j) : i,j \in \mathcal{V}, i < j, N_{i,j}(T) = 1\}$. The sum of the first three terms can be seen as the complete-data log-likelihood, and the last two terms are penalties for cluster memberships and connecting intensities. In the second penalty term, we use ℓ_0 to represent the degrees of freedom in each connecting intensity. For optimization problem (3.6), we select K and γ by maximizing ICL with $\ell_0 = 2$, where the two degrees of freedom

correspond to the shape and the scale. For optimization problem (3.7), we select K, γ, and ℓ_0 by maximizing ICL. In practice, heuristic methods can be employed for computationally efficient selection of tuning parameters. Further details are provided in Section B.4.1 in Appendix B.

REMARK 3.3.1. *We can generalize the proposal to networks with recurrent edges, where the edges are events that can happen more than one time. Instances of such networks include, but are not limited to, email exchange networks, payment transaction networks, and bike sharing networks [see, e.g., Diesner and Carley, 2005, Loupos et al., 2019, Lin et al., 2018]. When the connecting intensities* $\mathbf{f}$ *have bounded supports, the optimization problems* (3.6) *and* (3.7) *can be directly applied without modification, and the identifiability results in Section 3.2 hold. When* $\mathbf{f}$ *have unbounded supports, identifiability requires additional assumptions on the intensities* $\mathbf{f}$*. For instance, we can assume that intensities in* $\mathbf{f}$ *are periodic and a full period is observed for each intensity in* $\mathbf{f}$ *[see, e.g., Bigot et al., 2013, Bigot and Gendre, 2013]. Furthermore, the optimization problem* (3.7) *can be applied directly, when the intensity is periodic.*

3.4. Algorithm

We now present an algorithm for solving the optimization problem (3.6). The key observation in (3.6) is that it is straight-forward to solve one set of parameters conditioning on the others. The resulting algorithm resembles the classic k-means algorithm that iterates between the centering and clustering steps. In particular, we can write the two steps as follows

$$\text{(centering step)} \quad \hat{\mathbf{F}}, \hat{\mathbf{v}} = \arg\min_{\mathbf{F},\mathbf{v}} M_{\mathrm{C}}(\mathbf{F}, \hat{\mathbf{z}}, \mathbf{v}; \gamma), \tag{3.9}$$

$$\text{(clustering step)} \quad \hat{\mathbf{z}} = \arg\min_{\mathbf{z}} M_{\mathrm{C}}(\hat{\mathbf{F}}, \mathbf{z}, \hat{\mathbf{v}}; \gamma). \tag{3.10}$$

In the centering step (3.9), we update $\hat{\mathbf{F}}$ and $\hat{\mathbf{v}}$ conditioning on the current estimates of cluster memberships $\hat{\mathbf{z}}$. Given $\hat{\mathbf{z}}$, the optimization problem (3.6) reduces to a well-studied problem in the literature of shape invariant models [Bigot and Gadat, 2010, Bigot et al., 2013, Bigot and Gendre, 2013]. The reduced optimization problem can be decoupled into two parts concerning $\hat{\mathbf{F}}$ and $\hat{\mathbf{v}}$, respectively. In the first part, we have the analytical solution of $\hat{F}_{q,k}$ as

$$\hat{F}_{q,k} = \sum_{i\in\hat{\mathcal{C}}_q, j\in\hat{\mathcal{C}}_k, i\neq j} \frac{S^{-\max(\hat{v}_i,\hat{v}_j)} N_{i,j}}{\big|\hat{\mathcal{C}}_q\big|\big|\hat{\mathcal{C}}_k\big| - \mathbf{I}(q=k)\big|\hat{\mathcal{C}}_q\big|}, \quad q,k \in [K], \tag{3.11}$$

where $\hat{\mathcal{C}}_k \equiv \{i \in \mathcal{V} : \hat{z}_i = k\}$. The estimator in (3.11) is the empirical cumulative intensity of edges between clusters q and k, where self-edges are eliminated if $q = k$.

In the second part, we solve for $\hat{\mathbf{v}}$ via

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}} \sum_{i \in \mathcal{V}} \sum_{k \in [K]} \left\{ \left| \hat{J}_{i,k} \right| \overline{N}_{i,k}(T) \frac{1}{T} \int_0^T \left| \frac{\overline{N}_{i,k}(t)}{\overline{N}_{i,k}(T)} - \frac{S^{v_i} \hat{F}_{\hat{z}_i,k}(t)}{\hat{F}_{\hat{z}_i,k}(T)} \right|^2 \mathrm{d}t \right\}, \tag{3.12}$$

where $\hat{J}_{i,k} \equiv \{j \in \mathcal{V} : \hat{z}_j = k, j \neq i, v_j \leq v_i\}$ is the collection of nodes in the k-th cluster whose time shifts are no larger than node i, and $\overline{N}_{i,k}(t) \equiv |\hat{J}_{i,k}|^{-1} \sum_{j \in \hat{J}_{i,k}} N_{i,j}(t)$ is the empirical cumulative intensity of edges between node i and nodes in $\hat{J}_{i,k}$. The optimization problem in (3.12) can be solved by utilizing the well-known Parseval's theorem to reframe the problem in the Frequency domain. The Frequency domain offers a suitable environment for the application of the gradient descent algorithm, which provides a solution to the problem. Details of the gradients are available in Section B.3 in Appendix B.

Finally, we set the global shift as $\hat{c} \equiv \min\{t_{i,j} - \max(\hat{v}_i, \hat{v}_j) : i, j \in \mathcal{V}\}$, which is an approximation of $\inf\{t \in [0, T] : \mathrm{d}\hat{F}_{q,k}(t) > 0 \text{ for some } q, k \in [K]\}$. Specifically, we update $\hat{\mathbf{F}}$ and $\hat{\mathbf{v}}$ as follows

$$\hat{F}_{q,k} \leftarrow S^{-\hat{c}} \hat{F}_{q,k}, \ q, k \in [K]; \quad \hat{v}_i \leftarrow \hat{v}_i + \hat{c}, \ i \in \mathcal{V}. \tag{3.13}$$

The global shift $\hat{c}$ is the minimum of p^2 numbers. As such, for small values of p, $\hat{c}$ might be subject to a large error, which might further compromise the accuracy of $\hat{\mathbf{F}}$ and $\hat{\mathbf{v}}$.

In the clustering step (3.10), we update each cluster membership $\hat{z}_i$ conditioning on current $\hat{\mathbf{F}}$, $\hat{\mathbf{v}}$ and $\{\hat{z}_j : j \in \mathcal{V}, j \neq i\}$. The optimization problem in (3.6) reduces to, for $i \in \mathcal{V}$,

$$\begin{aligned} \hat{z}_i = \arg\min_{z_i \in [K]} \sum_{k \in [K]} \Bigg\{ & \left| \hat{\mathcal{C}}_{k,-i} \right| \overline{F}_{i,k}(T) \frac{1}{T} \int_0^T \left| \frac{\overline{F}_{i,k}(t)}{\overline{F}_{i,k}(T)} - \frac{\hat{F}_{z_i,k}(t)}{\hat{F}_{z_i,k}(T)} \right|^2 \mathrm{d}t \\ & + \gamma \left| \hat{\mathcal{C}}_{k,-i} \right| \left| \overline{F}_{i,k}(T) - \hat{F}_{z_i,k}(T) \right|^2 \Bigg\}, \end{aligned} \tag{3.14}$$

where $\hat{\mathcal{C}}_{k,-i} \equiv \hat{\mathcal{C}}_k \backslash \{i\}$ and $\overline{F}_{i,k}(\cdot) \equiv \left| \hat{\mathcal{C}}_{k,-i} \right|^{-1} \sum_{j \in \hat{\mathcal{C}}_{k,-i}} S^{-\max(\hat{v}_i, \hat{v}_j)} N_{i,j}(\cdot)$. The objective function in (3.14) can be seen as a "distance" between node i and cluster $\hat{\mathcal{C}}_{z_i}$ that involves differences in both shapes and scales.

The optimization problem in (3.6) is non-convex, hence the initialization scheme and the convergence criterion are crucial for the performance of the algorithm (see Remarks 3.4.1 and 3.4.2). The estimation procedure is summarized in Algorithm 3.1.

Algorithm 3.1: Iterative algorithm for SidSBM-C

Input: $\mathbf{N}, K, \varepsilon, \gamma$

Initialize $\hat{\mathbf{v}}^{(0)}, \hat{\mathbf{z}}^{(0)}$ via (3.15) (3.16), and $s = 0$;

while *not converge* **do**

Update $\hat{\mathbf{v}}^{(s+1)}$ and $\hat{\mathbf{F}}^{(s+1)}$ via (3.11) - (3.13) with $(\mathbf{z}, \mathbf{v}) = (\hat{\mathbf{z}}^{(s)}, \hat{\mathbf{v}}^{(s)})$;

Update $\hat{\mathbf{z}}^{(s+1)}$ via (3.14) with $(\mathbf{v}, \mathbf{F}, \mathbf{z}) = (\hat{\mathbf{v}}^{(s+1)}, \hat{\mathbf{F}}^{(s+1)}, \hat{\mathbf{z}}^{(s)})$;

Evaluate the stopping criterion via (3.18);

$s = s + 1$;

end

Output: $\hat{\mathbf{F}}^{(s)}, \hat{\mathbf{z}}^{(s)}, \hat{\mathbf{v}}^{(s)}$.

The algorithm to solve the optimization problem (3.7) closely resembles Algorithm 3.1, which is detailed in Section B.2. The objective function in (3.7) might have plateaus in v, while the one in (3.6) does not encounter this issue due to the utilization of cumulative intensities (see Figure B.1).

REMARK 3.4.1. *We initialize time shifts by the time of the first edge associated with each node, that is*

$$\hat{v}_i^{(0)} \equiv \min\{t_{i,j} : j \in \mathcal{V}\}, \quad i \in \mathcal{V}. \tag{3.15}$$

Letting $F_i(t \mid \mathbf{v}) \equiv (p-1)^{-1} \sum_{j \in \mathcal{V} \setminus \{i\}} S^{-\max(v_i, v_j)} N_{i,j}(t)$, *we initialize cluster memberships by solving the following k-medoids problem*

$$\hat{\mathbf{z}}^{(0)}, \hat{\boldsymbol{\mu}}^{(0)} = \underset{\mathbf{z} \in [K]^p, \boldsymbol{\mu} \in \mathcal{C}_1 \times \ldots \times \mathcal{C}_K}{\arg\min} \sum_{i \in \mathcal{V}} \int_0^T \left| \frac{F_i(t \mid \hat{\mathbf{v}}^{(0)})}{F_i(T \mid \hat{\mathbf{v}}^{(0)})} - \frac{F_{\mu_{z_i}}(t \mid \hat{\mathbf{v}}^{(0)})}{F_{\mu_{z_i}}(T \mid \hat{\mathbf{v}}^{(0)})} \right|^2 \mathrm{d}t, \tag{3.16}$$

where $\{\mu_q : q \in [K]\}$ *are the medoids of the* K *clusters. Optimization problem* (3.16) *is a simplification of problem* (3.6), *where* $\mathbf{v}$ *is given as* $\hat{\mathbf{v}}^{(0)}$, $\{F_{q,k}(t) : q, k \in [K]\}$ *take values in the known set* $\{F_i(t \mid \hat{\mathbf{v}}^{(0)}) : i \in \mathcal{V}\}$, *and* γ *is set as zero. In practice,* $F_i(t \mid \hat{\mathbf{v}}^{(0)})$ *is evaluated on an equispaced time grid as an* L*-vector*

$$\frac{1}{p-1} \sum_{j \in \mathcal{V} \setminus \{i\}} \left(\mathbf{I}\left[\tilde{t}_{i,j} \le \frac{T}{L}\right], \mathbf{I}\left[\tilde{t}_{i,j} \le \frac{2T}{L}\right], \cdots, \mathbf{I}\left[\tilde{t}_{i,j} \le T\right] \right)^\top, \tag{3.17}$$

where $\tilde{t}_{i,j} \equiv t_{i,j} - \max(\hat{v}_i^{(0)}, \hat{v}_j^{(0)})$ *is the adjusted connecting time. The tuning parameter* L *can be chosen such that* T/L *is close to temporal resolution of the data. We note that the proposed initialization scheme can not guarantee a global minimum. If feasible, we recommend to use multiple random restarts by jittering the initial time shifts and shuffling the initial cluster memberships for a small proportion of nodes.*

REMARK 3.4.2. *The proposed algorithm is stopped when, for a user-specified* ε*,*

$$\left[\left\{1 - \mathrm{ARI}(\hat{\mathbf{z}}^{(s)}, \hat{\mathbf{z}}^{(s+1)})\right\} + \frac{\|\hat{\mathbf{v}}^{(s)} - \hat{\mathbf{v}}^{(s+1)}\|_2}{\|\hat{\mathbf{v}}^{(s)}\|_2} + \frac{\|\hat{\mathbf{F}}^{(s)} - \hat{\mathbf{F}}^{(s+1)}\|_2}{\|\hat{\mathbf{F}}^{(s)}\|_2}\right] \leq \varepsilon, \tag{3.18}$$

where $\|\cdot\|_2$ *is the* L^2*-norm and* $\hat{\mathbf{z}}^{(s)}, \hat{\mathbf{v}}^{(s)}, \hat{\mathbf{F}}^{(s)}$ *are estimates in* s*-th iteration. This criterion measures the overall change of model parameters between two consecutive iterations. We use the change of model parameters rather than the change of loss function because the stabilization of model parameters can ensure the stabilization of loss function. The change of cluster memberships is measured using the Adjusted Rand Index* (ARI) *[Hubert and Arabie, 1985]. In particular, for two partitions* $\mathcal{C} \equiv \{\mathcal{C}_q : q \in [K]\}$ *and* $\mathcal{C}' \equiv \{\mathcal{C}'_k : k \in [K']\}$ *such that* $\mathcal{V} = \cup_{q\in[K]}\mathcal{C}_q = \cup_{k\in[K']}\mathcal{C}'_k$*, denoting* $a_q \equiv |\mathcal{C}_q|$*,* $b_k \equiv |\mathcal{C}'_k|$*, and* $d_{qk} \equiv |\mathcal{C}_q \cap \mathcal{C}'_k|$ *for* $q \in [K]$ *and* $k \in [K']$*, then*

$$\mathrm{ARI} \equiv \frac{\sum_{q,k}\binom{d_{qk}}{2} - \left[\sum_q \binom{a_q}{2}\sum_k \binom{b_k}{2}\right]/\binom{p}{2}}{\frac{1}{2}\left[\sum_q \binom{a_q}{2} + \sum_k \binom{b_k}{2}\right] - \left[\sum_q \binom{a_q}{2}\sum_k \binom{b_k}{2}\right]/\binom{p}{2}}. \tag{3.19}$$

If the two partitions are identical up to a permutation of cluster labels, then ARI *equals one. If one partition is purely random and the other one is fixed, then the expectation of* ARI *is zero.*

3.5. Simulation

3.5.1. Data generation and model specification. Random dynamic networks are generated from the proposed model. In particular, cluster memberships are generated by sequentially

TABLE 3.1. Mean (μ) and variance (σ^2) of the connecting intensities for generating synthetic data. Each column corresponds to one pair of (q, k). The value of β is specified in Section 3.5.2.

	$(1,1)$	$(1,2)$	$(1,3)$	$(2,2)$	$(2,3)$	$(3,3)$
$\mu_{q,k}$	20	20β	$20\beta^2$	$20\beta^2$	$20\beta^{1/2}$	$20\beta^{3/2}$
$\sigma^2_{q,k}$	100	$100\beta^{-2}$	$100\beta^{-2}$	100β	$100\beta^{-1}$	100β

allocating p nodes into $K = 3$ clusters of equal size, that is, $\mathcal{C}_q = \{(p/K)(q-1)+1, \ldots, (p/K)q\}$ for $q \in [K]$. For any $q, k \in [K]$, the representative connecting intensity between clusters q and k takes the form $f_{q,k} = 0.9 \times \Gamma(\mu_{q,k}^2/\sigma_{q,k}^2, \mu_{q,k}/\sigma_{q,k}^2)$, where $\Gamma(\mu_{q,k}^2/\sigma_{q,k}^2, \mu_{q,k}/\sigma_{q,k}^2)$ is the density of a Gamma distribution with mean $\mu_{q,k}$ and variance $\sigma_{q,k}^2$. Specifications of $\mu_{q,k}$'s and $\sigma_{q,k}^2$'s are provided in Table 3.1, where we introduce β to control the separability between representative connecting intensities. Time shifts are generated from a uniform distribution that $v_i \sim U(0, W), i \in \mathcal{V}$. The connecting time between any nodes i and j is generated as $t_{i,j} = X_{i,j} + \max(v_i, v_j)$, where $X_{i,j}$ is drawn from f_{z_i,z_j} for any pair (i, j).

By altering p, β, and W, we generate data with varying signal strengths. Intuitively, the signal strength is positively associated with p and β, since increasing p increases the sample size, and increasing β increases distinction among $(\mu_{q,k}, \sigma_{q,k}^2)$'s; the signal strength might decrease when W increases as W controls the variation in time shifts.

We compare the performance of three methods, SidSBM-C (3.6), SidSBM-P (3.7), and the Poisson process stochastic block model (PPSBM) by Matias et al. [2018]. We apply SidSBM-C with $K = 3$, $\epsilon = 0.01$, $\gamma = 0.01$ and $L = 200$. The same set of tuning parameters is used for SidSBM-P except for $\gamma = 0.0001$ and $\ell_0 = 4$. We apply PPSBM using the implementation in `R` by Giorgi et al. [2018] with the number of bins being 64.

3.5.2. Simulation results. We use two criteria to evaluate the performance of three methods described in Section 3.5.1. To measure the clustering performance, we calculate the ARI introduced in (3.19). To measure the estimation performance, we calculate the mean integrated squared error (MISE) between the estimated and the true connecting intensities

$$\text{MISE} = \frac{1}{(K^2+K)/2} \sum_{q,k \in [K], q \le k} \mathbb{E} \| \hat{f}_{q,k} - f_{q,k} \|_2^2, \tag{3.20}$$

where $\| \cdot \|_2$ is the L^2-norm. Note that SidSBM-C estimates cumulative intensities rather than intensities. In order to make a fair comparison, we estimate intensities for SidSBM-C using Gaussian kernel density estimators with bandwidth selected as suggested in Silverman [2017].

Simulation results under various settings are provided in Figure 3.4. In all plots, a smaller y-coordinate indicates a better performance. Across all settings, both proposed methods outperform PPSBM, because PPSBM suffers from the presence of unknown time shifts. In addition, SidSBM-C

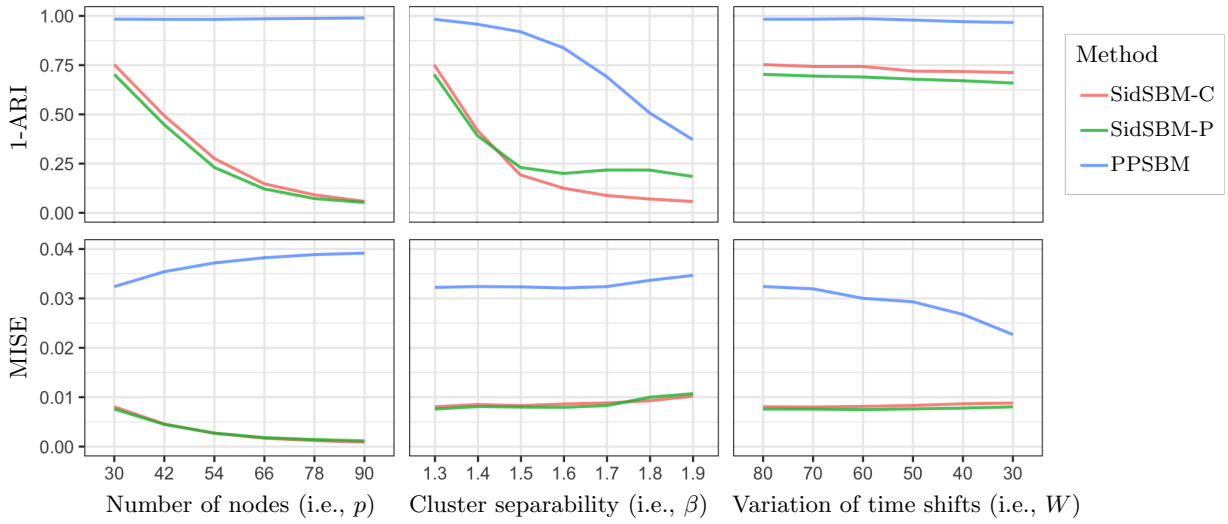

FIGURE 3.4. Performance of SidSBM-C, SidSBM-P and PPSBM in simulation with 1500 replicates. Simulated data is generated under various p, β, W with default setting being $p = 30$, $\beta = 1.3$ and $W = 80$. The y-axis is 1-ARI in the first row and MISE in the second row, where smaller values indicate better performance. The proposals outperform PPSBM in all settings due to their effectiveness in handling time shifts. Between the two proposals, SidSBM-C and SidSBM-P have similar performance except for when $\beta > 1.5$ due to limited intensity support ranges and large time shifts.

and SidSBM-P exhibit comparable performance, except when β reaches high values. This discrepancy may be caused by the limited ranges of supports of intensities when the value of β is high, as well as the large values of W. Supporting this hypothesis, Figure B.5 in Appendix B demonstrates that a reduction in W results in a corresponding decrease of the discrepancy between the performance of SidSBM-C and SidSBM-P. Interestingly, both proposed methods still outperform the PPSBM in settings without time shifts (see Figure B.8 in Appendix B).

Furthermore, Figure 3.4 shows that the overall performance of the three methods improves as signal strengths increase. However, we note that there are four seemingly counterintuitive observations. Firstly, as p increases, PPSBM has stable ARI and increasing MISE. This is because PPSBM is not able to handle time shifts, thus will converge to erroneous estimators as p increases. Secondly, increasing β enlarges MISE of the three methods, due to the increasing average L^2-norm of the representative connecting intensities. Thirdly, the proposed methods have stable performance as W decreases because estimators of cluster memberships and connecting intensities in proposed methods are invariant to time shifts. Fourthly, the MISE values appear to exhibit limited sensitivity

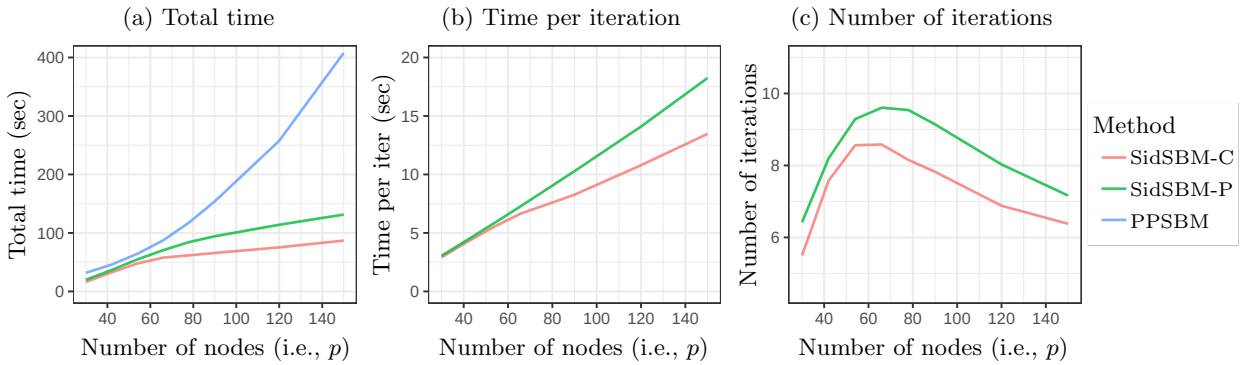

FIGURE 3.5. Computing time of SidSBM-C, SidSBM-P and PPSBM in simulation with 1500 replicates. Simulated data is generated under $\beta = 1.3$, $W = 80$ and a range of p. The proposals are computationally more efficient than PPSBM and might be scalable to large networks.

to improved clustering performance. This pattern can be attributed to several factors, including the small default values of p and β in the simulation setting, as well as the wide range of the y-axis in the figure. In general cases, the MISE value should exhibit sensitivity to improvements in clustering performance. Further supporting evidence is provided in Section B.4.4 in Appendix B.

The computing time of the three methods under consideration is shown in Figure 3.5(a). The proposed methods outperform PPSBM in terms of computing time as well as scalability. In particular, the computing time of both SidSBMs seems to increase sublinearly with respect to the size of network p. To explain this phenomenon, we break down the total computing time into the number of iterations between the centering and clustering steps, and the computing time per iteration shown in Figure 3.5(b) and (c). We find that the computing time per iteration is roughly linear in p, whereas the number of iterations decreases as p increases for $p \geq 60$. This is a promising sign that the proposed methods can potentially handle large networks with the computing cost being linear in the size of network.

One computational concern in non-convex optimization is the requirement of multiple initializations to achieve global minimum. During our simulation experiment, we observe that the initialization scheme in Remark 3.4.1 might lead to local minimums, e.g., empty clusters. Therefore, when the size of any estimated cluster is less than $p/10$, we restart the algorithm by jittering the initial values as discussed in Remark 3.4.1. In Figure 3.6(a), we see that the proposed initialization scheme offers comparable performance to random initialization with three restarts, and that keep increasing the number of random restarts can only result in diminishing improvement in performance.

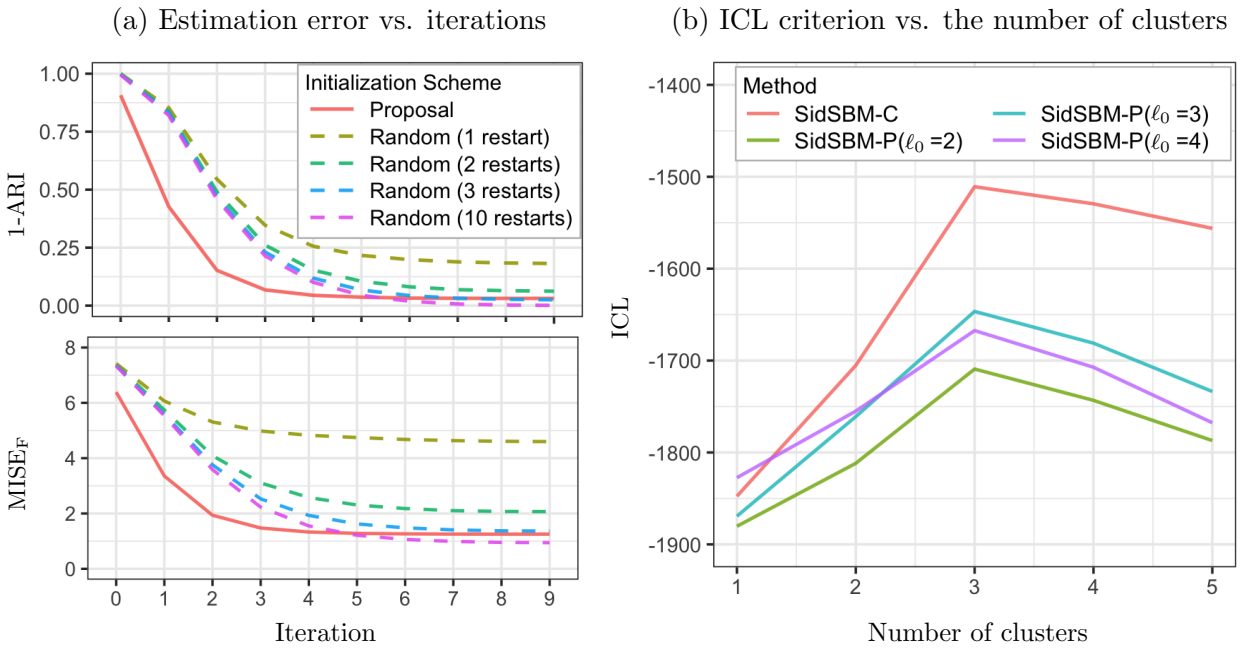

FIGURE 3.6. Performance of proposed initialization scheme and model selection criterion (i.e., ICL) in simulation with over 1500 replicates. Simulated data is generated under the setting with $p = 30$, $\beta = 1.9$, and $V = 80$. Panel (a) displays trajectories of 1-ARI and MISE as iterations proceed, using proposed initialization or random initialization with restarts. When multiple restarts are applied, the best result is selected using the objective function in (3.6). The performance of proposed initialization scheme is comparable to random initialization with ten restarts. Panel (b) displays values of the ICL criterion for various number of clusters. ICL is able to select the correct number of clusters, i.e., $K = 3$.

Furthermore, the proposed initialization scheme requires only 1.07 restarts on average across the 1500 replicates, which is computationally more efficient than random initialization with three or ten restarts.

Finally, we demonstrate the performance of integrated classification likelihood (ICL) criterion in Figure 3.6(b). We observe that for both SidSBMs, ICL values reach maximum at the correct number of clusters. More specifically, for SidSBM-C, ICL correctly selects $\hat{K} = 3$ in all 1500 replicates, and for SidSBM-P, ICL selects $\hat{K} = 3$ in 84.6% of the 1500 replicates.

3.6. Real Data Application

We utilize the proposed model to analyze the neural data set from Wan et al. [2019]. Wan and colleagues recorded the development of motor circuits in the spinal cord of zebrafish. The developmental process was observed from approximately 17.5 hours to 22 hours post fertilization.

In the data set that we analyze, a total of 96 neurons are observed in the spinal cord. The spatial distribution of these neurons is depicted in Figure B.11. Among the observed neurons, 40 neurons are located in the left spine, while the remaining 56 neurons are located in the right spine. The activities and locations of neurons were recorded with calcium imaging and processed by Wan et al. [2019]. During the recorded period, neurons displayed increasing synchronization over time that indicated the functional maturation of the motor circuit. At around 22 hours post fertilization, the motor circuit was able to instruct early motor behaviors. We refer interested readers to Wan et al. [2019] for more experiment details. In this analysis, we aim to learning the roles of neurons during motor circuit development.

We apply SidSBM-C to the recorded development of neural networks where nodes represent neurons and edges represent synchronization between two neurons. Roughly speaking, we define the connecting time between neurons (i.e. $t_{i,j}$'s) to be the time point when the two neurons have correlated activities. Details of preprocessing are available in Section B.5 in Appendix B. Neural networks from the left and right spines are treated as two separate networks in our analysis. To be specific, we apply SidSBM-C with $\varepsilon = 10^{-3}$, $\gamma = 0.1$ and $L = 336$ on the two networks separately, with estimates $(\hat{\mathbf{f}}_L, \hat{\mathbf{z}}_L, \hat{\mathbf{v}}_L)$ and $(\hat{\mathbf{f}}_R, \hat{\mathbf{z}}_R, \hat{\mathbf{v}}_R)$. Here the value of γ is selected using ICL, and it is noteworthy that the data analysis results remain unchanged while γ ranges from 0.01 to 0.3. We simultaneously select the number of clusters (i.e., K) for both spines using the sum of ICLs for the two networks. Each fit of the method takes about 30 seconds (using a machine with an AMD Epyc 7351 processor, 4GB RAM, R 3.6.3). When accounting for the selection of both tuning parameters (i.e., $\gamma \in \{0.01, 0.03, 0.1, 0.3\}$ and $K \in \{2, 3, 4, 5\}$), the overall computational time for the obtained real data results is approximately eight minutes. Recall that Proposition 3.2.1 shows that cluster memberships and intensities are *identifiable up to a permutation* of cluster labels. Hence we match the cluster memberships for the left spine (i.e., $\hat{\mathbf{z}}_L$) and the right spine (i.e., $\hat{\mathbf{z}}_R$) by maximizing the sum of cross-correlations between $\hat{\mathbf{f}}_L$ and $\hat{\mathbf{f}}_R$.

Figure 3.7 shows that estimated representative connecting intensities for left and right spines, i.e., $\hat{\mathbf{f}}_L$ and $\hat{\mathbf{f}}_R$, are highly consistent, although the estimation was done separately on the two networks. In particular, intensities $(\hat{f}_{L,1,1}, \hat{f}_{L,1,2}, \hat{f}_{L,2,2})$ and $(\hat{f}_{R,1,1}, \hat{f}_{R,1,2}, \hat{f}_{R,2,2})$ are almost identical. Furthermore, intensities $\hat{f}_{L,3,k}$ and $\hat{f}_{R,3,k}$ have similar shapes for $k = 1, 2, 3$ despite having a time lag (see Figure B.15 in Appendix B). We hypothesize that the time lag might be due to estimation

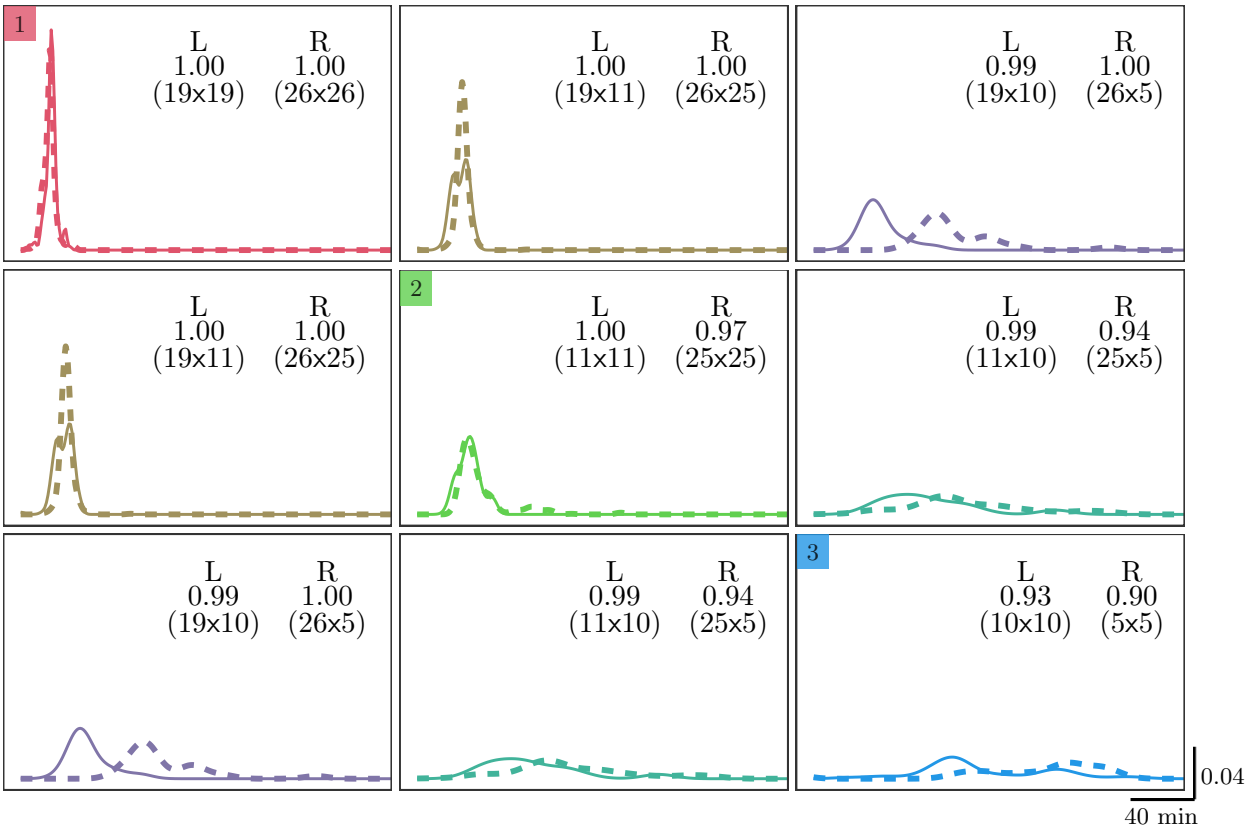

FIGURE 3.7. Estimated connecting intensities on the neural data set from Wan et al. [2019]. The rows and columns are indexed by clusters as shown in the top left squares on diagonal. For $q, k = 1, 2, 3$, solid and dashed curves at position (q, k) represent estimated representative connecting intensity $\hat{f}_{L,q,k}$ and $\hat{f}_{R,q,k}$ from the left and right spines, respectively. Each subfigure contains a tabular legend, where the first row shows "L" for "left" and "R" for "right", the second row shows connecting probabilities, the third row shows the sizes of pairs of clusters. Networks in the left and right spines are fitted separately, but the estimated representative connecting intensities are highly consistent.

errors given the large dispersion of $\hat{f}_{3,k}$'s and the small sizes of the third clusters. Since the two networks are from the same zebrafish, it is natural to expect consistency between them. However, it is worthwhile to emphasize that the estimation procedure does not enforce any constraints to encourage consistency between estimated representative connecting intensities of the two networks. The consistent representative connecting intensities from two sides of spine are likely reflecting the true mechanism of neural circuit development.

In addition, time shifts of neurons in the same cluster can be very different. Indeed, time shifts seem to be positively correlated with the AP-coordinates of neurons as shown in Figure 3.8(a).

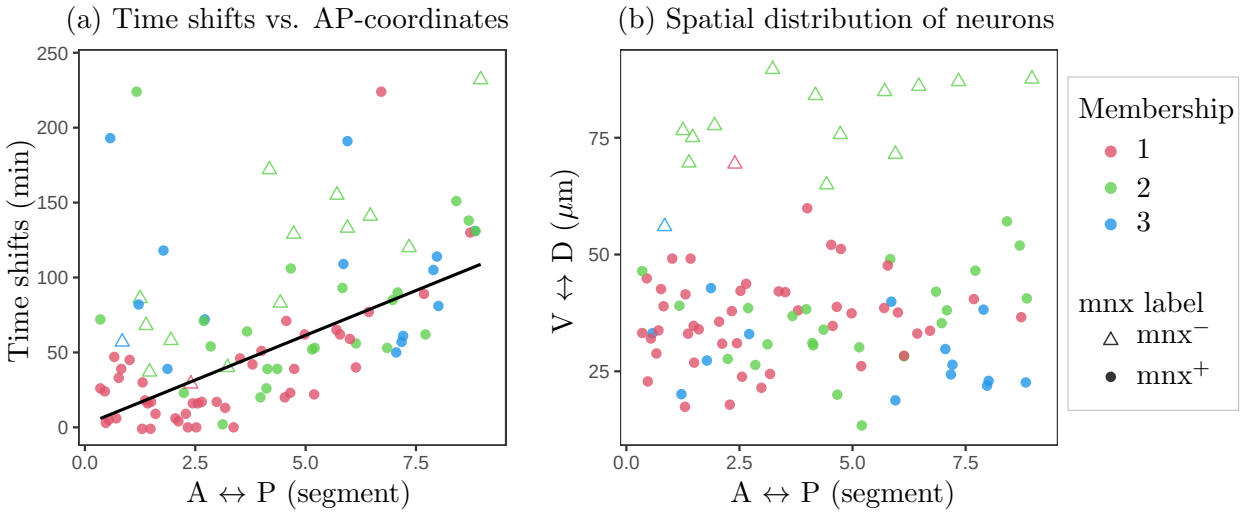

FIGURE 3.8. Estimated time shifts and spatial distribution of neurons in both left and right spines. Each dot represents a neuron, whose color indicates the neuron's estimated cluster membership and the shape indicates the *mnx* label provided in the data set (see Wan et al. [2019] for details). Panel (a) shows that the time shifts might be positively correlated with the AP-coordinates of neurons, where the solid line is the regression line with least absolute deviation (i.e., median regression). Panel (b) shows that most of neurons in the dorsal side that belong to the *mnx*- population have the same estimated cluster membership.

Methods that do not account for time shifts may group neurons with similar time shifts into the same cluster, leading to less consistent estimates of the intensities between the left and right spines. For completeness, we provide the results of applying PPSBM on this data set in Section B.6.5.

The estimated representative connecting intensities and time shifts reveal different roles of the clusters during the development of neural circuits. Firstly, neurons in Cluster 1 might be responsible for initiating neural circuit development, since they are the first to build edges both in the representative connecting intensities and have the smallest time shifts. This early-active behavior aligns with the recognized role of motoneurons [see, e.g., Wenner and O'Donovan, 2001, Blankenship and Feller, 2010, Wan et al., 2019]. A noteworthy association exists between neurons in Cluster 1 and motoneurons: 98% of neurons in Cluster 1 exhibit *mnx1* expression, which is commonly associated with motoneurons [Jao et al., 2012, Wan et al., 2019]. Secondly, neurons in Cluster 2 are likely to be recruited by neurons in Cluster 1, because Cluster 2 tends to establish edges with Cluster 1 faster than among themselves. The recruitment phenomenon of neurons aligns with findings in prior studies [Nishimaru et al., 2005, Blankenship and Feller, 2010, Wan et al., 2019]. Lastly, neurons in Cluster 3 have lower connecting probabilities and slower edge formation speed than neurons in

the other two clusters. We hypothesize that neurons in Cluster 3 may not manifest their functions during the experiment period (i.e., 17.5 hours to 22 hours post fertilization), instead, they might play essential roles in a later stage of the development, for instance, generation and coordination of rhythmic motor patterns [Menelaou and McLean, 2019, Song et al., 2020].

REMARK 3.6.1. *Interestingly, Figure 3.8(b) shows that* 86.7% *of the* 15 mnx^- *neurons (see Wan et al. [2019] for details), which are located at the dorsal side of the spine, are grouped into the same cluster. It is worthwhile to emphasize that the proposed method does not use the spatial locations or the* mnx *labels of neurons. We suspect that the estimated clusters might be biologically meaningful that warrants further investigation.*

REMARK 3.6.2. *The data analysis results substantiate the viability of the conditions in Assumption 3.2.1. We can see from Figure 3.7 that the three clusters demonstrate unique intensity shapes, all intensities deviate from the constant zero function, and the values of* $\hat{f}_{L,1,1}(t)$ *and* $\hat{f}_{R,1,1}(t)$ *become non-zero as soon as* t *surpasses zero, which suggest the feasibility of Conditions C3.1, C3.3, and C3.4 in this data set. Additionally, Figure B.16 in Appendix B presents the distribution of time shifts for neurons within the three clusters, which confirms the attainability of Condition C3.2 in the real data context.*

3.7. Discussion

In this article, we study the problem of estimating cluster memberships and representative connecting intensities from dynamic networks with unknown time shifts. We note that there are a few directions that can be explored in future works.

First of all, the proposed method assumes that edges are persistent and the network has stabilized by the end of the experiment. These assumptions rule out networks where edges can vanish or constantly change, such as biological neural networks during learning. We can generalize the proposed model by introducing two separate intensities to characterize the stochastic formation of edges and stochastic deletion of edges as in Krivitsky and Handcock [2014], Kreiß et al. [2019].

Secondly, we only consider one network in this analysis. A natural extension is to consider networks in a multiple-subject setting [Zhang et al., 2020, Paul and Chen, 2020, Pavlović et al., 2020]. For instance, neural circuits from different zebrafish may share similar connecting intensities

that can be estimated simultaneously. In this direction, a key issue to address is how to match cluster labels across subjects that have different sets of nodes.

Finally, the theoretical properties of the proposed estimators have not been examined. For static stochastic block models, the minimax convergence rate of the estimated connecting probabilities has been well-studied (see, e.g., Gao et al., 2015). Recently, Longepierre and Matias [2019] and Pensky [2019] establish the consistency and minimax optimality for two specific dynamic stochastic block models. Based on the established framework [Gao et al., 2015, Pensky, 2019], we may adapt the theory of shape invariant models [Bigot et al., 2013, Bigot and Gendre, 2013] to discuss the theoretical properties of joint estimation of time shifts and connecting intensities.

APPENDIX A

Appendix for Chapter 2

A.1. Proof of Proposition 2.2.1

Suppose $(a, \mathbf{f}, \mathbf{v}) \in \Theta_0$ are such that for all $t \in [0, T], \mathbf{w}^* \in [0, W]^M, i \in [n]$ the following equality holds:

$$a^* + \sum_{m \in [M]} S^{v^*_{i,m} + w^*_m} f^*_m(t) = a + \sum_{m \in [M]} S^{v_{i,m} + w^*_m} f_m(t). \tag{A.1}$$

Noting that $T_0 + V + W \leq T$ by Assumption C2.1, the equation in (A.1) can be formulated in the frequency domain as follows: for $\xi \neq 0$, $\mathbf{w}^* \in [0, W]^M$, $i \in [n]$,

$$\sum_{m \in [M]} \exp\left\{-\mathrm{j}\, 2\pi\xi(v^*_{i,m} + w^*_m)\right\} \phi^*_m(\xi) = \sum_{m \in [M]} \exp\left\{-\mathrm{j}\, 2\pi\xi(v_{i,m} + w^*_m)\right\} \phi_m(\xi), \tag{A.2}$$

where $\phi^*_m(\xi)$ and $\phi_m(\xi)$ are the Fourier coefficients of $f^*_m(t)$ and $f_m(t)$ at frequency ξ.

We first show an intermediate result that $\exp\left\{-\mathrm{j}\, 2\pi\xi(v^*_{i,m} - v_{i,m})\right\} \phi^*_m(\xi) = \phi_m(\xi)$. By employing matrix notations, (A.2) can be written as, for $\xi \neq 0$, $\mathbf{w}^* \in [0, W]^M$, $i \in [n]$,

$$\boldsymbol{\eta}^*(\xi)^\top \boldsymbol{\psi}^*_i(\xi) = \boldsymbol{\eta}^*(\xi)^\top \boldsymbol{\psi}_i(\xi), \tag{A.3}$$

where

$$\boldsymbol{\eta}^*(\xi) \equiv \left[\exp\left\{-\mathrm{j}\, 2\pi\xi w^*_1\right\} \ \cdots \ \exp\left\{-\mathrm{j}\, 2\pi\xi w^*_M\right\}\right]^\top, \tag{A.4}$$

$$\boldsymbol{\psi}^*_i(\xi) \equiv \left[\exp\left\{-\mathrm{j}\, 2\pi\xi v^*_{i,1}\right\} \phi^*_1(\xi) \ \cdots \ \exp\left\{-\mathrm{j}\, 2\pi\xi v^*_{i,M}\right\} \phi^*_M(\xi)\right]^\top, \tag{A.5}$$

$$\boldsymbol{\psi}_i(\xi) \equiv \left[\exp\left\{-\mathrm{j}\, 2\pi\xi v_{i,1}\right\} \phi_1(\xi) \ \cdots \ \exp\left\{-\mathrm{j}\, 2\pi\xi v_{i,M}\right\} \phi_M(\xi)\right]^\top. \tag{A.6}$$

From Assumption C2.2, we know that $\mathbb{E}[\overline{\boldsymbol{\eta}^*(\xi)}\boldsymbol{\eta}^*(\xi)^\top]$ is invertible for $\xi \in \mathbb{R} \setminus \{0\}$. Therefore, we know from (A.3) that, for $\xi \neq 0$, $i \in [n]$,

$$\boldsymbol{\psi}^*_i(\xi) = \boldsymbol{\psi}_i(\xi). \tag{A.7}$$

Substituting the definitions of $\boldsymbol{\psi}_i^*(\xi)$ and $\boldsymbol{\psi}_i(\xi)$ in (A.5) and (A.6) into (A.7), we obtain that, for $\xi \neq 0$, $i \in [n]$, $m \in [M]$,

$$\exp\left\{-\mathrm{j}\, 2\pi\xi(v_{i,m}^* - v_{i,m})\right\} \phi_m^*(\xi) = \phi_m(\xi). \tag{A.8}$$

Now we show that, for $m \in [M]$ such that the set $\{t : f_m^*(t) \neq 0\}$ is of positive measure, the subject-specific time shifts are identifiable up to a constant, i.e., $v_{i,m}^* - v_{i,m} = c_m$ for $i \in [n]$, where c_m is a constant independent of i. Based on (A.8), it can be derived that for any $\xi \neq 0$, $m \in [M]$, $i, i' \in [n]$,

$$\exp\left\{-\mathrm{j}\, 2\pi\xi(v_{i,m}^* - v_{i,m})\right\} \phi_m^*(\xi) = \exp\left\{-\mathrm{j}\, 2\pi\xi(v_{i',m}^* - v_{i',m})\right\} \phi_m^*(\xi). \tag{A.9}$$

Consider function $g(t) \equiv S^{v_{i,m}^* - v_{i,m}} f_m^*(t) - S^{v_{i',m}^* - v_{i',m}} f_m^*(t)$. From (A.9) we know that the Fourier coefficient of $g(t)$ is zero at frequency $\xi \neq 0$. Moreover, from the definition of the set $\mathcal{F}$, we know that the function $g(t)$ has a bounded support. By applying Lemma A.8.0.3 to the function $g(t)$, we deduce that $g(t) = 0$ almost everywhere. Consequently, for any $i, i' \in [n]$ and $m \in [M]$,

$$S^{v_{i,m}^* - v_{i,m}} f_m^*(t) - S^{v_{i',m}^* - v_{i',m}} f_m^*(t) = 0 \quad \text{a.e.} \tag{A.10}$$

or equivalently,

$$S^{(v_{i,m}^* - v_{i,m}) - (v_{i',m}^* - v_{i',m})} f_m^*(t) = f_m^*(t), \quad \text{a.e.} \tag{A.11}$$

The equation in (A.11) implies that, for $m \in [M]$ such that the set $\{t : f_m^*(t) \neq 0\}$ is of positive measure, $(v_{i,m}^* - v_{i,m}) - (v_{i',m}^* - v_{i',m}) = 0$ for any $i, i' \in [n]$. Consequently, for $i \in [n]$,

$$v_{i,m}^* - v_{i,m} = c_m, \tag{A.12}$$

where c_m is a constant independent of i. In other words, Statement P2.2 is proved.

Next we show that, for $m \in [M]$, $S^{c_m} f_m^*(t) = f_m(t)$ for almost every $t \in \mathbb{R}$. Plugging (A.12) into (A.8), we obtain that, for $m \in [M]$ such that the set $\{t : f_m^*(t) \neq 0\}$ is of positive measure and $\xi \neq 0$,

$$\exp\left\{-\mathrm{j}\, 2\pi\xi c_m\right\} \phi_m^*(\xi) = \phi_m(\xi). \tag{A.13}$$

Applying the inverse Fourier transformation to (A.13) yields

$$S^{c_m} f_m^*(t) = f_m(t) + c \quad \text{a.e.}, \tag{A.14}$$

where $c \in \mathbb{R}$ is a constant. From the definition of $\mathcal{F}$, we know that

$$f_m^*(t) = f_m(t) = 0, \text{ for } t \in \mathbb{R} \setminus [0, T_0]. \tag{A.15}$$

Combining (A.14) and (A.15), we can derive that $c = 0$. Inserting the value of c to (A.14) yields that, for $m \in [M]$ such that the set $\{t : f_m^*(t) \neq 0\}$ is of positive measure,

$$S^{c_m} f_m^*(t) = f_m(t) \quad \text{a.e.} \tag{A.16}$$

Notably, for $m \in [M]$ such that $f_m^* = 0$ almost everywhere, (A.16) always holds. Thus (A.16) holds for all $m \in [M]$, in other words, Statement P2.1 is proved.

Finally, substituting (A.16) and (A.12) into (A.1), we derive that,

$$a^* = a. \tag{A.17}$$

In other words, Statement P2.3 is proved. □

A.2. Proof of Proposition 2.3.1

Suppose there exist $(\mathbf{z}, \boldsymbol{a}, \mathbf{f}, \mathbf{v}) \in \Theta_1$ such that for all $t \in [0, T]$, $\mathbf{w}^* \in [0, W]^M$, $i \in [n]$, the following equation holds:

$$a_{z_i^*}^* + \sum_{m \in [M]} S^{v_{i,m}^* + w_m^*} f_{z_i^*, m}^*(t) = a_{z_i} + \sum_{m \in [M]} S^{v_{i,m} + w_m^*} f_{z_i, m}(t). \tag{A.18}$$

Using $T_0 + V + W \leq T$ by Assumption C2.1, (A.18) can be formulated in the frequency domain as follows: for $\xi \neq 0$, $\mathbf{w}^* \in [0, W]^M$, $i \in [n]$,

$$\sum_{m \in [M]} \exp\left\{-\mathrm{j}\, 2\pi\xi (v_{i,m}^* + w_m^*)\right\} \phi_{z_i^*, m}^*(\xi) = \sum_{m \in [M]} \exp\left\{-\mathrm{j}\, 2\pi\xi (v_{i,m} + w_m^*)\right\} \phi_{z_i, m}(\xi), \tag{A.19}$$

where $\phi_{k,m}^*(\xi)$ and $\phi_{k,m}(\xi)$ are the Fourier coefficients of $f_{k,m}^*$ and $f_{k,m}$ at frequency ξ.

We first show an intermediate result that $\exp\left\{-\mathrm{j}\,2\pi\xi(v^*_{i,m}-v_{i,m})\right\}\phi^*_{z^*_i,m}(\xi) = \phi_{z_i,m}(\xi)$. By employing matrix notations, (A.19) can be written as

$$\boldsymbol{\eta^*}(\xi)^\top \boldsymbol{\psi}^*_i(\xi) = \boldsymbol{\eta^*}(\xi)^\top \boldsymbol{\psi}_i(\xi), \tag{A.20}$$

where

$$\boldsymbol{\eta}^*(\xi) \equiv [\exp\{-\mathrm{j}\,2\pi\xi w^*_1\}\ \cdots\ \exp\{-\mathrm{j}\,2\pi\xi w^*_M\}]^\top, \tag{A.21}$$

$$\boldsymbol{\psi}_i{}^*(\xi) \equiv \left[\exp\left\{-\mathrm{j}\,2\pi\xi v^*_{i,1}\right\}\phi^*_{z^*_i,1}(\xi)\ \cdots\ \exp\left\{-\mathrm{j}\,2\pi\xi v^*_{i,M}\right\}\phi^*_{z^*_i,M}(\xi)\right]^\top, \tag{A.22}$$

$$\boldsymbol{\psi}_i(\xi) \equiv [\exp\{-\mathrm{j}\,2\pi\xi v_{i,1}\}\,\phi_{z_i,1}(\xi)\ \cdots\ \exp\{-\mathrm{j}\,2\pi\xi v_{i,M}\}\,\phi_{z_i,M}(\xi)]^\top. \tag{A.23}$$

From Assumption C2.2 we know that $\mathbb{E}[\overline{\boldsymbol{\eta}^*(\xi)}\boldsymbol{\eta}^*(\xi)^\top]$ is invertible for $\xi \in \mathbb{R}\setminus\{0\}$. Therefore, we have that, for $\xi \neq 0$,

$$\boldsymbol{\psi}^*_i(\xi) = \boldsymbol{\psi}_i(\xi), \tag{A.24}$$

Substituting the definitions of $\boldsymbol{\psi}^*_i(\xi)$ and $\boldsymbol{\psi}_i(\xi)$ in (A.22) and (A.23) into (A.24), we obtain that, for $\xi \neq 0$, $i \in [n]$, $m \in [M]$,

$$\exp\left\{-\mathrm{j}\,2\pi\xi(v^*_{i,m}-v_{i,m})\right\}\phi^*_{z^*_i,m}(\xi) = \phi_{z_i,m}(\xi). \tag{A.25}$$

Now we show that cluster memberships are identifiable up to a permutation of cluster labels, i.e., $z_i = \sigma(z^*_i)$ where $\sigma : [K] \to [K]$ is a permutation of $[K]$. To achieve this, we prove the following two statements: (i) $z_i = z_{i'} \Rightarrow z^*_i = z^*_{i'}$; and (ii) $z_i \neq z_{i'} \Rightarrow z^*_i \neq z^*_{i'}$. First, based on (A.25), we can derive that for $i, i' \in [n]$ such that $z_i = z_{i'}$, $\xi \neq 0$, and $m \in [M]$,

$$\exp\left\{-\mathrm{j}\,2\pi\xi(v^*_{i,m}-v_{i,m})\right\}\phi^*_{z^*_i,m}(\xi) = \exp\left\{-\mathrm{j}\,2\pi\xi(v^*_{i',m}-v_{i',m})\right\}\phi^*_{z^*_{i'},m}(\xi). \tag{A.26}$$

Consider a function $g(t)$ defined as $g(t) \equiv S^{v^*_{i,m}-v_{i,m}} f^*_{z^*_i,m}(t) - S^{v^*_{i',m}-v_{i',m}} f^*_{z^*_{i'},m}(t)$. From (A.26) we know that the Fourier transform of $g(t)$ is zero for $\xi \neq 0$. Moreover, from the definition of $\mathcal{F}$ we know that the function $g(t)$ has a bounded support. Applying Lemma A.8.0.3 to $g(t)$, we deduce that $g(t) = 0$ almost everywhere. Consequently, for any $i, i' \in [n]$ that $z_i = z_{i'}$ and $m \in [M]$, the

following equation holds:

$$S^{v^*_{i,m}-v_{i,m}} f^*_{z^*_i,m}(t) - S^{v^*_{i',m}-v_{i',m}} f^*_{z^*_{i'},m}(t) = 0, \quad \text{a.e.} \tag{A.27}$$

According to Assumption C2.3, if $z^*_i \neq z^*_{i'}$, then there exists $m_0 \in [M]$ such that for any $x \in \mathbb{R}$, $\{t \in \mathbb{R} : S^x f^*_{z^*_i,m_0}(t) \neq f^*_{z^*_{i'},m_0}(t)\}$ has a positive measure. Therefore, in order for (A.27) to hold, we must have $z^*_i = z^*_{i'}$. In other words,

$$z_i = z_{i'} \Rightarrow z^*_i = z^*_{i'}. \tag{A.28}$$

Second, suppose there exist i_0, i_1 such that $z_{i_0} \neq z_{i_1}$ and $z^*_{i_0} = z^*_{i_1}$. Given that there are K non-empty clusters by definition, it is always possible to find indices $i_2, \ldots, i_K$ such that $z^*_{i_1}, \ldots, z^*_{i_K}$ are pairwise distinct. Based on (A.28), it follows that $z_{i_1}, \ldots, z_{i_K}$ are also pairwise distinct. Since $z_{i_0} \neq z_{i_1}$ by definitions of i_0 and i_1, there must exist $k \in [K] \setminus \{1\}$ such that $z_{i_0} = z_{i_k}$, which, according to (A.28), implies $z^*_{i_0} = z^*_{i_k}$. Thus, by definitions of i_0 and i_1, we have $z^*_{i_0} = z^*_{i_1} = z^*_{i_k}$. This contradicts with the definition of $i_2, \ldots, i_K$, which asserts that $z^*_{i_1}$ is different from $z^*_{i_2}, \ldots, z^*_{i_K}$. Consequently, such i_0 and i_1 cannot exist. Hence we know that, for all $i, i' \in [n]$,

$$z_i \neq z_{i'} \Rightarrow z^*_i \neq z^*_{i'}. \tag{A.29}$$

Combining (A.29) and (A.28), we obtain that $z^*_i = z^*_{i'}$ if and only if $z_i = z_{i'}$. This further implies that there exists a permutation of $[K]$, denoted by $\sigma : [K] \to [K]$, such that for $i \in [n]$,

$$z_i = \sigma(z^*_i). \tag{A.30}$$

In other words, Statement P2.4 is proved.

Now we show that, for $k \in [K], m \in [M]$ such that the set $\{t : f^*_{k,m}(t) \neq 0\}$ is of positive measure, the subject-specific time shifts are identifiable up to a constant, i.e., $v^*_{i,m} - v_{i,m} = c_{k,m}$ for $i \in \{i : z^*_i = k\}$, where $c_{k,m}$ is a constant independent of i. Plugging (A.30) into (A.27), we have that, for $i, i' \in [n]$ such that $z^*_i = z^*_{i'} = k$ and $m \in [M]$,

$$S^{v^*_{i,m}-v_{i,m}} f^*_{k,m}(t) = S^{v^*_{i',m}-v_{i',m}} f^*_{k,m}(t), \quad \text{a.e.} \tag{A.31}$$

From the definition of $\mathcal{F}$ we know that $f^*_{k,m}(t)$ has a bounded support for $k \in [K]$ and $m \in [M]$. Therefore, (A.31) implies that, for $k \in [K]$ and $m \in [M]$ such that the set $\{t : f^*_{k,m}(t) \neq 0\}$ is of positive measure, and $i, i' \in [n]$ such that $z^*_i = z^*_{i'}$,

$$v^*_{i,m} - v^*_{i',m} = v_{i,m} - v_{i',m}. \tag{A.32}$$

It follows from (A.32) that

$$v^*_{i,m} - v_{i,m} = c_{k,m}, \tag{A.33}$$

where $c_{k,m} \in \mathbb{R}$ is a constant independent of i. In other words, Statement P2.7 is proved.

Next we show that $S^{c_{k,m}} f^*_{k,m}(t) = f_{\sigma(k),m}(t)$ for almost every $t \in \mathbb{R}$. Plugging (A.33) and (A.30) into (A.25), we obtain that, for $k \in [K]$ and $m \in [M]$ such that the set $\{t : f^*_{k,m}(t) \neq 0\}$ is of positive measure and $\xi \neq 0$,

$$\exp\{-\mathrm{j}\, 2\pi\xi c_{k,m}\}\, \phi^*_{k,m}(\xi) = \phi_{\sigma(k),m}(\xi), \tag{A.34}$$

Applying the inverse Fourier transformation to (A.34) yields

$$S^{c_{k,m}} f^*_{k,m}(t) = f_{\sigma(k),m}(t) + c \quad \text{a.e.}, \tag{A.35}$$

where $c \in \mathbb{R}$ is a constant. From the definition of $\mathcal{F}$, we know that

$$f^*_{k,m}(t) = f_{k,m}(t) = 0, \text{ for } t \in \mathbb{R} \setminus [0, T_0]. \tag{A.36}$$

Combining (A.35) and (A.36), we can derive that $c = 0$. Inserting the value of c to (A.35) yields that, for $k \in [K]$ and $m \in [M]$ such that the set $\{t : f^*_{k,m}(t) \neq 0\}$ is of positive measure,

$$S^{c_{k,m}} f^*_{k,m}(t) = f_{\sigma(k),m}(t) \quad \text{a.e.} \tag{A.37}$$

Notably, for $k \in [K]$ and $m \in [M]$ such that $f^*_{k,m} = 0$ almost everywhere, (A.37) always holds. In other words, (A.37) holds for all $k \in [K]$ and $m \in [M]$. Therefore, Statement P2.6 is proved.

Finally, substituting (A.37), (A.33) and (A.30) into (A.18), we derive that, for $k \in [K]$,

$$a^*_k = a_{\sigma(k)}. \tag{A.38}$$

In other words, Statement P2.5 is proved. □

A.3. Connection Between the Additive Shape Invariant Model and FPCA

When the variances of $v_{i,m}$'s and $w^*_{r,m}$'s are both close to zero, the proposed model in (2.1) can be approximated using the Taylor expansion:

$$\begin{aligned}
\lambda_{i,r}(t) &= a + \sum_{m\in[M]} S^{u_{i,r,m}} f_m(t) \\
&= a + \sum_{m\in[M]} f_m\left(\{t - \mathbb{E}u_{i,r,m}\} - \{u_{i,r,m} - \mathbb{E}u_{i,r,m}\}\right) \\
&\approx a + \sum_{m\in[M]} \{f_m(t - \mathbb{E}u_{i,r,m}) - (u_{i,r,m} - \mathbb{E}u_{i,r,m}) Df_m(t - \mathbb{E}u_{i,r,m})\} \\
&= a + \sum_{m\in[M]} f_m(t - \mathbb{E}u_{i,r,m}) - \sum_{m\in[M]} (u_{i,r,m} - \mathbb{E}u_{i,r,m}) Df_m(t - \mathbb{E}u_{i,r,m}),
\end{aligned} \tag{A.39}$$

where $u_{i,r,m} \equiv v_{i,m} + w^*_{r,m}$, and Df_m denotes the first order derivative of f_m. In (A.39), the first equality follows from the definition of $u_{i,r,m}$, the second equality follows from the definition of the shift operator, the approximation in the third line follows from the Taylor expansion. To elucidate the connection between the proposed additive shape invariant model and the FPCA, we define a new set of parameters:

$$\mu(t) \equiv a + \sum_{m\in[M]} f_m(t - \mathbb{E}u_{i,r,m}), \tag{A.40}$$

$$\zeta_{i,r,m} \equiv -(u_{i,r,m} - \mathbb{E}u_{i,r,m}) \left\|Df_m(t - \mathbb{E}u_{i,r,m})\right\|_t, \tag{A.41}$$

$$\psi_m(t) \equiv Df_m(t - \mathbb{E}u_{i,r,m}) \left\|Df_m(t - \mathbb{E}u_{i,r,m})\right\|_t^{-1}. \tag{A.42}$$

By definitions of $\zeta_{i,r,m}$ and $\psi_m(t)$ in (A.41) and (A.42), we know that $\mathbb{E}\zeta_{i,r,m} = 0$, and $\|\psi_m(t)\|_t = 1$. Using the new set of parameters in (A.40), (A.41) and (A.42), the model approximation in (A.39) can be expressed as follows:

$$\lambda_{i,r}(t) \approx \mu(t) + \sum_{m\in[M]} \zeta_{i,r,m}\ \psi_m(t). \tag{A.43}$$

When $\{f_m(t - \mathbb{E}u_{i,r,m}) : m \in [M]\}$ have non-overlapping supports, it follows that $\psi_m(t)$'s are mutually orthogonal. Consequently, the approximate model in (A.43) is an FPCA model, where

$\psi_1(t), \cdots, \psi_M(t)$ correspond to the first M eigenfunctions, and $\zeta_{i,r,1}, \ldots, \zeta_{i,r,M}$ correspond to the principal components associated with the eigenfunctions.

A.4. Detailed Derivation of Solutions in (2.14) and (2.15)

The first optimization problem in the centering step (2.11) can be elaborated as follows:

$$\begin{aligned}\hat{\mathbf{a}}', \hat{\mathbf{f}}' &= \arg\min_{\mathbf{a}', \mathbf{f}'} L_1(\hat{\mathbf{z}}, \mathbf{a}', \mathbf{f}', \hat{\mathbf{v}}) \\ &= \arg\min_{\mathbf{a}', \mathbf{f}'} \sum_{i\in[n], r\in[R]} \beta_{i,r} \frac{1}{T} \left\| \frac{y_{i,r}(t)}{N_{i,r}(T)} - \left\{ a'_{\hat{z}_i} + \sum_{m\in[M]} S^{\hat{v}_{i,m}+w^*_{r,m}} f'_{\hat{z}_i,m}(t) \right\} \right\|_t^2 .\end{aligned} \tag{A.44}$$

Utilizing the renowned Parseval's theorem, the optimization problem in (A.44) can be formulated as follows:

$$\begin{aligned}\hat{\mathbf{a}}', \hat{\boldsymbol{\phi}}' = \arg\min_{\mathbf{a}', \boldsymbol{\phi}'} \sum_{i\in[n], r\in[R]} \beta_{i,r} \sum_{l\in\mathbb{Z}} \Bigg| \frac{\eta_{i,r,l}}{N_{i,r}(T)} - \Bigg\{ a'_{\hat{z}_i} \mathbf{1}(l=0) \; + \\ \sum_{m\in[M]} \exp\left\{ -\mathrm{j}\, 2\pi l (\hat{v}_{i,m} + w^*_{r,m}) T^{-1} \right\} \phi'_{\hat{z}_i,m,l} \Bigg\} \Bigg|^2 ,\end{aligned} \tag{A.45}$$

where $\boldsymbol{\phi}' \equiv (\phi'_{k,m,l})_{k\in[K], m\in[M], l\in\mathbb{Z}}$, $\{\phi'_{\hat{z}_i,m,l} : l \in \mathbb{Z}\}$ denotes the Fourier coefficients of $f'_{z_i,m}(t)$, $\{\eta_{i,r,l} : l \in \mathbb{Z}\}$ denotes the Fourier coefficients of $y_{i,r}(t)$, and j denotes the imaginary unit.

The optimization problem in (A.45) can be solved by breaking it down to smaller independent problems. For $k \in [K]$ and $l \in \mathbb{Z}$, let $\boldsymbol{\phi}'_{k,*,l} \equiv (\phi'_{k,m,l})_{m\in[M]}$. For $l \neq 0$, the objective function associated with $\boldsymbol{\phi}'_{k,*,l}$ is essentially a weighted sum of squares:

$$\hat{\boldsymbol{\phi}}'_{k,*,l} = \arg\min_{\boldsymbol{\phi}'_{k,*,l}} (\mathbf{h}_{k,l} - \mathbf{E}_{k,l}\boldsymbol{\phi}'_{k,*,l})^\top \mathbf{B}_k \, \overline{(\mathbf{h}_{k,l} - \mathbf{E}_{k,l}\boldsymbol{\phi}'_{k,*,l})}, \tag{A.46}$$

where $\mathbf{h}_{k,l} \equiv (\eta_{i,r,l} N_{i,r}(T)^{-1})_{(i,r)\in\hat{\mathcal{C}}_k\times[R]}$, $\hat{\mathcal{C}}_k \equiv \{i \in [n] : \hat{z}_i = k\}$, $\mathbf{E}_{k,l} \equiv [\exp\{-\mathrm{j}\, 2\pi l(\hat{v}_{i,m} + w^*_{r,m})T^{-1}\}]_{(i,r)\in\hat{\mathcal{C}}_k\times[R], m\in[M]}$, $\mathbf{B}_k$ is a diagonal matrix of $(\beta_{i,r})_{(i,r)\in\hat{\mathcal{C}}_k\times[R]}$, and $\overline{z}$ denotes the complex conjugate of z for any $z \in \mathbb{C}$. As a result, the solution to (A.46) is:

$$\hat{\boldsymbol{\phi}}'_{k,*,l} = \left(\overline{\mathbf{E}_{k,l}}^\top \mathbf{B}_k \, \mathbf{E}_{k,l} \right)^{-1} \left(\overline{\mathbf{E}_{k,l}}^\top \mathbf{B}_k \, \mathbf{h}_{k,l} \right), \quad \text{for } l \neq 0. \tag{A.47}$$

For $l = 0$, the parameter $\boldsymbol{\phi}'_{k,*,0}$ can be estimated by exploiting the definition of $\mathcal{F}$. From the definition of $\mathcal{F}$ we know that $f_{k,m}(0) = 0$ for $k \in [K], m \in [M]$. Consequently, $f'_{k,m}(0) = f_{k,m}(0)\Lambda_k^{-1} = 0$.

In addition, using the inverse Fourier transformation, we can derive that $\sum_{l\in\mathbb{Z}}\phi'_{k,m,l} = f'_{k,m}(0)$. Hence we know that $\sum_{l\in\mathbb{Z}}\phi'_{k,m,l} = 0$. This leads to the estimate of $\boldsymbol{\phi}'_{k,*,0}$ as

$$\hat{\boldsymbol{\phi}}'_{k,*,0} = -\sum_{|l|\le\ell_0, l\neq 0}\hat{\boldsymbol{\phi}}'_{k,*,l}, \tag{A.48}$$

where ℓ_0 is the truncation frequency to facilitate the numerical feasibility of computing $\hat{\boldsymbol{\phi}}'_{k,*,0}$.

Based on $\hat{\boldsymbol{\phi}}'$, it is straightforward to obtain estimation for $\mathbf{a}'$ as follows:

$$\begin{aligned}\hat{a}'_k &= \arg\min_{a'_k}\sum_{i\in\hat{\mathcal{C}}_k, r\in[R]}\beta_{i,r}\left|\frac{\eta_{i,r,0}}{N_{i,r}(T)} - \left\{a'_k + \sum_{m\in[M]}\hat{\phi}'_{k,m,0}\right\}\right|^2 \\ &= \left[\sum_{i\in\hat{\mathcal{C}}_k, r\in[R]}\beta_{i,r}\left\{\frac{\eta_{i,r,0}}{N_{i,r}(T)} - \sum_{m\in[M]}\hat{\phi}'_{k,m,0}\right\}\right]\left[\sum_{i\in\hat{\mathcal{C}}_k, r\in[R]}\beta_{i,r}\right]^{-1} \\ &= T^{-1} - \sum_{m\in[M]}\hat{\phi}'_{k,m,0},\end{aligned} \tag{A.49}$$

where in the third equality we use $\eta_{i,r,0} = T^{-1}\int_0^T y_{i,r}(t)\mathrm{d}t = T^{-1}N_{i,r}(T)$. Moreover, using the inverse Fourier transform, the estimation for $\mathbf{f}'$ can be derived as follows:

$$\hat{f}'_{k,m}(t) = \sum_{|l|\le\ell_0}\hat{\phi}_{k,m,l}\exp(\,\mathrm{j}\,2\pi l t T^{-1}). \tag{A.50}$$

A.5. Implementation of Newton's Method in the Clustering Step

We employ the Newton's method to solve the optimization problem in (2.23) of the main text. The objective function in (2.23) can be formulated in the frequency domain:

$$L_{1,i}(k, \hat{\mathbf{a}}', \hat{\mathbf{f}}', \mathbf{v}_i) = \sum_{r\in[R]}\beta_{i,r}\sum_{l\in\mathbb{Z}}L_{1,i,l}(\mathbf{v}_i), \tag{A.51}$$

where $L_{1,i,l}(\mathbf{v}_i)$ is defined as

$$L_{1,i,l}(\mathbf{v}_i) \equiv \left|\frac{\eta_{i,r,l}}{N_{i,r}(T)} - \left\{\hat{a}'_k\mathbf{1}(l=0) + \sum_{m\in[M]}\exp\left\{-\,\mathrm{j}\,2\pi l(v_{i,m} + w^*_{r,m})T^{-1}\right\}\hat{\phi}'_{k,m,l}\right\}\right|^2. \tag{A.52}$$

The definition in (A.52) suggests that $L_{1,i,l}(\mathbf{v}_i)$ remains constant with respect to $\mathbf{v}_i$ when $l = 0$. Moreover, since $\hat{f}_{k,m}$ is calculated using the truncated Fourier series (see (2.18) of the main text), it follows that $\hat{\phi}'_{k,m,l} = 0$ for $|l| > \ell_0$. As a result, $L_{1,i,l}(\mathbf{v}_i)$ remains constant with respect to $\mathbf{v}_i$ for

$|l| > \ell_0$. Therefore, the optimization problem in (2.23) can be formulated in the frequency domain as

$$\begin{aligned}\tilde{\mathbf{v}}_{i|k} &= \arg\min_{\mathbf{v}_i} \sum_{r\in[R]} \beta_{i,r} \sum_{|l|\le\ell_0, l\neq 0} L_{1,i,l}(\mathbf{v}_i) \\ &= \arg\min_{\mathbf{v}_i} \sum_{r\in[R]} \beta_{i,r} \sum_{|l|\le\ell_0, l\neq 0} \left| \frac{\eta_{i,r,l}}{N_{i,r}(T)} - \sum_{m\in[M]} \exp\big\{ -\mathrm{j}\, 2\pi l (v_{i,m} + w^*_{r,m}) T^{-1} \big\} \hat{\phi}'_{k,m,l} \right|^2 \\ &\equiv \arg\min_{\mathbf{v}_i} Q_i(\mathbf{v}_i).\end{aligned} \tag{A.53}$$

where the first equality follows from the fact that $L_{1,i,l}(\mathbf{v}_i)$ remains constant with respect to $\mathbf{v}_i$ for $l = 0$ or $|l| > \ell_0$, and the second equality follows from the definition of $L_{1,i,l}(\mathbf{v}_i)$ in (A.52).

We solve the optimization problem in (A.53) using the Newton's method. Specifically, the estimate of $\mathbf{v}_i$ is updated iteratively via

$$\hat{\mathbf{v}}_i \leftarrow \hat{\mathbf{v}}_i - \mathrm{trunc}\{[\nabla^2 Q_i(\hat{\mathbf{v}}_i)]^{-1} \nabla Q_i(\hat{\mathbf{v}}_i)\}, \tag{A.54}$$

where $\mathrm{trunc}\{x\}$ is a function defined as

$$\mathrm{trunc}\{x\} \equiv \begin{cases} -T/10, & \text{if } x < -T/10, \\ x, & \text{if } x \in [-T/10, T/10], \\ T/10, & \text{if } x > T/10, \end{cases} \tag{A.55}$$

$\nabla^2 Q_i(\hat{\mathbf{v}}_i) \equiv (\partial^2 Q_i(\mathbf{v}_i)/\partial v_{i,m_1} \partial v_{i,m_2})_{(m_1,m_2)\in[M]^2}$ denotes the Hessian matrix, and $\nabla Q_i(\hat{\mathbf{v}}_i) \equiv (\partial Q_i(\mathbf{v}_i)/\partial v_{i,m})_{m\in[M]}$ denotes the gradient. The gradient can be calculated through the following partial derivatives: for $m \in [M]$,

$$\begin{aligned}\frac{\partial Q_i(\mathbf{v}_i)}{\partial v_{i,m}} = -2 \sum_{r\in[R]} \beta_{i,r} \sum_{|l|\le\ell_0, l\neq 0} \Re e \Bigg(& \Big[(-\mathrm{j}\, 2\pi l T^{-1}) \exp\big\{ -\mathrm{j}\, 2\pi l (v_{i,m} + w^*_{r,m}) T^{-1} \big\} \hat{\phi}'_{k,m,l} \Big] \\ & \times \overline{\Big[\frac{\eta_{i,r,l}}{N_{i,r}(T)} - \sum_{m'\in[M]\setminus\{m\}} \exp\big\{ -\mathrm{j}\, 2\pi l (v_{i,m'} + w^*_{r,m'}) T^{-1} \big\} \hat{\phi}'_{k,m',l} \Big]} \Bigg).\end{aligned} \tag{A.56}$$

The Hessian matrix can be calculated through the following second order partial derivatives: for $m \in [M]$,

$$\frac{\partial^2 Q_i(\mathbf{v}_i)}{\partial v_{i,m}^2} = -2 \sum_{r\in[R]} \beta_{i,r} \sum_{|l|\le \ell_0, l\neq 0} \Re e\Bigg(\Big[(-\,\mathrm{j}\, 2\pi l T^{-1})^2 \exp\big\{ -\,\mathrm{j}\, 2\pi l (v_{i,m} + w^*_{r,m}) T^{-1} \big\} \hat{\phi}'_{k,m,l} \Big] \times \overline{\Big[\frac{\eta_{i,r,l}}{N_{i,r}(T)} - \sum_{m'\in[M]\setminus\{m\}} \exp\big\{ -\,\mathrm{j}\, 2\pi l (v_{i,m'} + w^*_{r,m'}) T^{-1} \big\} \hat{\phi}'_{k,m',l} \Big]} \Bigg), \tag{A.57}$$

while for $m_1, m_2 \in [M]$ such that $m_1 \neq m_2$,

$$\frac{\partial^2 Q_i(\mathbf{v}_i)}{\partial v_{i,m_1} \partial v_{i,m_2}} = 2 \sum_{r\in[R]} \beta_{i,r} \sum_{|l|\le \ell_0, l\neq 0} \Re e\Bigg(\Big[(-\,\mathrm{j}\, 2\pi l T^{-1}) \exp\big\{ -\,\mathrm{j}\, 2\pi l (v_{i,m_1} + w^*_{r,m_1}) T^{-1} \big\} \hat{\phi}'_{k,m_1,l} \Big] \times \overline{\Big[(-\,\mathrm{j}\, 2\pi l T^{-1}) \exp\big\{ -\,\mathrm{j}\, 2\pi l (v_{i,m_2} + w^*_{r,m_2}) T^{-1} \big\} \hat{\phi}'_{k,m_2,l} \Big]} \Bigg). \tag{A.58}$$

A.6. Additional Simulation Results

A.6.1. The fourth experiment. In the fourth experiment, we evaluate the performance of the proposed method under conditions simulating the scale of intensity components in the neural data. Specifically, we set $n = 40$, $K = 4$, $a_k = 2$, and set $f_{k,m}$'s as shown in Table A.1. The remaining parameters are consistent with those used in the second and third experiments. We apply the proposed method with $K = 4$ and $\gamma = 0.01$. The results are presented in Figure A.1. It is evident that both the clustering and intensity estimation performances improve with increasing values of R and τ. This finding is consistent with the observations from Experiments 1 and 3. Furthermore, in comparison to Experiment 3, the magnitudes of a_k's and the scales of $f_{k,m}$'s are reduced by a factor of 10. Consequently, Figure A.1 illustrates that achieving good performance requires $R = 100$ observations, which is 10 times more than what was needed in Experiment 3.

A.6.2. Heuristic selection method for γ. Figure A.2 shows the trend of L_1 and L_2 as γ changes, where the candidate range of γ is discussed in Section A.9. From Figure A.2, we can see consistent trends of L_1 and L_2 across the designated number of clusters $\hat{K}$, indicating that the choice of γ using the proposed heuristic method is insensitive to the change in $\hat{K}$.

Figure A.3 shows the clustering performance as γ changes. We can see that the clustering performance slightly improves when γ increases from 10^{-4} to 10^{-2}. This is because the within-cluster

TABLE A.1. True values of $\{f_{k,m}(t) : k \in [K], m \in [M]\}$ in Experiment 4. The parameter ρ controls the distinctiveness across clusters, whose value is altered in the experiment. The functions $q_1(t)$ and $q_2(t)$ are defined in (2.28), $h_1(x) \equiv |\max(x, 0)|^{1/2}$, and $h_2(x) \equiv 1 + \min(x, 0)$.

	$m = 1$	$m = 2$
$k = 1$	$5.25 \times q_1(t)$	$5.25 \times q_2(t)$
$k = 2$	$6 \times [1 - h_1(2\rho - 1)] \times q_1(t)$ $+4.8 \times h_2(2\rho - 1) \times q_2(2[t - 0.8])$	$6 \times [1 + h_1(2\rho - 1)] \times q_2(t)$ $-4.8 \times h_2(2\rho - 1) \times q_2(2t)$
$k = 3$	$6.75 \times (1 + 0.5\rho) \times q_1(t)$	$6.75 \times (1 - 0.5\rho) \times q_2(t)$
$k = 4$	$7.5 \times (1 + \rho) \times q_1(t)$	$7.5 \times (1 - \rho) \times q_2(t)$

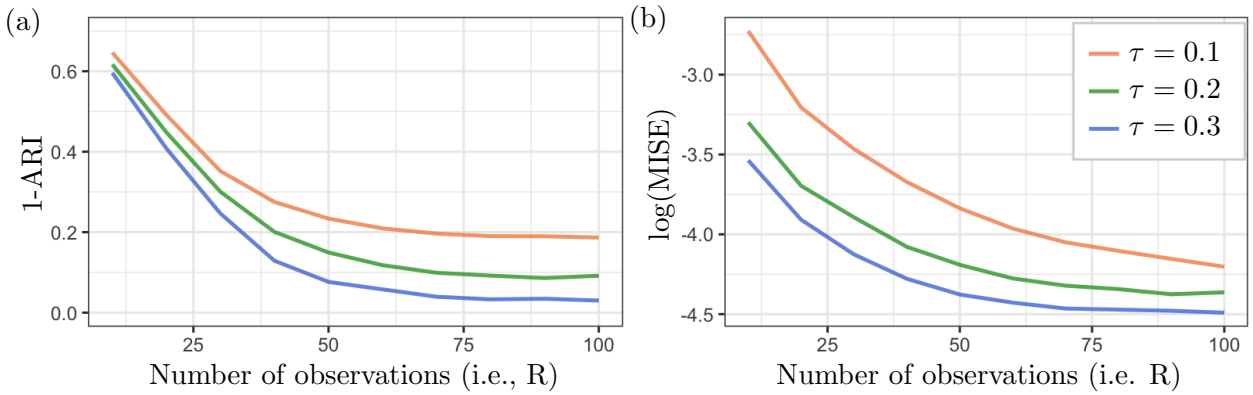

FIGURE A.1. Performance of proposed method in Experiment 4 with 1000 replicates. Synthetic data is generated with $K = 4$, $n = 40$, $\rho = 0.5$, and varying R and τ. Panel (a) and panel (b) display the clustering performance and intensity estimation performance, respectively.

heterogeneity of the distribution of event times remains unchanged (see Figure A.2(1a)), while the within-cluster heterogeneity of event counts decreases (see Figure A.2(1b)). Moreover, we can see from Figure A.3 that once γ exceeds 0.01, the clustering performance rapidly deteriorates. This is because the estimated clusters have increasing within-cluster heterogeneity of the distribution of event times when γ exceeds 0.01 (see Figure A.2(1a)). Therefore, by choosing the largest γ before observing a significant increase in L_1, we are able to achieve optimal clustering performance.

A.6.3. Sensitivity analysis with respect to ℓ_0. Figure A.4 shows how the performance of clustering and intensity estimation changes with different values of ℓ_0. Both criteria decrease when ℓ_0 increases from 2 to 8, because higher values of ℓ_0 allow the estimated intensities to capture more information in the distribution of event times. As ℓ_0 further increases, both 1-ARI and MISE

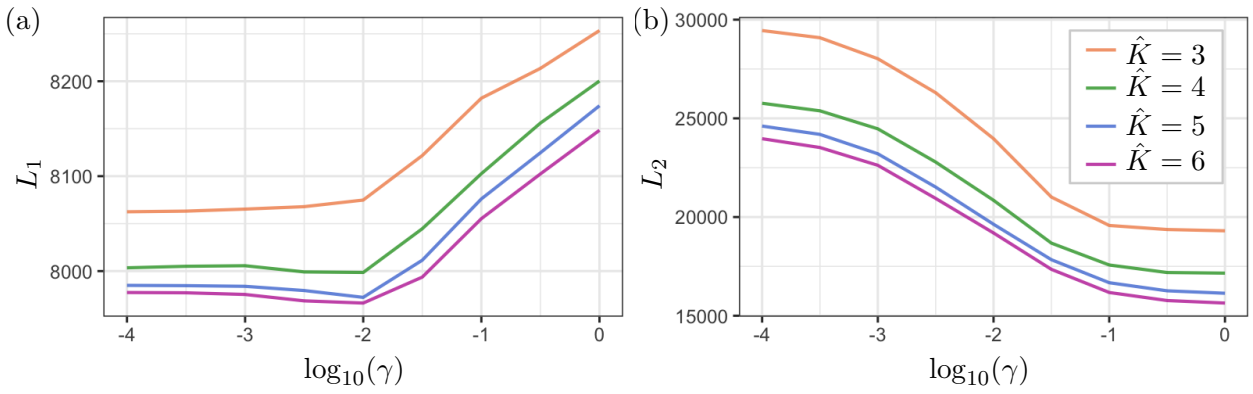

FIGURE A.2. Trends of L_1 and L_2 as γ changes averaged across 1000 replicates. Synthetic data is generated with $K = 4$, $n = 40$, $R = 3$, $\tau = 0.3$, $\rho = 0.5$. In the legend, $\hat{K}$ represents the designated number of clusters as input of the algorithm. It is evident that the trend of L_1 and L_2 is consistent across the designated number of clusters.

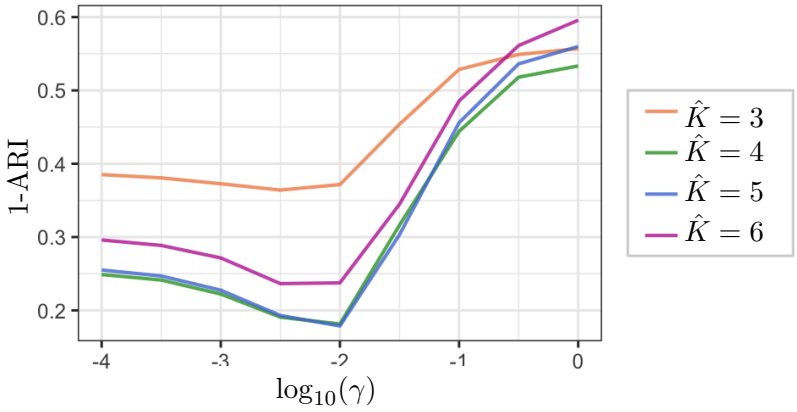

FIGURE A.3. Clustering performance as γ changes averaged across 1000 replicates. Synthetic data is generated with $K = 4$, $n = 40$, $R = 3$, $\tau = 0.3$, $\rho = 0.5$. In the legend, $\hat{K}$ represents the designated number of clusters as input of the algorithm.

stabilize, indicating that the estimation results are not significantly affected by further changes in ℓ_0 as long as ℓ_0 is sufficiently large to capture the signal in the distribution of event times, for instance, for the synthetic data in Figure A.4, $\ell_0 = 8$ is large enough.

A.6.4. Effectiveness of proposed initialization scheme. Figure A.5 shows the efficacy of proposed initialization scheme compared to a random initialization scheme. In the random initialization scheme, initial time shifts are set using $\text{Unif}(0, \hat{v}_{i,m}^{(0)})$, and initial cluster memberships randomly assigned. The results in Figure A.5 demonstrate that the proposed initialization method yields better estimation than the random initialization with multiple restarts. Moreover, keep

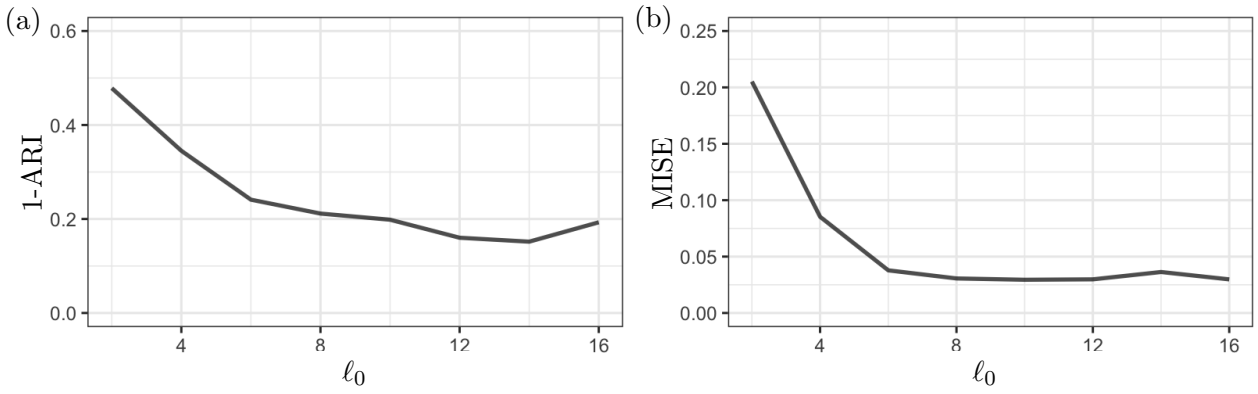

FIGURE A.4. Sensitivity analysis with respect to ℓ_0 averaged over 1000 replicates. Synthetic data is generated with $K = 4$, $n = 40$, $R = 3$, $\tau = 0.3$, $\rho = 0.5$. Panel (a) and (b) show that the performance of clustering and intensity component estimation become stable when $\ell_0 \geq 8$.

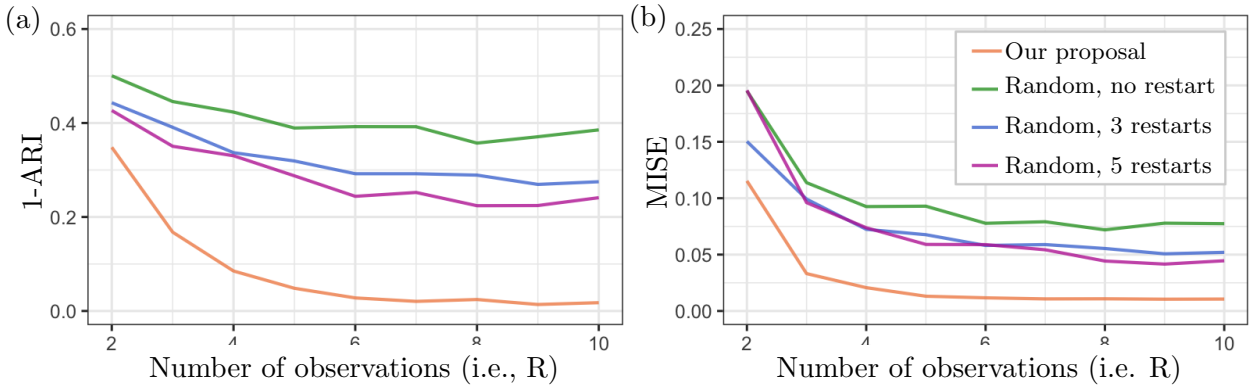

FIGURE A.5. Performance of proposed initialization scheme averaged over 1000 replicates. Synthetic data is generated with $K = 4$, $n = 40$, $\tau = 0.3$, $\rho = 0.5$, and varying R. In the legend, "our proposal" stands for the proposed initialization scheme, "random" stands for the random initialization scheme with restarts. For the random initialization scheme, time shifts are initialized using $\text{Unif}(0, \hat{v}_{i,m}^{(0)})$, and cluster memberships are initialized randomly. The best result among restarts is selected according to the smallest objective function. Panel (a) and (b) show that the proposed initialization scheme leads to better estimation than the random initialization scheme with 5 restarts.

increasing the number of restarts only results in diminishing improvement in performance. Overall, using proposed initialization scheme is computationally efficient and achieve promising estimation performance.

A.7. Supplement for Real Data Analysis

A.7.1. Data preprocessing.

TABLE A.2. Experimental conditions and total numbers of trials. In the "scenario" column, "L" and "R" denote higher contrast in the left and right gratings, respectively. The "choice" column shows "L" for moving the left grating towards the center and "R" for moving the right grating The "feedback" column shows "1" for reward and "-1" for penalty.

Scenario	Choice	Feedback	# trials
L	L	1	102
R	R	1	81
R	L	-1	8
L	R	-1	3

Experimental trials. Table A.2 presents the feedback types under different experimental conditions, along with the number of trials for each condition. The trials where the left visual grating was of higher contrast and the feedback was reward are analyzed using our proposed method. We refer to these trials as *training trials*. The rest of trials, which are not used to fit the proposed model, are employed to explore the roles of identified clusters under different tasks.

Selection of time window. Due to the design of the experiment, the duration of time between stimulus onset and feedback delivery across trials. In order to create comparable samples, we analyze a 3.5-second time window whose center is positioned at the midpoint between 0.1 seconds before the visual stimuli onset and 2 seconds after the feedback delivery. To be more specific, for observation $r \in \{1, \cdots, R\}$, the start and end time of observation r is set as

$$\text{ObsStartTime}_r \equiv \{(\text{VisTime}_r - 0.1) + (\text{FeedTime}_r + 2)\}/2 - 3.5/2, \tag{A.59}$$

$$\text{ObsEndTime}_r \equiv \{(\text{VisTime}_r - 0.1) + (\text{FeedTime}_r + 2)\}/2 + 3.5/2, \tag{A.60}$$

where ObsStartTime_r denotes the start time of the time window for trial r, VisTime_r denotes the visual stimulus onset time of trial r, FeedTime_r denotes the feedback delivery time of trial r, and ObsEndTime_r denotes the end time of the time window for trial r. In cases where the time window started within 1 second after the feedback delivery time of the previous trial, or ended after the visual stimulus onset time of the next trial, we augment the spike trains as detailed in the next section.

Here we analyze the firing activity within a $T = 3.5$ seconds time window due to the following reasons. Figure A.6 displays the average neural firing rate post feedback delivery over all training trials. It is evident that the average firing rate stabilizes at around 2 seconds post feedback delivery.

Moreover, among the 102 training trials, the maximum duration between visual stimulus onset and feedback delivery is 1.4 seconds. Therefore, by setting $T = 3.5$, we can ensure that each observation includes at least 0.1 second before visual stimulus onset and 2 seconds after feedback delivery.

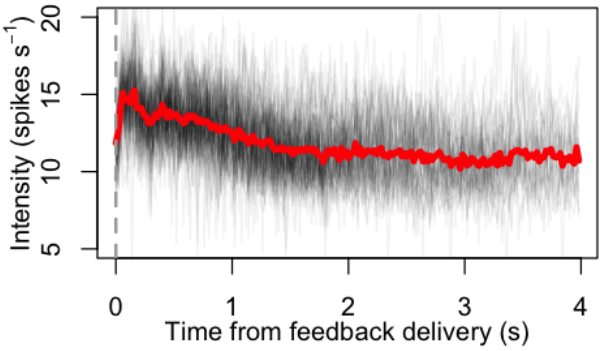

FIGURE A.6. Average neural firing intensities over neurons and trials. Each gray curve represents an average firing intensity of the midbrain region in a single training trial. All trials are aligned based on the feedback delivery time. The red curve represents the average firing intensity across all training trials. It is observed that the average firing intensity stabilizes approximately 2 seconds after feedback delivery.

Augmentation of spike trains. For trials with $\text{ObsStartTime}_r < \text{ConcluTime}_{r-1}$, where ConcluTime_{r-1} denotes the trial conclusion time (i.e., 1 second post feedback delivery time) of trial $r-1$, we augment the spike trains to impute the missing data. Figure A.7(a) shows that, for trials with $\text{ObsStartTime}_r < \text{ConcluTime}_{r-1}$, the average intensity remains stable within 0.4 second post the conclusion time of previous trial. Therefore, we augment spikes by making shifted copies of spikes within 0.4 second post ConcluTime_{r-1}. We append the following artificial spikes to the firing activity of neuron $i \in [n]$ in observation $r \in \{r : \text{ObsStartTime}_r < \text{ConcluTime}_{r-1}\}$:

$$\Big\{ t_{i,r,j} - 0.4 \times k : t_{i,r,j} \in [\text{ConcluTime}_{r-1}, \text{ConcluTime}_{r-1} + 0.4], \\ t_{i,r,j} - 0.4 \times k \in [\text{ObsStartTime}_r, \text{ConcluTime}_{r-1}], k \in \mathbb{N} \Big\}. \tag{A.61}$$

Additionally, we augment the spike trains for trials with $\text{ObsEndTime}_r > \text{VisTime}_{r+1}$. Figure A.7(b) shows that, for trials with $\text{ObsEndTime}_r > \text{VisTime}_{r+1}$, the average intensity remains stable within 0.4 seconds before the visual stimulus onset of the next trials. Therefore, we append the following spikes to the firing activity of neuron $i \in [n]$ in observation $r \in \{r : \text{ObsEndTime}_r >$

$\text{VisTime}_{r+1}\}$:

$$\Big\{t_{i,r,j} + 0.4 \times k : t_{i,r,j} \in [\text{VisTime}_{r+1} - 0.4, \text{VisTime}_{r+1}], \\ t_{i,r,j} + 0.4 \times k \in [\text{VisTime}_{r+1}, \text{ObsEndTime}_r], k \in \mathbb{N}\Big\} \tag{A.62}$$

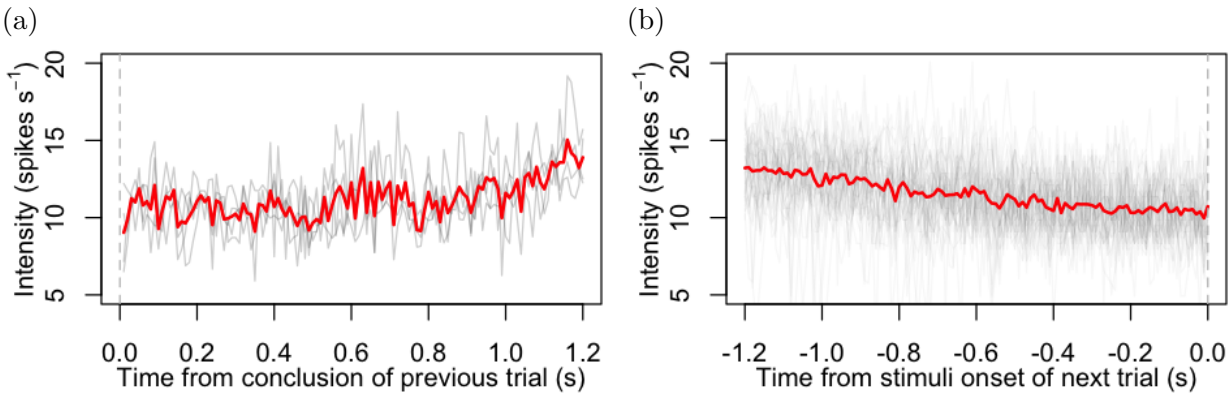

FIGURE A.7. Neural firing patterns after the previous trial's conclusion or before the next trial's visual stimulus onset. Each gray curve represents an average firing intensity of the midbrain region in one trial. The red curves represent the average of gray curves. Panel (a) shows the average neural firing intensities post conclusion of the previous trials for the trial set $\{r : \text{ObsStartTime}_r < \text{ConcluTime}_{r-1}\}$. The trials are aligned by the conclusion time of the previous trials. Panel (b) shows the average neural firing intensities prior visual stimuli onset of the next trials for the trial set $\{r : \text{ObsEndTime}_r > \text{VisTime}_{r+1}\}$. The trials are aligned by the visual stimulus onset time of the next trials.

A.7.2. Feasibility of proposed model on the neural data. We investigate the feasibility of the *shape invariant* assumption on the neural data in Steinmetz et al. [2019]. The assumption posits that neural firing intensities exhibit identical shapes but with subject-specific and trial-specific time shifts. Figure A.8 presents the empirical intensities of the four example neurons in Figure 2.1. It is evident from Figure A.8 that, for neuron 1 and neuron 2, their firing intensities exhibit similar shapes across trials when trials are aligned by stimulus onset time or feedback delivery time. In addition, for neuron 3 and neuron 4, their empirical firing intensities have similar shapes across trials when trials are aligned by the wheel movement onset time. Therefore, the assumption that neural firing intensities have similar shapes across trials, subject to trial-specific time shifts, appears to be feasible.

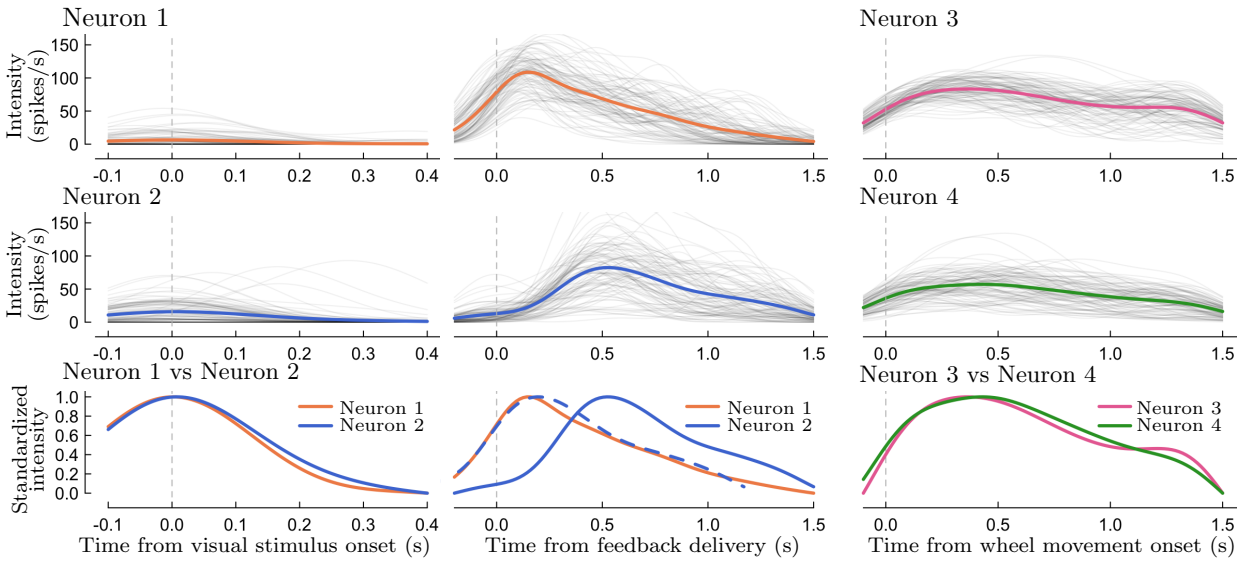

FIGURE A.8. Empirical firing intensities of the four example neurons in Figure 2.1. The first two rows present the empirical intensities for all training trials, with each gray line representing one trial. The thick colored lines indicate the average empirical intensity across all training trials. The third row displays the standardized average intensities, with line colors matching those of the thick lines in the first two rows. The third row displays the standardized average intensities. The colors of the lines are consistent with the colors of the thick lines in the first two rows. In the third row, the dashed blue line in the second column represents the manually shifted blue line, adjusted to best align with the orange line.

Furthermore, Figure A.8 presents a comparison of the standardized average empirical intensities between neuron 1 and neuron 2, as well as between neuron 3 and neuron 4. It is evident from the figure that the average empirical intensities of neuron 1 and neuron 2 exhibit similar shapes up to a scaling factor, as do the average empirical intensities of neuron 3 and neuron 4. The variation in the scales of firing intensities can be effectively addressed during the estimation stage by adjusting the value of γ. Consequently, it seems feasible to assume that neural firing intensities maintain similar shapes across neurons within the same cluster.

A.7.3. Tuning parameter selection. We choose the values of γ and K using the heuristic method proposed in Section 2.3.3. Figure A.9 presents the results of the heuristic method that informs our choice of γ and K. Firstly, we establish a preliminary estimate of K by applying the k-means algorithm on $N_i(T)$'s where $N_i(T) \equiv R^{-1} \sum_{r\in[R]} N_{i,r}(T)$, and selecting K using the elbow method. Figure A.9(a) illustrates that the within-cluster variance has only marginal reduction as the number of clusters exceeds 3. Therefore, we set $K = 3$ as a preliminary estimation of

K. Secondly, given the preliminary estimation of K, we choose the largest γ before observing a significant increase in L_1. Figure A.9(b) shows that the value of L_1 a significant upward trend once γ exceeds 10^{-4}. Therefore, we set $\gamma = 10^{-4}$. Finally, with γ fixed at 10^{-4}, we refine the value of K by identifying the elbow point on the curve of overall objective function (i.e., $L_1 + \gamma L_2$) against K. Figure A.9(c) suggests that $K = 3$ is a feasible choice.

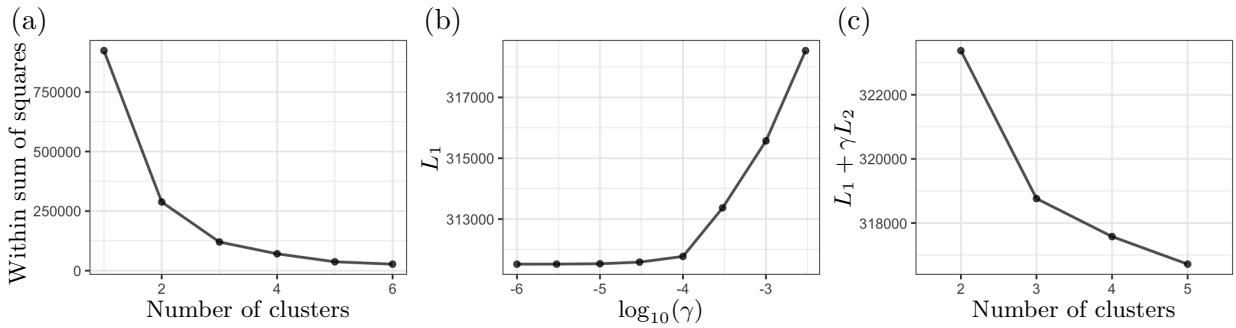

FIGURE A.9. Results of the heuristic method for selection of γ and K. Panel (a) shows the within cluster variance of $N_i(T)$'s obtained from the k-means algorithm against the number of clusters, where $N_i(T) \equiv R^{-1} \sum_{r \in [R]} N_{i,r}(T)$. Using the elbow method, we set $K = 3$ as a preliminary estimation of K. Panel (b) shows the values of L_1 as γ increases. The upward trend of L_1 when γ exceeds 10^{-4} suggests the choice of $\gamma = 10^{-4}$. Panel (c) shows the overall objective function value obtained from various numbers of clusters. The curve suggests a diminishing decrease in the objective function value when K exceeds 3.

A.7.4. Intensity component refinement. We refine the estimated intensity components by employing the proposed additive shape invariant model in (2.1) on each estimated cluster. For Cluster 1, we apply Algorithm 2.1 on spike trains of neurons in Cluster 1, denoted as $\mathcal{N}_1 \equiv \{N_{i,r}(t) : i \in \hat{\mathcal{C}}_1, r \in [R]\}$. We set $M = 1$ and let the observation-specific time shifts $w^*_{r,1}$'s to be the visual stimulus onset time. The rest of the parameters are set as follows: $K = 1$, $\gamma = 0$, $\ell_0 = 10$, $\epsilon = 0.005$. The algorithm is applied with 20 restarts, where each restart involves distinct initial values for subject-specific time shifts. These initial values are obtained by jittering the proposed initial subject-specific time shifts as follows: for $i \in \hat{\mathcal{C}}_1$, $m = 1$ and $x \in \{1, \cdots, 20\}$,

$$\tilde{v}^{(0)}_{i,m,x} \equiv \hat{v}^{(0)}_{i,m} + \varepsilon_{i,m,x}, \quad \varepsilon_{i,m,x} \sim \text{unif}(-T/50, T/50), \tag{A.63}$$

where $\tilde{v}^{(0)}_{i,m,x}$ represents the initial subject-specific time shifts associated subject i and stimulus m in restart x, and $\hat{v}^{(0)}_{i,m}$ represents the proposed initial subject-specific time shifts defined in (2.26).

The best result of intensity components among the 20 restarts is selected based on the smallest value of the objective function.

For Cluster 2 and 3, Algorithm 2.1 is applied to $\mathcal{N}_k \equiv \{N_{i,r}(t) : i \in \hat{\mathcal{C}}_k, r \in [R]\}$, $k \in \{2, 3\}$, in a similar manner as for $\mathcal{N}_1$. The only difference lies in the specification of $M = 2$ for Clusters 2 and 3, where $m = 1$ corresponds to the visual stimulus and $m = 2$ corresponds to the auditory tone cue.

A.7.5. Supplementary results of neural data analysis.

Neural firing intensity in condition "L,R". Figure A.10 displays the average firing patterns of the three clusters in four different experimental conditions. From the figure we can see that, there is high uncertainty in condition "L,R" since there are only 3 trials. However, the firing patterns in the "L,R" condition generally support our hypothesis regarding the roles of the clusters. For instance, in "L,R" trials, the firing rates of neurons in Cluster 1 show an upward trend before movement onset, which is consistent with our hypothesis that Cluster 1 neurons are responsible for executing the turning of the wheel. Moreover, the firing rates of neurons in Cluster 2 in "L,R" trials remain stationary after feedback delivery, aligning with our hypothesis that Cluster 2 neurons might respond to perceptions of stimuli such as rewards.

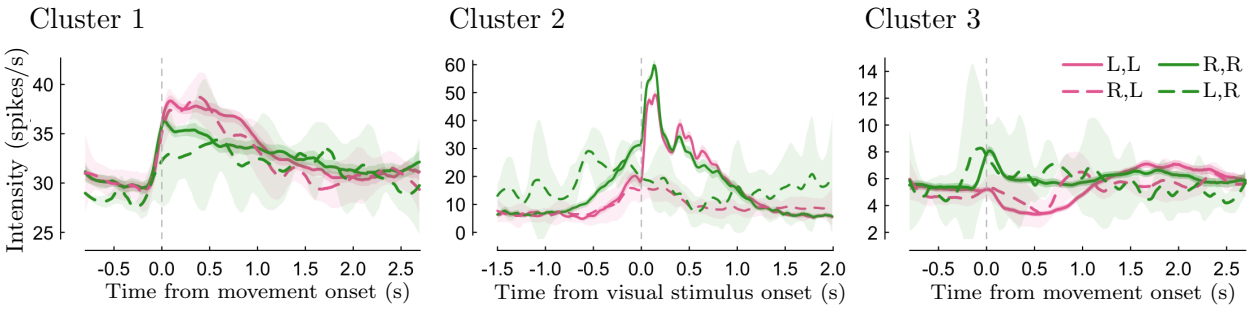

FIGURE A.10. Firing patterns of the three clusters in four different experimental conditions. The lines represent the mean firing intensities. The shaded regions represent the mean firing intensities plus or minus two standard errors of the mean. The legends represent "scenario, choice", for instance, "L,R" represents the trials where the left grating was of a higher contrast, and the mouse chose to move the right grating. In the "L,R" condition, there is high uncertainty due to the limited number of trials.

A.7.6. Results obtained using kCFC. For comparison, we apply the kCFC [Chiou and Li, 2007] to the neural data. Specifically, we apply the kCFC to the aggregated spike trains across the

TABLE A.3. Contingency table of the clusters from the proposed ASIMM and the clusters from the kCFC.

	ASIMM		
kCFC	Cluster 1	Cluster 2	Cluster 3
Cluster 1	60	2	71
Cluster 2	3	14	10
Cluster 3	0	7	58

training trials with $K = 3$, and set the remaining parameter values consistent with those specified in Section 2.5.3 of the main text. Table A.3 shows the association between clusters obtained from the kCFC and clusters obtained from the proposed ASIMM. From the table we see that the clusters from the kCFC are mixtures of the clusters from ASIMM. Particularly, for Cluster 1 and 2 identified by kCFC, approximately half of the neurons in these clusters are not grouped together in ASIMM's results.

Figure A.11 shows the firing patterns of clusters from the kCFC. We observe that firing patterns bear some resemblance to those obtained from ASIMM. However, Cluster 1 and 2 from the kCFC do not exhibit laterality. This lack of laterality is likely because Cluster 1 and 2 are mixtures of clusters from the ASIMM. For instance, the firing intensity of Cluster 1 from the ASIMM has a higher peak when the mouse chose the *left* visual grating (see Figure 2.6(1b)), whereas the firing intensity of Cluster 3 from the ASIMM has a higher peak when the mouse chose the *right* grating (see Figure 2.6(3b)). Consequently, when Cluster 1 and 3 from our proposed method are mixed, their individual lateral biases are counteracted, resulting in the absence of clear laterality.

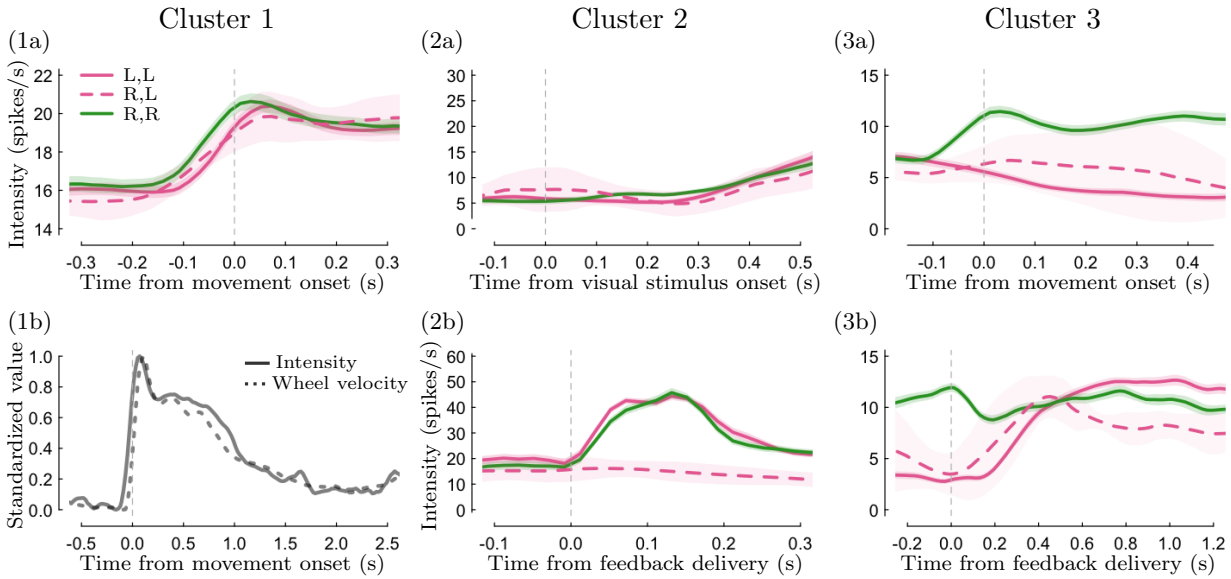

FIGURE A.11. Firing patterns of clusters from kCFC. The three columns correspond to the three estimated clusters. The shaded area represents the mean firing intensities plus or minus two standard errors of the mean. The legend in panel (1a) represents "scenario, choice", for instance, "R,L" represents the trials where the right grating was of a higher contrast, and the mouse chose to move the left grating. Panel (1b) illustrates the average firing intensity and wheel velocity, where both firing intensity and wheel velocity are standardized to range from 0 to 1. The firing intensities of Cluster 1 and 2 do not exhibit laterality.

A.8. Auxiliary Lemmas

LEMMA A.8.0.1. *Let* $\eta_m \equiv \exp\{-\mathrm{j}\,2\pi\xi w_m^*\}$ *and* $\boldsymbol{\eta} \equiv (\eta_m)_{m\in[M]}$, *where* $\xi \in \mathbb{R}\setminus\{0\}$. *If* $\{w_m^* \in \mathbb{C} : m \in [M]\}$ *are independent random variables, and among them, at least* $M-1$ *variables have non-zero variance, i.e.,* $|\{m \in [M] : \mathrm{var}(w_m^*) > 0\}| \geq M-1$, *then the matrix* $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]$ *is invertible.*

PROOF. Without loss of generality, we assume that $\mathrm{var}(w_m^*) > 0$ for $m = 2, \ldots, M$. Suppose $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]$ is not invertible. By the definition of invertible matrix, there exists $\mathbf{x} \in \mathbb{C}^M \setminus \{\mathbf{0}\}$ such that $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]\mathbf{x} = \mathbf{0}$. Thus we can derive that

$$\mathbb{E}[\overline{\mathbf{x}}^\top\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top\mathbf{x}] = 0. \tag{A.64}$$

Moreover, we have

$$\mathbb{E}[\overline{\mathbf{x}}^\top\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top\mathbf{x}] = \mathbb{E}\left[\left|\sum_{m\in[M]}\eta_m x_m\right|^2\right] \geq \mathrm{var}\left(\sum_{m\in[M]}\eta_m x_m\right) = \sum_{m\in[M]}|x_m|^2\,\mathrm{var}(\eta_m) \geq 0, \tag{A.65}$$

where the second equality follows from the assumption that $\{w_m^* : m \in [M]\}$ are independent. In order for (A.64) and (A.65) to hold simultaneously, we must have $|x_m|^2 \operatorname{var}(\eta_m) = 0$ for $m \in [M]$. From the assumption that $\operatorname{var}(w_m^*) > 0$ for $m = 2, \ldots, M$, we obtain that $\operatorname{var}(\eta_m) > 0$ for $m = 2, \ldots, M$. Thus, we know that $x_m = 0$ for $m = 2, \ldots, M$. Since $\mathbf{x} \neq 0$ by definition, we deduce that $x_1 \neq 0$. Hence, in order for $|x_1|^2 \operatorname{var}(\eta_1) = 0$ to hold we must have

$$\operatorname{var}(\eta_1) = 0. \tag{A.66}$$

Additionally, in order for (A.64) and (A.65) to hold simultaneously, we must have

$$\mathbb{E}\left[\left|\sum_{m\in[M]} \eta_m x_m\right|^2\right] = \operatorname{var}\left(\sum_{m\in[M]} \eta_m x_m\right),$$

or equivalently,

$$\left|\mathbb{E}\left(\sum_{m\in[M]} \eta_m x_m\right)\right|^2 = 0. \tag{A.67}$$

Plugging $x_m = 0$ for $m = 2, \ldots, M$ into (A.67), we obtain that

$$\mathbb{E}\left(\sum_{m\in[M]} \eta_m x_m\right) = \mathbb{E}(\eta_1) x_1. \tag{A.68}$$

Combining (A.67), (A.68), and that $x_1 \neq 0$, we deduce that

$$\mathbb{E}[\eta_1] = 0. \tag{A.69}$$

Combining (A.66) and (A.69), we can derive that $\eta_1 = 0$. However, we know that $|\eta_1| = 1$ by definition, implying that $\eta_1 = 0$ is impossible. Therefore, $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]$ must be invertible. □

LEMMA A.8.0.2. *Let $\eta_m \equiv \exp\{-\mathrm{j}\, 2\pi\xi w_m^*\}$ and $\boldsymbol{\eta} \equiv (\eta_m)_{m\in[M]}$, where $\xi \in \mathbb{R} \setminus \{0\}$. If $w_m^* = w_{m-1}^* + \delta_{m-1}$ for $m = 2, \ldots, M$, where $\{\delta_m \in \mathbb{C} : m \in [M-1]\}$ are independent random variables with non-zero variance, then the matrix $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]$ is invertible.*

PROOF. Suppose $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]$ is not invertible. By the definition of invertible matrix, there exists $\mathbf{x} \in \mathbb{C}^M \setminus \{\mathbf{0}\}$ such that $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]\mathbf{x} = \mathbf{0}$. Thus we can derive that

$$\mathbb{E}[\overline{\mathbf{x}}^\top \overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top \mathbf{x}] = 0. \tag{A.70}$$

On the other hand, we have

$$\begin{aligned}
\mathbb{E}[\overline{\mathbf{x}}^\top \overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top \mathbf{x}] &= \mathbb{E}\left|\sum_{m\in[M]} \eta_m x_m\right|^2 \\
&= \mathbb{E}\left|\sum_{m\in[M]} \exp\{-\mathrm{j}\,2\pi\xi w_m^*\} x_m\right|^2 \\
&= \mathbb{E}\left|\exp\{-\mathrm{j}\,2\pi\xi w_1^*\}\left[x_1 + \sum_{m=2}^{M} \exp\{-\mathrm{j}\,2\pi\xi(\delta_1+\cdots+\delta_{m-1})\} x_m\right]\right|^2 \\
&= \mathbb{E}\left|x_1 + \sum_{m=2}^{M} \exp\{-\mathrm{j}\,2\pi\xi(\delta_1+\cdots+\delta_{m-1})\} x_m\right|^2
\end{aligned} \tag{A.71}$$

where the second equation follows from the definition of η_m, the third equation follows from the assumption that $w_m^* = w_{m-1}^* + \delta_{m-1}$, and the last equation follows from the fact that $|\exp\{-\mathrm{j}\,\theta\}| = 1$ for any $\theta \in \mathbb{R}$. Define $A_m \equiv x_m + \sum_{m'=m+1}^{M} \exp\{-\mathrm{j}\,2\pi\xi(\delta_m + \cdots + \delta_{m'-1})\} x_{m'}$ for $m \in [M-1]$, and $A_M \equiv x_M$. Combining (A.70) and (A.71) we have $\mathbb{E}|A_1|^2 = 0$, which further implies that

$$A_1 = 0. \tag{A.72}$$

From the definition of A_m's we know that $A_m = x_m + \exp\{-\mathrm{j}\,2\pi\xi\delta_m\} A_{m+1}$. Thus (A.72) implies that $x_1 + \exp\{-\mathrm{j}\,2\pi\xi\delta_1\} A_2 = 0$, or equivalently,

$$A_2 = -x_1 \exp\{\,\mathrm{j}\,2\pi\xi\delta_1\}. \tag{A.73}$$

From the definition of A_2, we know that A_2 is a function of $\{\delta_2, \ldots, \delta_M\}$, which are independent to δ_1 by our assumption. Furthermore, δ_1 has a positive variance based on our assumption. Therefore, in order for (A.73) to hold, it must holds that $x_1 = 0$, or equivalently,

$$A_2 = 0. \tag{A.74}$$

Similarly, we can derive that $A_3 = ... = A_M = 0$. Therefore, using the definition of A_m's, we know that,

$$x_m = A_m - \exp\{-\mathrm{j}\,2\pi\xi\delta_m\} A_{m+1} = 0, \qquad m = 1, \cdots, M-1, \tag{A.75}$$

$$x_M = A_M = 0, \qquad m = M. \tag{A.76}$$

Equation (A.75) and (A.76) contradict the definition of $\mathbf{x}$ that asserts $\mathbf{x} \neq \mathbf{0}$. Therefore, $\mathbb{E}[\overline{\boldsymbol{\eta}}\boldsymbol{\eta}^\top]$ must be invertible.

□

LEMMA A.8.0.3. *Let $f \subseteq L^1(\mathbb{R})$ and $\hat{f}$ be its Fourier transform, and let $A \equiv \{x \in \mathbb{R} : f(x) \neq 0\}$ and $B \equiv \{\xi \in \mathbb{R} : \hat{f}(\xi) \neq 0\}$. Then*

$$m(A) < \infty \text{ and } m(B) < \infty \Rightarrow f = 0 \quad a.e. \ , \tag{A.77}$$

where m denotes the Lebesgue measure.

The proof of Lemma A.8.0.3 can be found in Benedicks [1985].

A.9. Reasonable Range of γ

To find a reasonable range for γ, we explore the magnitude of $\mathbb{E}[L_1(\mathbf{z}^*, \mathbf{a}^{*\prime}, \mathbf{f}^{*\prime}, \mathbf{v}^*)]$ and $\mathbb{E}[L_2(\mathbf{z}^*, \boldsymbol{\Lambda}^*)]$, where $\mathbf{z}^*, \boldsymbol{a}^{*\prime}, \mathbf{f}^{*\prime}, \mathbf{v}^*, \boldsymbol{\Lambda}^*$ denote the true parameters. The magnitude of $\mathbb{E}[L_1(\mathbf{z}^*, \mathbf{a}^{*\prime}, \mathbf{f}^{*\prime}, \mathbf{v}^*)]$ and $\mathbb{E}[L_2(\mathbf{z}^*, \boldsymbol{\Lambda}^*)]$ can be approximated as follows:

$$\mathbb{E}L_1(\mathbf{z}^*, \mathbf{a}^{*\prime}, \mathbf{f}^{*\prime}, \mathbf{v}^*) \lessapprox (nR)(T\Delta t)^{-1}, \tag{A.78}$$

$$\mathbb{E}L_2(\mathbf{z}^*, \boldsymbol{\Lambda}^*) \approx \sum_{i\in[n], r\in[R]} N_{i,r}(T). \tag{A.79}$$

The derivation of (A.78) and (A.79) is provided later in this section. Combining (A.78) and (A.79), we obtain that

$$\frac{\mathbb{E}L_1(\mathbf{z}^*, \mathbf{a}^{*\prime}, \mathbf{f}^{*\prime}, \mathbf{v}^*)}{\mathbb{E}L_2(\mathbf{z}^*, \boldsymbol{\Lambda}^*)} \lessapprox (nR)(T\Delta t)^{-1}\left\{\sum_{i\in[n], r\in[R]} N_{i,r}(T)\right\}^{-1} \equiv \gamma_0. \tag{A.80}$$

Consequently, we suggested to explore γ in the range $[10^{-5} \times \gamma_0, 10 \times \gamma_0]$.

Derivation of (A.78). From the definition of L_1 in (2.9) of the main text, we know that

$$\mathbb{E}L_1(\mathbf{z}^*, \mathbf{a}^{*\prime}, \mathbf{f}^{*\prime}, \mathbf{v}^*) = \sum_{i\in[n], r\in[R]} \mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{y_{i,r}(t)}{N_{i,r}(T)} - \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2\right) \tag{A.81}$$

where $\lambda^*_{i,r}(t) \equiv a^*_{z_i} + \sum_{m\in[M]} S^{v^*_{i,m}+w^*_{r,m}} f^*_{z^*_i,m}(t)$, and $\Lambda^*_{i,r}(T) \equiv \int_0^T \lambda^*_{i,r}(t)\mathrm{d}t$. Therefore, it suffices to show that

$$\mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{y_{i,r}(t)}{N_{i,r}(T)} - \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2\right) \lessapprox (T\Delta t)^{-1}. \tag{A.82}$$

To this end, we consider the following conditional expectation:

$$\begin{aligned}
&\mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{y_{i,r}(t)}{N_{i,r}(T)} - \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2 \middle| N_{i,r}(T)\right)\\
&= \mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{N_{i,r}(t+\Delta t)-N_{i,r}(t)}{N_{i,r}(T)\Delta t} - \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2 \middle| N_{i,r}(T)\right)\\
&= \mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{\sum_{j=1}^{N_{i,r}(T)}\mathbf{1}(t<t_{i,r,j}\le t+\Delta t)}{N_{i,r}(T)\Delta t} - \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2 \middle| N_{i,r}(T)\right)\\
&= \frac{N_{i,r}(T)}{T}\int_0^T \mathbb{E}\left(\left|\frac{\sum_{j=1}^{N_{i,r}(T)}\mathbf{1}(t<t_{i,r,j}\le t+\Delta t)}{N_{i,r}(T)\Delta t} - \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right|^2 \middle| N_{i,r}(T)\right)\mathrm{d}t,\\
&\equiv \frac{N_{i,r}(T)}{T}\int_0^T \mathbb{E}\left(\left|X(t)-Y(t)\right|^2 \middle| N_{i,r}(T)\right)\mathrm{d}t,
\end{aligned} \tag{A.83}$$

where the first equality follows from the definition of $y_{i,r}(t)$, the second equality follows from the definition of $N_{i,r}(t)$, and the third equality follows from the definition of L^2-norm. By definitions of $X(t)$ and $Y(t)$, we know that

$$\begin{aligned}
\mathbb{E}\left[X(t)|N_{i,r}(T)\right] &= \mathbb{E}\left(\frac{\sum_{j=1}^{N_{i,r}(T)}\mathbf{1}(t<t_{i,r,j}\le t+\Delta t)}{N_{i,r}(T)\Delta t}\middle| N_{i,r}(T)\right)\\
&= \mathbb{E}\left(\frac{\mathbf{1}(t<t_{i,r,1}\le t+\Delta t)}{\Delta t}\middle| N_{i,r}(T)\right)\\
&= \frac{\mathbb{P}\big(t<t_{i,r,1}\le t+\Delta t \mid N_{i,r}(T)\big)}{\Delta t}\\
&\approx \frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)} = Y(t)
\end{aligned} \tag{A.84}$$

where the first equality follows from the definition of $X(t)$. The second equality holds because, for Poisson processes, the event times $\{t_{i,r,j} : j = 1, \cdots, N_{i,r}(T)\}$ can be treated as i.i.d. samples given the total event count $N_{i,r}(T)$. The approximation in the fourth line follows from the definition of

$\lambda^*_{i,r}(t)$ and $\Lambda^*_{i,r}(T)$. Therefore we have

$$
\begin{aligned}
&\mathbb{E}\left(\left|X(t)-Y(t)\right|^2\middle|N_{i,r}(T)\right)\\
&=\operatorname{var}\left(X(t)\middle|N_{i,r}(T)\right)\\
&=\frac{1}{|\Delta t|^2}\left(\frac{\operatorname{var}\{\mathbf{1}(t<t_{i,r,j}\le t+\Delta t)\mid N_{i,r}(T)\}}{N_{i,r}(T)}\right)\\
&=\frac{1}{|\Delta t|^2}\left(\frac{\mathbb{P}\{t<t_{i,r,j}\le t+\Delta t\mid N_{i,r}(T)\}[1-\mathbb{P}\{t<t_{i,r,j}\le t+\Delta t\mid N_{i,r}(T)\}]}{N_{i,r}(T)}\right)\\
&\approx\frac{1}{|\Delta t|^2}\frac{1}{N_{i,r}(T)}\left(\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\Delta t\right)\left(1-\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\Delta t\right),
\end{aligned}
\tag{A.85}
$$

where the second equality follows from the definition of $X(t)$, the approximation in the fourth line follows from the definition of $\lambda^*_{i,r}(t)$ and $\Lambda^*_{i,r}(T)$.

Plugging (A.85) into (A.83), we can derive that

$$
\begin{aligned}
&\mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{y_{i,r}(t)}{N_{i,r}(T)}-\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|^2\middle|N_{i,r}(T)\right)\\
&\approx\frac{N_{i,r}(T)}{T}\int_0^T\frac{1}{|\Delta t|^2}\frac{1}{N_{i,r}(T)}\left(\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\Delta t\right)\left(1-\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\Delta t\right)\mathrm{d}t\\
&=\frac{1}{T}\left(\frac{1}{\Delta t}-\left\|\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|^2\right)\\
&\lessapprox(T\Delta t)^{-1}.
\end{aligned}
\tag{A.86}
$$

Finally, by taking the expectation of (A.86), we obtain

$$
\begin{aligned}
&\mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{y_{i,r}(t)}{N_{i,r}(T)}-\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2\right)\\
&=\mathbb{E}\left[\mathbb{E}\left(\frac{N_{i,r}(T)}{T}\left\|\frac{y_{i,r}(t)}{N_{i,r}(T)}-\frac{\lambda^*_{i,r}(t)}{\Lambda^*_{i,r}(T)}\right\|_t^2\middle|N_{i,r}(T)\right)\right]\\
&\lessapprox(T\Delta t)^{-1}.
\end{aligned}
\tag{A.87}
$$

Derivation of (A.79). Employing the definition of L_2 in (2.10) of the main text, along with the variance expression for the Poisson distribution, we can derive that

$$\text{(A.88)}\quad \mathbb{E}[L_2(\mathbf{z}^*, \mathbf{\Lambda}^*)] = \mathbb{E}\left[\sum_{i\in[n],r\in[R]} \left|N_{i,r}(T) - \Lambda^*_{i,r}(T)\right|^2\right] = \sum_{i\in[n],r\in[R]} \Lambda^*_{i,r}(T) \approx \sum_{i\in[n],r\in[R]} N_{i,r}(T).$$

APPENDIX B

Appendix for Chapter 3

B.1. Proof of Proposition 3.2.1

Proof: Without loss of generality, we set $T = 1$ and assume that $\{f_{q,k} : q, k \in [K]\}$ have supports on $[0, 1/2]$ and the time shifts $\{v_i : i \in \mathcal{V}\}$ take values in $[0, 1/2]$. When there is only one cluster, i.e., $K = 1$, identifiability follows from the classic argument in shape invariant model [Bigot and Gendre, 2013]. Therefore, we consider the case when $K \geq 2$ in this proof.

For each $q, k \in [K]$, we define the equivalence class of $f_{q,k}$ by the time shift as

$$\{f_{q,k}\}_{\mathbb{S}} \equiv \big\{S^w f_{q,k} : w \in \mathbb{R}\big\}, \tag{B.1}$$

where $S^w f_{q,k}(t) = f_{q,k}(t-w)$ as defined in the main text. Denote the the set of permutations of $[K]$ by $\mathfrak{S}_K$ that any $\sigma \in \mathfrak{S}_K$ maps $\{1, \cdots, K\}$ onto itself. With a slight abuse of notation, we define $\sigma^{-1} \circ \mathbf{f} \equiv \big(f_{\sigma^{-1}(q),\sigma^{-1}(k)}\big)_{K\times K}$ for $\sigma \in \mathfrak{S}_K$ and $\mathbf{f} = (f_{q,k})_{K\times K}$. We can then define the equivalence class of $(\mathbf{f}, \mathbf{z})$ by the permutation of cluster labels as

$$\{(\mathbf{f}, \mathbf{z})\}_{\mathfrak{S}_K} \equiv \big\{(\sigma^{-1} \circ \mathbf{f}, \sigma \circ \mathbf{z}) : \sigma \in \mathfrak{S}_K\big\},$$

where $\sigma \circ \mathbf{z} \equiv \big(\sigma(z_i)\big)_{i\in\mathcal{V}}$. Recalling the equivalent class introduced in (B.1), we can define

$$\{\mathbf{f}\}_{\mathbb{S}} \equiv \{(S^{w_{q,k}} f_{q,k})_{K\times K} : w_{q,k} \in \mathbb{R}, q, k \in [K]\}$$

to be the equivalence class of $\mathbf{f}$ by time shifts, and define

$$\{(\{\mathbf{f}\}_{\mathbb{S}}, \mathbf{z})\}_{\mathfrak{S}_K} \equiv \{(\sigma^{-1} \circ \mathbf{f}', \sigma \circ \mathbf{z}) : \sigma \in \mathfrak{S}_K, \mathbf{f}' \in \{\mathbf{f}\}_{\mathbb{S}}\}$$

to be the equivalence class of $(\mathbf{f}, \mathbf{z})$ by the permutation of cluster labels and time shifts.

Let $(\mathbf{f}^*, \mathbf{z}^*, \mathbf{v}^*)$ be the true parameters, and $(\mathbf{f}, \mathbf{z}, \mathbf{v})$ be arbitrary solutions to $\{\lambda_{i,j} = S^{\max(v_i,v_j)} f_{z_i,z_j} : i,j \in \mathcal{V}\}$. We first show that $(\mathbf{f}, \mathbf{z}) \in \{(\{\mathbf{f}^*\}_{\mathbb{S}}, \mathbf{z}^*)\}_{\mathfrak{S}_K}$. Since

$$S^{\max(v_i,v_j)} f_{z_i,z_j} = \lambda_{i,j} = S^{\max(v_i^*,v_j^*)} f^*_{z_i^*,z_j^*}, \quad i,j \in \mathcal{V}, \tag{B.2}$$

we know that, for any $i,j \in \mathcal{V}$,

$$\{f_{z_i,z_j}\}_{\mathbb{S}} = \{f^*_{z_i^*,z_j^*}\}_{\mathbb{S}}, \tag{B.3}$$

and hence

$$(\{f_{z_i,z_1}\}_{\mathbb{S}}, \cdots, \{f_{z_i,z_p}\}_{\mathbb{S}}) = (\{f^*_{z_i^*,z_1^*}\}_{\mathbb{S}}, \cdots, \{f^*_{z_i^*,z_p^*}\}_{\mathbb{S}}). \tag{B.4}$$

For i, i' such that $z_i^* \neq z_{i'}^*$, according to Condition C3.1, there exists $k \in [K]$ such that $\{f^*_{z_i^*,k}\}_{\mathbb{S}} \neq \{f^*_{z_{i'}^*,k}\}_{\mathbb{S}}$. Without loss of generality, let node i belongs to cluster k, i.e., $z_1^* = k$. Then, we know that $\{f^*_{z_i^*,z_1^*}\}_{\mathbb{S}} \neq \{f^*_{z_{i'}^*,z_1^*}\}_{\mathbb{S}}$. This implies that

$$\begin{aligned}(\{f_{z_i,z_1}\}_{\mathbb{S}}, \cdots, \{f_{z_i,z_p}\}_{\mathbb{S}}) &= (\{f^*_{z_i^*,z_1^*}\}_{\mathbb{S}}, \cdots, \{f^*_{z_i^*,z_p^*}\}_{\mathbb{S}}) \\ &\neq (\{f^*_{z_{i'}^*,z_1^*}\}_{\mathbb{S}}, \cdots, \{f^*_{z_{i'}^*,z_p^*}\}_{\mathbb{S}}) = (\{f_{z_{i'},z_1}\}_{\mathbb{S}}, \cdots, \{f_{z_{i'},z_p}\}_{\mathbb{S}}),\end{aligned}$$

where the first and last equalities are due to (B.4). Hence we know that if $z_i^* \neq z_{i'}^*$, then $z_i \neq z_{i'}$. Since clusters are non-empty by definition, we can always find K nodes with pairwise distinct true cluster memberships given by $\mathbf{z}^*$. Based on the previous argument, we know that those K nodes must have distinct cluster memberships given by $\mathbf{z}$. Thus both $\mathbf{z}$ and $\mathbf{z}^*$ include exactly K clusters, and we have $\mathbf{z} = \sigma \circ \mathbf{z}^*$ for some $\sigma \in \mathfrak{S}_K$. Plugging $\mathbf{z} = \sigma \circ \mathbf{z}^*$ into (B.3), we have $\{f_{\sigma(z_i^*),\sigma(z_j^*)}\}_{\mathbb{S}} = \{f^*_{z_i^*,z_j^*}\}_{\mathbb{S}}$ for any $i,j \in \mathcal{V}$, or equivalently $\{f_{q,k}\}_{\mathbb{S}} = \{f^*_{\sigma^{-1}(q),\sigma^{-1}(k)}\}_{\mathbb{S}}$ for any $q,k \in [K]$. To summarize, we have shown that $(\mathbf{f}, \mathbf{z}) = (\sigma^{-1} \circ \mathbf{f}', \sigma \circ \mathbf{z}^*)$ for some $\sigma \in \mathfrak{S}_K$ and $\mathbf{f}' \in \{\mathbf{f}^*\}_{\mathbb{S}}$, that is $(\mathbf{f}, \mathbf{z}) \in \{(\{\mathbf{f}^*\}_{\mathbb{S}}, \mathbf{z}^*)\}_{\mathfrak{S}_K}$. In the following proof, we use σ to denote the permutation such that $\mathbf{z} = \sigma \circ \mathbf{z}^*$.

Next, we will show that $(\mathbf{f}, \mathbf{z}) \in \{(\mathbf{f}^*, \mathbf{z}^*)\}_{\mathfrak{S}_K}$, that is, $f_{q,k} = f^*_{\sigma^{-1}(q),\sigma^{-1}(k)}$ for $q,k \in [K]$. The proof involves two steps: we first show that there exists one global shift w_0 such that $S^{w_0} f_{q,k} = f^*_{\sigma^{-1}(q),\sigma^{-1}(k)}$ for all $q,k \in [K]$; then, we can verify that the global shift $w_0 = 0$ under Condition C3.4.

For any pair $q, k \in [K]$, recall that combining (B.3) and $\mathbf{z} = \sigma \circ \mathbf{z}^*$ gives $\{f^*_{q,k}\}_{\mathbb{S}} = \{f_{\sigma(q),\sigma(k)}\}_{\mathbb{S}}$, thus $f^*_{q,k} = S^{w_{q,k}} f_{\sigma(q),\sigma(k)}$ for some $w_{q,k} \in \mathbb{R}$. Combining this equality with (B.2), we know that, for $i, j \in \mathcal{V}$,

$$\begin{aligned} S^{\max(v_i, v_j)} f_{z_i, z_j} &= S^{\max(v^*_i, v^*_j)} f^*_{z^*_i, z^*_j} \\ &= S^{\max(v^*_i, v^*_j)} S^{w_{z^*_i, z^*_j}} f_{\sigma(z^*_i), \sigma(z^*_j)} \\ &= S^{\max(v^*_i, v^*_j) + w_{z^*_i, z^*_j}} f_{z_i, z_j}. \end{aligned}$$

Hence, for any $q, k \in [K]$, if $f^*_{q,k} \not\equiv 0$, then $f_{\sigma(q),\sigma(k)} \not\equiv 0$ and

$$\max(v_i, v_j) = \max(v^*_i, v^*_j) + w_{q,k}, \quad i \in \mathcal{C}^*_q, j \in \mathcal{C}^*_k. \tag{B.5}$$

Furthermore, if $f^*_{q,k} \not\equiv 0$, we know that $\max\{v^*_j : j \in \mathcal{C}^*_k\} > \min\{v^*_i : i \in \mathcal{C}^*_q\}$ according to Condition C3.2. Let $j_0 \equiv \arg\max_{j \in \mathcal{C}^*_k}\{v^*_j\}$ be the node in the cluster k with largest true time shift, then there exists a node $i_0 \in \mathcal{C}^*_q$ such that $v^*_{j_0} > v^*_{i_0}$. Thus by (B.5) we know that

$$v_{j_0} \leq \max(v_{i_0}, v_{j_0}) = \max(v^*_{i_0}, v^*_{j_0}) + w_{q,k} = v^*_{j_0} + w_{q,k}. \tag{B.6}$$

Suppose $v_{j_0} < v^*_{j_0} + w_{q,k}$. In order for (B.5) to hold, we must have $\max(v_{i_0}, v_{j_0}) = v_{i_0}$, hence (B.6) gives $v_{i_0} = v^*_{j_0} + w_{q,k}$. Then, from (B.5), we know that for any $j' \in \mathcal{C}^*_k$,

$$\max(v^*_{i_0}, v^*_{j'}) + w_{q,k} = \max(v_{i_0}, v_{j'}) \geq v_{i_0} = v^*_{j_0} + w_{q,k}. \tag{B.7}$$

By definitions of j_0 and i_0, we know that $v^*_{j_0} > v^*_{i_0}$, thus $\max(v^*_{i_0}, v^*_{j'}) = v^*_{j'}$. Then by (B.7) we have $v^*_{j'} \geq v^*_{j_0}$, which contradicts with the definition of j_0. Therefore, it must hold that

$$v_{j_0} = v^*_{j_0} + w_{q,k}. \tag{B.8}$$

Applying the above argument to any triplet of clusters $q, k, l \in [K]$ with $q \neq l$, $f^*_{q,k} \not\equiv 0$ and $f^*_{k,l} \not\equiv 0$, we have

$$v_{j_0} = v^*_{j_0} + w_{q,k}; \qquad v_{j_0} = v^*_{j_0} + w_{k,l},$$

which indicates $w_{q,k} = w_{k,l}$. Therefore, for any $q, k, l \in [K]$ such that $f^*_{q,k} \not\equiv 0$ and $f^*_{k,l} \not\equiv 0$,

$$(f^*_{q,k}, f^*_{k,l}) = (S^{w_k} f_{\sigma(q),\sigma(k)}, S^{w_k} f_{\sigma(k),\sigma(l)}) \tag{B.9}$$

where $w_k \equiv w_{q,k} = w_{k,l}$.

Now we fix q, k and consider two arbitrary clusters $l, m \in [K]$ such that $\{q, k\} \cap \{l, m\} = \emptyset$, $f^*_{q,k} \not\equiv 0$ and $f^*_{l,m} \not\equiv 0$ ($q = k$ and/or $l = m$ is allowed). According to Condition C3.3, there exists a sequence of clusters $q_1, \cdots, q_a \in \{1, \cdots, K\}$ with $a \geq 2$, $q_1 = k$ and $q_a = l$, such that $f^*_{q_i,q_{i+1}} \not\equiv 0$ for $i = 1, \cdots, a-1$. Further denote $q_0 = q, q_{a+1} = m$. Then applying the result for triplets for each three consecutive clusters in the sequence yields

$$(f^*_{q,k}, f^*_{l,m}) = (S^{w_k} f_{\sigma(q),\sigma(k)}, S^{w_k} f_{\sigma(l),\sigma(m)}).$$

Since $l, m \in [K]$ are arbitrary, we obtain that, for any $l, m \in [K]$ such that $f^*_{l,m} \not\equiv 0$,

$$f^*_{l,m} = S^{w_0} f_{\sigma(l),\sigma(m)}, \tag{B.10}$$

where $w_0 \in \mathbb{R}$ is a global time shift.

According to Condition C3.4, we have $\inf\{t : f^*_{q,k}(t) > 0 \text{ for some } q, k \in [K]\} = 0$. Plugging in (B.10), we know that $\inf\{t : f_{\sigma(q),\sigma(k)}(t - w_0) > 0 \text{ for some } q, k \in [K]\} = 0$. Given Condition C3.4, we know that $w_0 = 0$. Thus for $q, k \in [K]$ that $f^*_{q,k} \neq 0$, we have

$$f^*_{q,k} = f_{\sigma(q),\sigma(k)}. \tag{B.11}$$

For $f_{q,k} \equiv 0$, the equality (B.11) naturally holds. Therefore we conclude that $f^*_{q,k} = f_{\sigma(q),\sigma(k)}$ for any $q, k \in [K]$, which suggests $(\mathbf{f}, \mathbf{z}) \in \{(\mathbf{f}^*, \mathbf{z}^*)\}_{\mathfrak{S}_K}$.

Finally we will show the identifiability of $\{v^*_i\}_{i \in \mathcal{V}}$ except for nodes in the set $\mathcal{A} \equiv \{i \in \mathcal{V} : v^*_i \leq v^*_j \text{ for any } j \in \mathcal{C}^*_{z^*_i}, \text{ or } v^*_i \leq v^*_j \text{ for any } j \in \mathcal{V} \text{ s.t. } \lambda^*_{i,j} \not\equiv 0\}$. The proof is similar to the steps for proving (B.8). From $f^*_{q,k} = f_{\sigma(q),\sigma(k)}$ for any $q, k \in [K]$, we know that, for any $i, j \in \mathcal{V}$ such that $\lambda^*_{i,j} \not\equiv 0$,

$$\max(v_i, v_j) = \max(v^*_i, v^*_j). \tag{B.12}$$

For any node $i \in \mathcal{A}^c$, we need to show that $v_i^* = v_i$. We know that there exists a j such that $v_i^* > v_j^*$ and $\lambda_{i,j} \not\equiv 0$, which implies that $f^*_{z_i^*, z_j^*} \neq 0$. Suppose that $v_i > v_i^*$, then we have from (B.12) that

$$v_i^* = \max(v_i^*, v_j^*) = \max(v_i, v_j) \geq v_i,$$

which contradicts $v_i > v_i^*$. Thus, we know that $v_i \leq v_i^*$.

Now suppose that $v_i < v_i^*$. In order for (B.12) to hold, we must have that $\max(v_i, v_j) = v_j$. Applying (B.12) again gives $v_j = v_i^*$. Then, from (B.12), we know that for $i' \in \mathcal{V}$ s.t. $\lambda_{i',j} \not\equiv 0$,

$$\max(v_{i'}^*, v_j^*) = \max(v_{i'}, v_j) \geq v_j = v_i^*. \tag{B.13}$$

We know by definition that $v_j^* < v_i^*$, and thus $\max(v_{i'}^*, v_j^*) = v_{i'}^*$. Therefore, it must hold that $v_{i'}^* \geq v_i^*$. This means that $v_i^* = \min\{v_{i'}^* : i' \in \mathcal{C}_{z_i^*}\}$. This contradicts the requirement that $v_i^* \neq \min\{v_{i'}^* : i' \in \mathcal{C}_{z_i^*}\}$ since $i \in \mathcal{A}^c$. Thus, we know that $v_i \geq v_i^*$. Combining this with $v_i \leq v_i^*$ gives $v_i = v_i^*$ for any $i \in \mathcal{A}^c$. □

B.2. Algorithm for Intensities-based Optimization Problem

In this section, we present an algorithm for solving SidSBM-P in main text. The algorithm is almost identical to the one for SidSBM-C except for the use of f and $\mathrm{d}N$.

Recall the optimization problem takes the following form

$$\begin{aligned}\left(\hat{\mathbf{f}}, \hat{\mathbf{z}}, \hat{\mathbf{v}}\right) &= \arg\min_{\Theta_{\mathrm{P}}} M_{\mathrm{P}}(\mathbf{f}, \mathbf{z}, \mathbf{v}; \gamma) \\ &\equiv \arg\min_{\Theta_{\mathrm{P}}} \Bigg[\sum_{(i,j)\in\mathcal{E}(T)} \frac{1}{T} \left\{ \int_0^T \left| \frac{S^{w_{i,j}} f_{z_i,z_j}(t)}{F_{z_i,z_j}(T)} \right|^2 \mathrm{d}t - 2 \int_0^T \frac{S^{w_{i,j}} f_{z_i,z_j}(t)}{F_{z_i,z_j}(T)} \mathrm{d}N_{i,j}(t) \right\} \\ &\qquad + \gamma \sum_{i,j\in\mathcal{V}, i\neq j} \left\{ N_{i,j}(T) - F_{z_i,z_j}(T) \right\}^2 \Bigg],\end{aligned} \tag{B.14}$$

where $w_{i,j} = \max(v_i, v_j)$, $F_{z_i,z_j}(T) \equiv \int_0^T f_{z_i,z_j}(s)\mathrm{d}s$, and $\Theta_{\mathrm{P}} \equiv \left\{(\mathbf{f}, \mathbf{z}, \mathbf{v}) : \mathbf{f} \in \mathcal{B}_{1,\ell_0}^{K\times K}, \mathbf{z} \in [K]^p, F_{z_i,z_j}(T) \neq 0 \text{ for } (i,j) \in \mathcal{E}(T), \mathbf{v} \in [0, T/2]^p\right\}$. The algorithm alternates between two steps:

$$(\textit{centering} \text{ step}) \quad \hat{\mathbf{f}}, \hat{\mathbf{v}} = \arg\min_{\mathbf{f},\mathbf{v}} M_{\mathrm{P}}(\mathbf{f}, \hat{\mathbf{z}}, \mathbf{v}; \gamma), \tag{B.15}$$

$$(\textit{clustering} \text{ step}) \quad \hat{\mathbf{z}} = \arg\min_{\mathbf{z}} M_{\mathrm{P}}(\hat{\mathbf{f}}, \mathbf{z}, \hat{\mathbf{v}}; \gamma). \tag{B.16}$$

In the centering step (B.15), we update the connecting intensities $\hat{\mathbf{f}}$ and the time shifts $\hat{\mathbf{v}}$ conditioning on the current estimates of cluster memberships $\hat{\mathbf{z}}$. Given cluster memberships $\hat{\mathbf{z}}$, the optimization problem (B.15) can be decoupled into two parts. In the first part, we have the analytical form of $\hat{f}_{q,k}$ as a function of $\hat{\mathbf{v}}$

$$\hat{f}_{q,k}(t) = \sum_{i\in\hat{\mathcal{C}}_q, j\in\hat{\mathcal{C}}_k, i\neq j} \frac{S^{-\max(\hat{v}_i,\hat{v}_j)}\hat{f}_{\ell_0,i,j}(t)}{|\hat{\mathcal{C}}_q||\hat{\mathcal{C}}_k| - \mathbf{I}(q=k)|\hat{\mathcal{C}}_q|}, \qquad 1\le q,k\le K, \tag{B.17}$$

where $\hat{f}_{\ell_0,i,j}(t) \equiv \sum_{|\ell|\le\ell_0} \overline{\psi_\ell(t)}\ T^{-1}\int_0^T \psi_\ell(t)\mathrm{d}N_{i,j}(t)$ with $\psi_\ell(t)\equiv\exp(-\mathbf{i}2\pi\ell t/T)$ is the projection of point process $\mathrm{d}N_{i,j}$ on $\mathcal{B}_{1,\ell_0}$, and $\hat{\mathcal{C}}_k = \{i\in\mathcal{V}:\hat{z}_i = k\}$ is the estimated k-th cluster. In the second part, we solve for $\hat{\mathbf{v}}$ from

$$\hat{\mathbf{v}} = \arg\min_{\mathbf{v}} \sum_{i\in\mathcal{V}}\sum_{k\in[K]} \left(|\hat{J}_{i,k}|\overline{N}_{i,k}(T)\frac{1}{T}\left\{ \int_0^T \left|\frac{S^{v_i}\hat{f}_{z_i,k}(t)}{\hat{F}_{z_i,k}(T)}\right|^2 \mathrm{d}t - 2\int_0^T \frac{S^{v_i}\hat{f}_{z_i,k}(t)}{\hat{F}_{z_i,k}(T)}\frac{\mathrm{d}\overline{N}_{i,k}(t)}{\overline{N}_{i,k}(T)}\right\}\right), \tag{B.18}$$

where $\hat{f}_{z_i,k}(t)$'s are treated as functions of $\hat{\mathbf{v}}$ as shown in (B.17), $\hat{J}_{i,k}\equiv\{j\in\mathcal{V}:\hat{z}_j=k, j\neq i, v_j\le v_i\}$ is a collection of nodes in k-th cluster whose time shifts are no larger than node i, $\mathrm{d}\overline{N}_{i,k}(t)\equiv|\hat{J}_{i,k}|^{-1}\sum_{j\in\hat{J}_{i,k}}\mathrm{d}\overline{N}_{i,j}(t)$ is the empirical intensity over $\hat{J}_{i,k}$, and $\overline{N}_{i,k}(T)\equiv|\hat{J}_{i,k}|^{-1}\sum_{j\in\hat{J}_{i,k}}N_{i,j}(T)$ is the proportion of connections between node i and nodes in $\hat{J}_{i,k}$. Problem (B.18) can be solved using a gradient descent algorithm. Detailed forms of the gradients are given in Section B.3. Finally, we pin down the global shift $\hat{c} = \min\{t_{i,j} - \max(\hat{v}_i,\hat{v}_j) : i,j\in\mathcal{V}\}$, which is an approximation of $\inf\{t\in[0,T]:\hat{f}_{q,k}(t)>0 \text{ for some } q,k\in[K]\}$. Specifically, we update $\hat{\mathbf{f}}$ and $\hat{\mathbf{v}}$ as follows

$$\hat{f}_{q,k}\leftarrow S^{-\hat{c}}\hat{f}_{q,k},\ q,k\in[K]; \quad \hat{v}_i\leftarrow\hat{v}_i+\hat{c},\ i\in\mathcal{V}. \tag{B.19}$$

In the clustering step (B.16), we update each cluster membership $\hat{z}_i$ conditioning on the current $\hat{\mathbf{f}}$, $\hat{\mathbf{v}}$ and $\{\hat{z}_j\}_{j\in\mathcal{V}\setminus\{i\}}$. The optimization problem (B.14) reduces to, for $i\in\mathcal{V}$,

$$\begin{aligned}\hat{z}_i = \arg\min_{z_i\in[K]} \sum_{k\in[K]} \Bigg[&|\hat{\mathcal{C}}_{k,-i}|\overline{F}_{i,k}(T)\frac{1}{T}\left\{\int_0^T\left|\frac{\hat{f}_{z_i,k}(t)}{\hat{F}_{z_i,k}(T)}\right|^2\mathrm{d}t - 2\int_0^T\frac{\hat{f}_{z_i,k}(t)}{\hat{F}_{z_i,k}(T)}\frac{\mathrm{d}\overline{F}_{i,k}(t)}{\overline{F}_{i,k}(T)}\right\}\\ &+\gamma|\hat{\mathcal{C}}_{k,-i}|\left\{\overline{F}_{i,k}(T)-\hat{F}_{z_i,k}(T)\right\}^2\Bigg],\end{aligned} \tag{B.20}$$

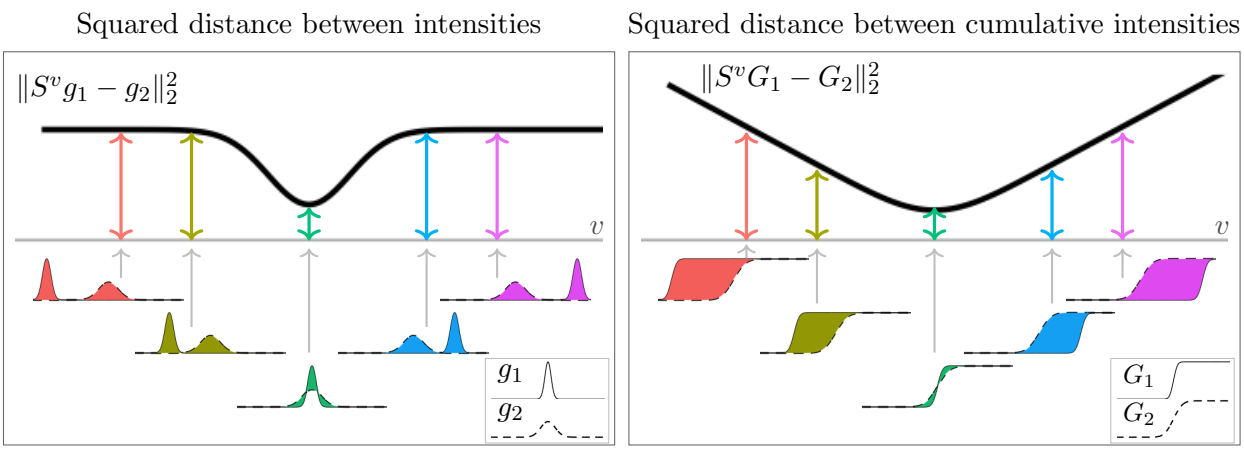

FIGURE B.1. Squared distances between two intensities or cumulative intensities as a function of time shift. The thick solid curves in the figure depict the values of squared distances, with the horizontal axes representing the time shift. The shadow regions below the time shift axes indicate the differentiation between the two intensities or cumulative intensities that contributes to the squared distances. In the left panel, a constant squared distance is observed when the time shift is large enough such that the intensities do not overlap. In contrast, the squared distance in the right panel shows no plateaus.

where $\hat{\mathcal{C}}_{k,-i} \equiv \hat{\mathcal{C}}_k \backslash \{i\}$, $\mathrm{d}\overline{F}_{i,k}(t) \equiv \left|\hat{\mathcal{C}}_{k,-i}\right|^{-1} \sum_{j \in \hat{\mathcal{C}}_{k,-i}} S^{-\max(\hat{v}_i, \hat{v}_j)} \mathrm{d}N_{i,j}(t)$ for $t \in [0, T]$ and $\overline{F}_{i,k}(T) \equiv \left|\hat{\mathcal{C}}_{k,-i}\right|^{-1} \sum_{j \in \hat{\mathcal{C}}_{k,-i}} S^{-\max(\hat{v}_i, \hat{v}_j)} N_{i,j}(T)$.

We can employ the same initialization scheme and a similar convergence criterion as those for SidSBM-C.

REMARK B.2.1. *It is worthwhile to mention that the objective function in* (B.18) *might suffer from plateaus. This can be illustrated by considering a simple scenario where we estimate the time shift between two intensities using the squared error. In this case, we aim to solve the problem* $\hat{v} = \arg\min_v |S^v g_1 - g_2|_2^2$, *where* g_1 *and* g_2 *are the two given intensities. The left panel of Figure B.1 shows that the value of* $|S^v g_1 - g_2|_2^2$ *remains constant when the value of* v *becomes sufficiently large such that there is no overlap between* $S^v g_1$ *and* g_2. *In practice, this plateau pattern can prevent the algorithm for SidSBM-P from finding the global minimum. As shown in the right panel of Figure B.1, the squared distance between cumulative intensities does not have any plateaus, which makes it easier for the algorithm for SidSBM-C to find the global minimum.*

B.3. Details of the Centering Step

In the centering step for SidSBM-C, we fix $\hat{\mathbf{z}}$ and update the cumulative intensities $\hat{\mathbf{F}}$ and time shifts $\hat{\mathbf{v}}$. The key observation is that when $\hat{\mathbf{v}}$ and $\hat{\mathbf{z}}$ are given, $\hat{\mathbf{F}}$ has a closed-form solution; and when $\hat{\mathbf{F}}$ and $\hat{\mathbf{z}}$ are given, $\hat{\mathbf{v}}$ can be estimated using gradient descent algorithm.

Given time shifts $\hat{\mathbf{v}}$, we update $\hat{\mathbf{F}}$ using Equation 3.11 in the main text, that is

$$\hat{F}_{q,k}(\cdot) = \sum_{i\in\hat{\mathcal{C}}_q, j\in\hat{\mathcal{C}}_k, i\neq j} \frac{S^{-\max(\hat{v}_i,\hat{v}_j)} N_{i,j}(\cdot)}{|\hat{\mathcal{C}}_q|\cdot|\hat{\mathcal{C}}_k| - \mathbf{I}(q=k)\cdot|\hat{\mathcal{C}}_q|}, \quad 1\le q,k\le K. \tag{B.21}$$

Given intensities $\hat{\mathbf{F}}$, we update $\hat{\mathbf{v}}$ by solving the optimization problem in (3.12), that is

$$\begin{aligned} \hat{v_i} &= \arg\min_{v_i} R_i(v_i) \\ &\equiv \arg\min_{v_i} \sum_{k\in[K]} \left(\left|\hat{J}_{i,k}\right| \cdot \overline{N}_{i,k}(T) \cdot \frac{1}{T}\int_0^T \left| \frac{\overline{N}_{i,k}(t)}{\overline{N}_{i,k}(T)} - \frac{S^{v_i}\hat{F}_{\hat{z}_i,k}(t)}{\hat{F}_{\hat{z}_i,k}(T)} \right|^2 \mathrm{d}t \right), \quad i\in\mathcal{V}. \end{aligned} \tag{B.22}$$

Let $\{\beta_{\ell,q,k} : \beta_{\ell,q,k} = T^{-1}\int_0^T \exp\{-\mathbf{i}2\pi\ell(s/T)\}\hat{F}_{q,k}(s)/\hat{F}_{q,k}(T)\mathrm{d}s, \ell\in\mathbb{Z}\}$ be the Fourier coefficients of $\hat{F}_{q,k}(\cdot)/\hat{F}_{q,k}(T)$. We can verify that Fourier coefficients of $S^{v_i}\hat{F}_{q,k}(\cdot)/\hat{F}_{q,k}(T)$ are, for $\ell\in\mathbb{Z}$,

$$\beta'_{\ell,q,k} = \exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,q,k} + \frac{\mathbf{I}(\ell\neq 0)}{\mathbf{i}2\pi\ell}\right) - \frac{1}{\mathbf{i}2\pi\ell}\mathbf{I}(\ell\neq 0) - \frac{v_i}{T}\mathbf{I}(\ell=0). \tag{B.23}$$

Further let $\{\alpha_{\ell,i,k} : \alpha_{\ell,i,k} = T^{-1}\int_0^T \exp\{-\mathbf{i}2\pi\ell(s/T)\}\overline{N}_{i,k}(s)/\overline{N}_{i,k}(T)\mathrm{d}s, \ell\in\mathbb{Z}\}$ be the Fourier coefficients of $\overline{N}_{i,k}(\cdot)/\overline{N}_{i,k}(T)$. From Parseval's theorem, we know that

$$R_i(v_i) = \sum_{k\in[K]} \left(\left|\hat{J}_{i,k}\right| \cdot \overline{N}_{i,k}(T) \cdot \sum_{l=-\infty}^{\infty} \left|\alpha_{\ell,i,k} - \beta'_{\ell,\hat{z}_i,k}\right|^2 \right).$$

Plugging in (B.23), we can see that

$$\begin{aligned} R_i(v_i) = &\sum_{k\in[K]} \left|\hat{J}_{i,k}\right| \overline{N}_{i,k}(T) \sum_{l\neq 0} \left| \alpha_{\ell,i,k} - \exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right) + \frac{1}{\mathbf{i}2\pi\ell} \right|^2 \\ &+ \sum_{k\in[K]} \left|\hat{J}_{i,k}\right| \overline{N}_{i,k}(T) \left| \alpha_{0,i,k} - \beta_{0,\hat{z}_i,k} + \frac{v_i}{T} \right|^2, \end{aligned}$$

where we separate the inner summation by whether ℓ is zero or not. For $\ell \neq 0$, we have the following gradient

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}v_i}\left|\alpha_{\ell,i,k} - \exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right) + \frac{1}{\mathbf{i}2\pi\ell}\right|^2 \\
&= \frac{\mathrm{d}}{\mathrm{d}v_i}\Bigg\{\left|\alpha_{\ell,i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right|^2 + \left|\exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)\right|^2 \\
&\qquad\qquad - 2\mathrm{Re}\left[\overline{\left(\alpha_{\ell,i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)}\exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)\right]\Bigg\} \\
&= -2\frac{\mathrm{d}}{\mathrm{d}v_i}\mathrm{Re}\left[\overline{\left(\alpha_{\ell,i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)}\exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)\right] \\
&= 2\mathrm{Re}\left[\frac{\mathbf{i}2\pi\ell}{T}\exp(-\mathbf{i}2\pi\ell v_i/T)\overline{\left(\alpha_{\ell,i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)}\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)\right].
\end{aligned}$$

Thus, we arrive at

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}v_i}R_i(v_i) = &\sum_{k\in[K]}\sum_{\ell\in\mathbb{Z},\ell\neq 0} 2|\hat{J}_{i,k}|\overline{N}_{i,k}(T)\mathrm{Re}\left[\frac{\mathbf{i}2\pi\ell}{T}\cdot\exp(-\mathbf{i}2\pi\ell v_i/T)\left(\beta_{\ell,\hat{z}_i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)\overline{\left(\alpha_{\ell,i,k} + \frac{1}{\mathbf{i}2\pi\ell}\right)}\right] \\
&- \sum_{k\in[K]} 2|\hat{J}_{i,k}|\overline{N}_{i,k}(T)\left(\beta_{0,\hat{z}_i,k} - \alpha_{0,i,k} - \frac{v_i}{T}\right)\frac{1}{T}. \qquad \text{(B.24)}
\end{aligned}$$

The gradient given in (B.24) is used in the gradient descent algorithm for solving (B.22). Note that the gradient includes a sum of infinite terms $\ell \in \mathbb{Z}$. In practice, we approximate the infinite summation with the sum over $\{\ell : |\ell| \leq L/2\}$, where L is the length of the equispaced time grid on which $\overline{N}_{i,k}$'s and $\hat{F}_{q,k}$'s are evaluated. Furthermore, the gradient is divided by $\sum_k |\hat{J}_{i,k}| \cdot \overline{N}_{i,k}(T)$ in order to make the magnitude of gradient, hence the learning rate, invariant to the number of nodes.

When applying the gradient descent algorithm for (B.22), we set the initialize time shifts as the current estimation of time shifts. The algorithm is stopped when the decrease of loss function value between two consecutive iterations is less than a given threshold

$$\frac{|R_i(\hat{v}_i^{(s+1)}) - R_i(\hat{v}_i^{(s)})|}{R_i(\hat{v}_i^{(s)})} < \varepsilon = 10^{-4}, \qquad \text{(B.25)}$$

or when the maximal number of iterations is reached. Here (s) and $(s+1)$ are the iteration index. The maximal number of iterations is set as 1000 in our experiments. We remark that the

convergence criterion (B.25) is only for solving (B.22), and is different from the criterion for the entire algorithm (i.e., Equation (3.18) in the main text).

To summarize, in the centering step, the algorithm iterates between (B.21) and (B.22) until the change of estimation of cumulative intensities is less than a given threshold

$$\frac{\|\hat{\mathbf{F}}^{(s+1)} - \hat{\mathbf{F}}^{(s)}\|_2}{\|\hat{\mathbf{F}}^{(s)}\|_2} < \varepsilon = 10^{-3}, \tag{B.26}$$

or when the maximal number of iterations is reached. Here (s) and $(s+1)$ are the iteration index. The maximal number of iterations is set as 10 in our experiments.

In a similar manner, we can show that the centering step for SidSBM-P is almost identical to the one for SidSBM-C except for the gradients. Let $c_{\ell,i,j}$ denote the Fourier coefficient of $N_{i,j}$, i.e., $c_{\ell,i,j} \equiv T^{-1}\int_0^T \exp\{-\mathbf{i}2\pi\ell s/T\}\mathrm{d}N_{i,j}(s)$, $\ell \in \mathbb{Z}$. The optimization problem (B.18) in the centering step reduces to

$$\begin{aligned}\hat{\mathbf{v}} &= \arg\min_{\mathbf{v}} R(\mathbf{v}) \\ &\equiv \arg\min_{\mathbf{v}} \sum_{i\in\mathcal{V}}\sum_{k\in[K]}\left(|\hat{J}_{i,k}|\cdot\overline{N}_{i,k}(T)\cdot\sum_{|\ell|\le\ell_0}\left|\overline{c}_{\ell,i,k}\exp(\mathbf{i}2\pi\ell v_i/T) - \phi_{\ell,\hat{z}_i,k}\right|_2^2\right),\end{aligned} \tag{B.27}$$

where

$$\overline{c}_{\ell,i,k} = \frac{\sum_{j\in\hat{J}_{i,k}:N_{i,j}(T)=1} c_{\ell,i,j}}{\#\{j\in\hat{J}_{i,k}: N_{i,j}(T)=1\}}$$

and

$$\phi_{\ell,q,k} = \frac{\sum_{i\in\hat{\mathcal{C}}_q, j\in\hat{\mathcal{C}}_k, N_{i,j}(T)=1, i\neq j} c_{\ell,i,j}\exp\{\mathbf{i}2\pi\ell\max(v_i,v_j)/T\}}{\#\{(i,j): i\in\hat{\mathcal{C}}_q, j\in\hat{\mathcal{C}}_k, N_{i,j}(T)=1, i\neq j\}}.$$

We can show that, for $j\in\mathcal{V}$,

$$\frac{\partial}{\partial v_j}R(\mathbf{v}) = \sum_{k\in[K]}\sum_{|\ell|\le\ell_0} 2\mathrm{Re}\left[\frac{\mathbf{i}2\pi\ell}{T}|\hat{J}_{j,k}|\overline{N}_{j,k}(T)\overline{\overline{c}_{\ell,j,k}\exp(\mathbf{i}2\pi\ell v_j/T)}\phi_{\ell,\hat{z}_j,k}\right], \tag{B.28}$$

where the detailed derivation is deferred to Section B.7.

B.4. Additional Simulation Results

B.4.1. Heuristic criterion for γ. In practice, the following heuristic criterion can be employed to obtain a preliminary estimation of γ. Initially, a small value of γ can be utilized to obtain

TABLE B.1. Connecting probabilities for synthetic data in Scenario B and Scenario C. Each column corresponds to one pair of (q, k).

	$(1,1)$	$(1,2)$	$(1,3)$	$(2,2)$	$(2,3)$	$(3,3)$
$\alpha_{q,k}$	0.10	0.10	0.50	0.50	0.90	0.90

a rough assessment of the shapes and scales. Subsequently, we can explore larger values of γ if the estimated scales exhibit significant variations among clusters, and otherwise try smaller γs to achieve better clustering performance.

We explore the performance of the proposed heuristic selection criterion of γ through three distinct simulated data scenarios. The connecting intensities are defined as $f_{q,k} = \alpha_{q,k} \times \Gamma(\mu_{q,k}^2/\sigma_{q,k}^2, \mu_{q,k}/\sigma_{q,k}^2)$, where $\mu_{q,k}$ and $\sigma_{q,k}$ are controlled by β, as explained in Table 3.1 in the main text. The specific values of $\alpha_{q,k}$ and β will be provided later. For all scenarios, we set $K = 3$, $p = 30$, and $W = 80$. In Scenario A, all $\alpha_{q,k}$ values are set to $\alpha_{q,k} = 0.9$ for $q, k \in [K]$, while β is assigned as $\beta = 1.9$. This means that the intensities have different shapes but share identical scales across clusters. Figure B.2 depicts the performance of SidSBM-C under varying values of γ in this scenario. The results indicate that small values of γ lead to the best performance. In Scenario B, we vary $\alpha_{q,k}$ across clusters as presented in Table B.1, while fixing $\beta = 1.9$. Consequently, the intensities exhibit different shapes and scales across clusters. Figure B.2 reveals that the performance of SidSBM-C improves with increasing values of γ up to an optimal point, after which the performance deteriorates. In Scenario C, we set $\alpha_{q,k}$ as the same as in Scenario B. However, for this scenario, β is set to $\beta = 1$, leading to intensities with identical shapes but different scales across clusters. Figure B.2 demonstrates that the best performance of SidSBM-C is achieved with the large values of γ in this case. Overall, the observations from Figure B.2 suggest that choosing larger values of γ is preferable when the variation in scales across clusters increases. This finding justifies the heuristic selection criteria of γ.

The heuristic criterion presented here can be used to select tuning parameters more computationally efficiently than the ICL criterion. We can use the heuristic criterion to obtain a preliminary estimate of γ. Then, by conditioning on the preliminary estimate of γ, we can estimate the number of clusters K using the ICL criterion. This step is valid because the K that maximizes the ICL

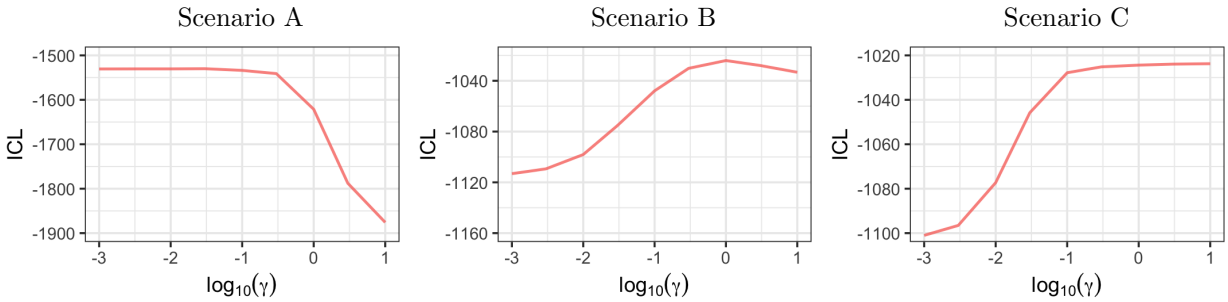

FIGURE B.2. Influence of the intensity configurations on the selection of γ across 1500 replicates. In Scenario A, where all intensities have identical scales, the ICL value is maximized by the smallest values of γ. In Scenario B, where intensities have varying shapes and scales across clusters, the value of ICL initially increases with increasing γ values, but eventually deteriorates. In Scenario C, where all intensities have identical shapes and varying scales, the ICL is maximized by the largest γ values. Overall, choosing larger values of γ is preferable when the variation in scales across clusters increases.

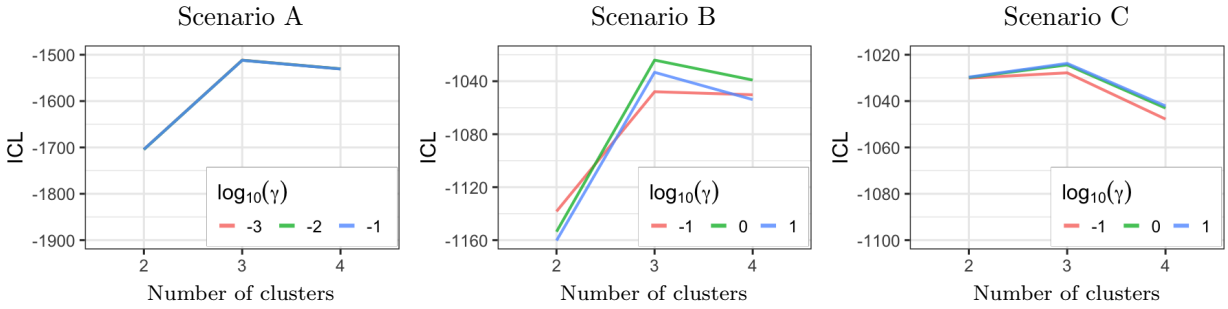

FIGURE B.3. Influence of the value of γ on the estimation of the number of clusters across 1500 replicates. In Scenario A, all intensities have identical scales and varying shapes. In Scenario B, intensities have varying shapes and scales across clusters. In Scenario C, all intensities have identical shapes and varying scales. The optimal number of clusters is consistent across different values of γ in all scenarios.

remains consistent as γ varies in a reasonable range, as shown in Figure B.3. Once K is determined, we can refine γ using the ICL criterion. This procedure avoids the joint selection of K and γ, resulting in improved computational efficiency for tuning parameter selection.

B.4.2. Sensitivity analysis with respect to γ. We conduct a sensitivity analysis with respect to γ. The results of this analysis are presented in Figure B.4. The figure illustrates that the values of ARI and MISE remain unchanged throughout the γ values that maximize the ICL value. This finding suggests that the estimation results are not sensitive to small variations in γ.

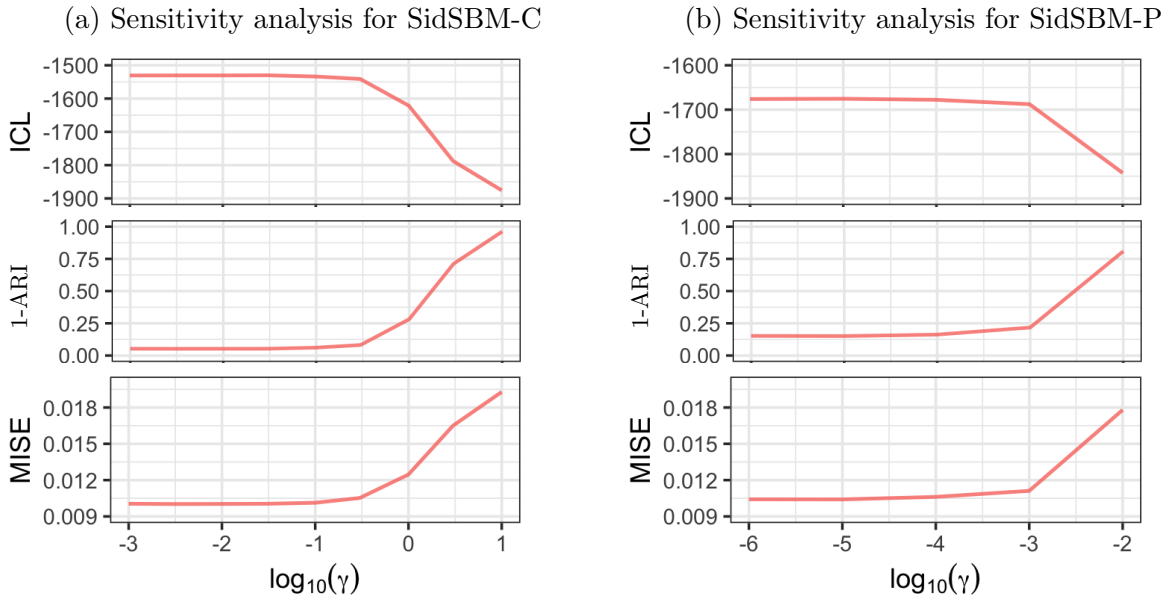

FIGURE B.4. Sensitivity of estimation results to the change of γ across 1500 replicates. Panel (a) and panel (b) display the performance of SidSBM-C and SidSBM-P, respectively. Simulated data is generated under the setting with $p = 30$, $\beta = 1.9$, and $W = 80$. For both proposed methods, the values of ARI and MISE remain unchanged throughout the values of γ that maximize the ICL.

B.4.3. Supporting evidence for explanation of performance discrepancy between SidSBM-C and SidSBM-P. The discrepancy in performance between SidSBM-C and SidSBM-P at high values of β in Figure 3.4 can be attributed to the plateaus in the objective function of SidSBM-P due to the limited ranges of intensity supports and the large values of W. To verify this hypothesis, we examine the performance of both SidSBM-C and SidSBM-P as W is progressively decreased. Figure B.5 illustrates that the performance discrepancy diminishes with decreasing W, which is aligned with our hypothesis.

B.4.4. Association between MISE and ARI. We examine the sensitivity of the MISE to changes in (1-ARI) by applying the SidSBM-C with different values of γ. By altering the value of γ, multiple estimates are generated, each with a distinct cluster estimation error and an intensity estimation error. The results, presented in Figure B.6, demonstrate a positive correlation between the MISE and (1-ARI) values. Moreover, we observe that MISE exhibits greater sensitivity to changes in (1-ARI) as the number of nodes increases.

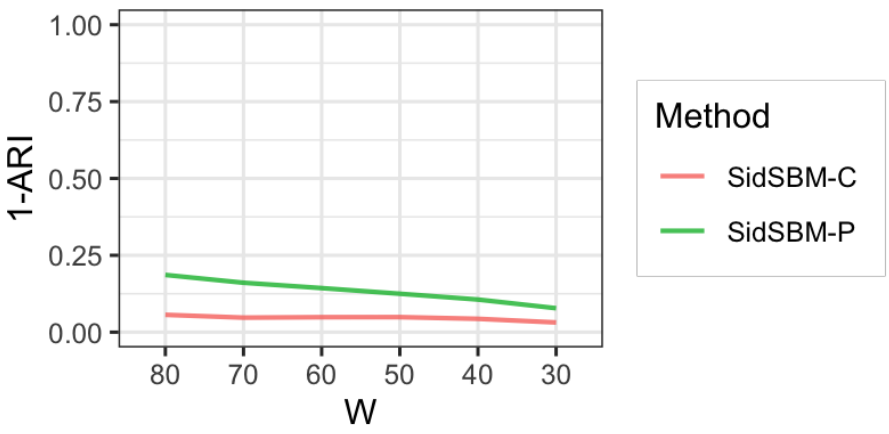

FIGURE B.5. Clustering performance of SidSBM-C and SidSBM-P in simulation with 1500 replicates. Simulated data is generated under various W with default setting being $p = 30$ and $\beta = 1.9$. The discrepancy between the performance of SidSBM-C and SidSBM-P decreases as W decreases.

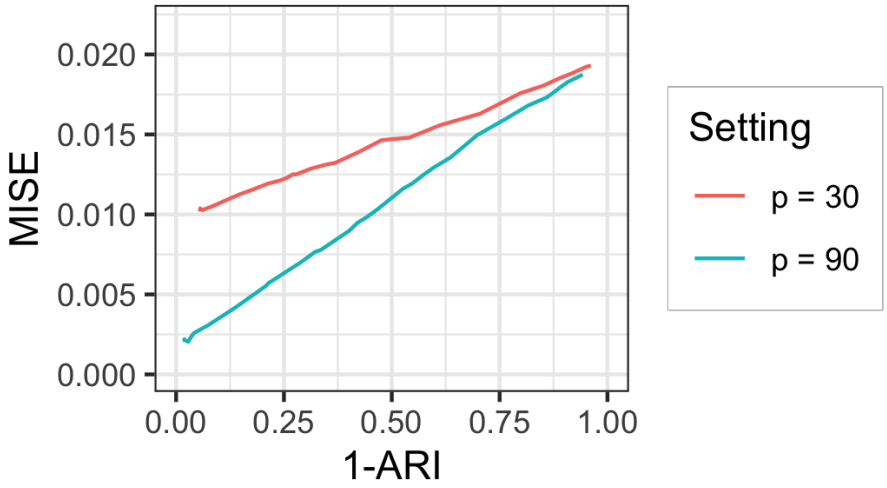

FIGURE B.6. Association between MISE and (1-ARI). Simulated data is generated under the setting $p = 30$ or 90, $\beta = 1.9$, and $W = 80$. The values of MISE and (1-ARI) are positively associated. Additionally, the sensitivity of MISE to changes in (1-ARI) becomes more pronounced as p increases.

We compare the proposed methods to the shape invariant model (SIM) in terms of intensity estimation error. The results, shown in Figure B.7, indicate that both proposed methods outperform the shape invariant model. This superiority can be attributed to the proposed methods' capability to estimate cluster memberships, whereas the shape invariant model assigns all nodes to a single

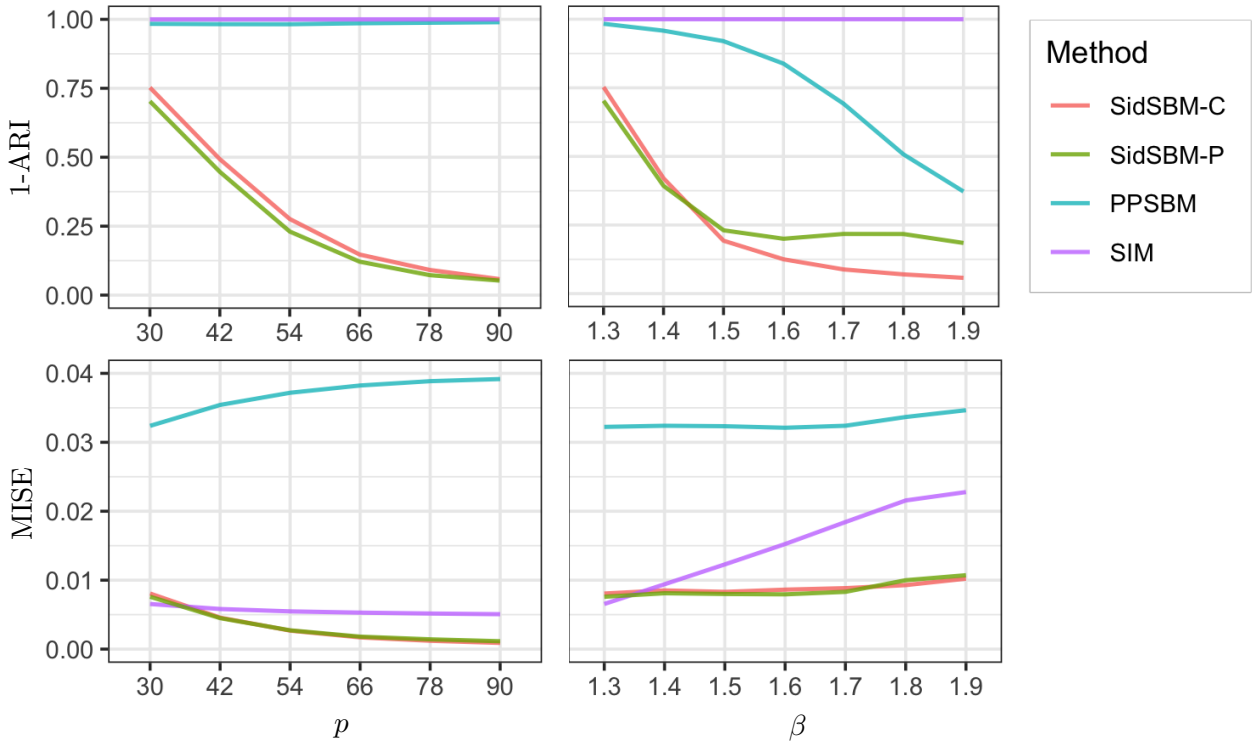

FIGURE B.7. Performance of SidSBM-C, SidSBM-P, PPSBM and SIM in simulation with 1500 replicates. Simulated data is generated under various p and β with default setting being $p = 30$, $\beta = 1.3$ and $W = 80$. Both proposed methods outperform the SIM in terms of MISE due to their capability in estimating cluster memberships.

cluster. Essentially, the enhancement in cluster estimation corresponds to the improvement in the intensity estimation.

B.4.5. Comparison of proposed methods and PPSBM in cases with no time shifts. We conduct additional simulation experiments to compare the proposed methods with the Poisson process stochastic block model (PPSBM) proposed by Matias et al. [2018]. in cases with no time shifts, that is, $W = 0$. The connecting intensities are set as $f_{q,k} = 0.9 \times \Gamma(\mu_{q,k}^2/\sigma_{q,k}^2, \mu_{q,k}/\sigma_{q,k}^2)$, where $\mu_{q,k}$ and $\sigma_{q,k}$ are the same as described in the main text. We apply SidSBM-C to simulated data with $K = 3$, $\epsilon = 0.01$, $\gamma = 0.1$ and $L = 200$. We apply SidSBM-P to simulated data with $\ell_0 = 7$, $\gamma = 0.001$, and rest parameters are the same as SidSBM-C. We apply PPSBM using the implementation in `R` by Giorgi et al. [2018] with the number of bins being 64.

We find that SidSBM-C still outperforms the other methods in the absence of time shifts as shown in Figure B.8. For completeness, we also provide the computing time of the three algorithms

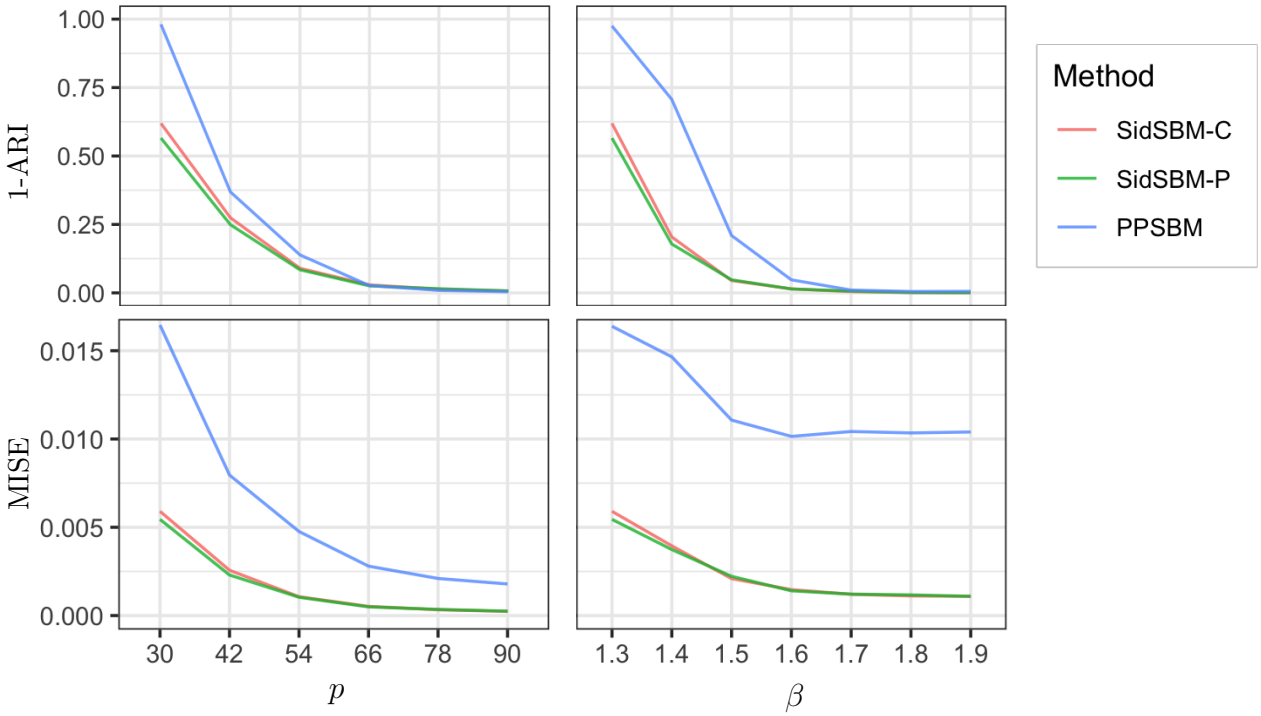

FIGURE B.8. Performance of SidSBM-C, SidSBM-P and PPSBM in simulation without time shifts across 1500 replicates. Simulated data is generated under various p and β with default setting being $p = 30$, $\beta = 1.3$. The y-axis is 1-ARI in the first row and MISE in the second row, where smaller values indicate better performance. SidSBMs exhibit better performance compared to PPSBM. Additionally, SidSBM-C and SidSBM-P demonstrate similar performance.

in Figure B.9. The proposed algorithms are much faster than PPSBM when there are no time shifts.

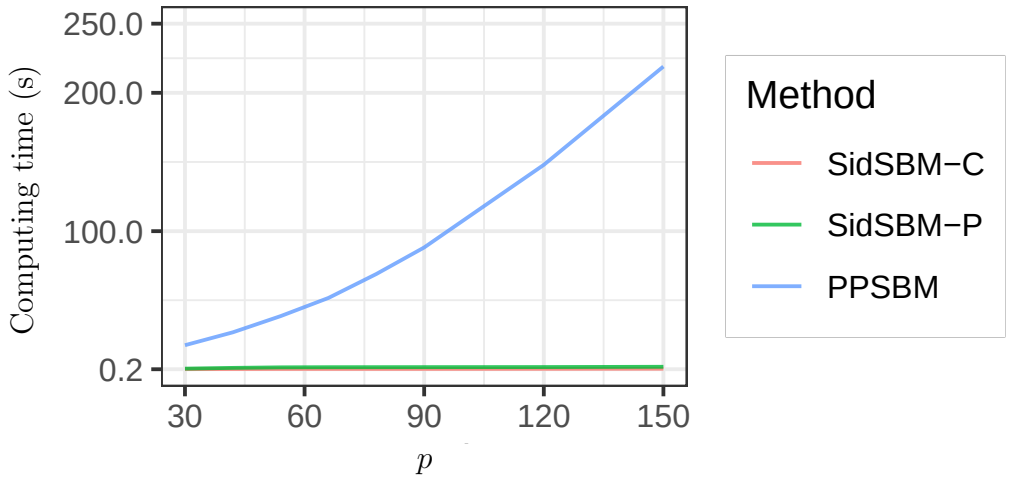

FIGURE B.9. Computing time of SidSBM-C, SidSBM-P and PPSBM in simulation without time shifts across 1500 replicates. Simulated data is generated under $\beta = 1.3$ and a range of p. The proposed methods are much more efficient than PPSBM when there is no need to fit the time shifts.

B.5. Real Data Preprocessing

B.5.1. Preprocessing steps. Neural activities are calculated following the protocal in Wan et al. [2019]. With a slight abuse of notation, we denote the fluorescence intensity of a particular neuron as F_{cell}. This notation is only used in this section and should not be confused with the representative cumulative intensities elsewhere. We calculate $\mathrm{d}F/F$ at time t as

$$\frac{\mathrm{d}F(t)}{F(t)} = \frac{F_{cell}(t) - F_{bsl}(t)}{F_{bsl}(t)},$$

where $F_{bsl}(t)$ is the baseline intensity at time t, defined as the 20^{th} percentile of F_{cell} in a sliding time window of 15 seconds (61 time points) centered at time t, that is $\{F_{cell}(s) : s \in [t - r_0, t + r_0]\}$ where $r_0 = 7.5$ seconds.

Figure B.10 presents three snapshots of the $\mathrm{d}F/F$ traces for two example pairs of neurons, showcasing the emergence and persistence of neural functional connectivity. At the 30th minute mark, the neurons do not exhibit synchronous activity. However, at 120th minute, neuron 1 and neuron 2, as well as neuron 3 and neuron 4, demonstrate synchronous activity, indicating functional connectivity between them. The synchronization persists in the third snapshot recorded at 240th minute, suggesting that the established functional connectivity between neurons endures over time.

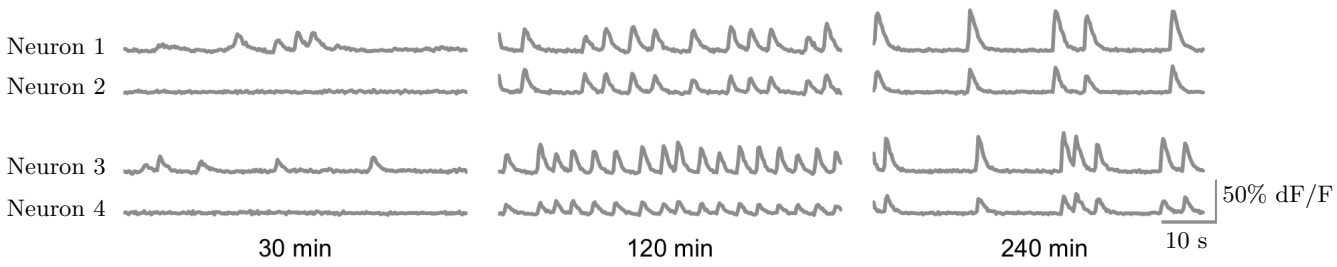

FIGURE B.10. Emergence and persistence of neural activity synchronization in two example pairs of neurons. The curves represent the dF/F traces of neurons, with time stamps indicating the time since the experiment starts. At the 30th minute, the neurons do not display synchronous activity. However, at the 120th minute, neuron 1 and neuron 2 display synchronous activity, and neuron 3 and neuron 4 also demonstrate synchronization. The synchronization between these neuron pairs persists until the 240th minute.

We estimate the functional connectivity between neurons as follows. For each pair of neurons, we compute the Pearson correlation of their $\mathrm{d}F/F$ traces for every 1-minute interval, or 240 time points at 4 Hz. Then the correlation trace is smoothed by taking average of the previous 5 correlations at each time stamp. The connecting time of two neurons is defined as the earliest time that their activity's smoothed correlation exceeds a threshold $\rho = 0.4$. Upon closer investigation into the connecting times, we exclude five neurons due to their abnormal behavior in connection patterns, e.g. way fewer (less than 5) edges than average (around 40).

The neuronal networks are estimated separately for the left and right spines. Figure B.11 illustrates the spatial distribution of recorded neurons in the zebrafish. Neurons are differentiated into the left and right spines based on their LR-coordinate. Specifically, neurons with an LR-coordinate smaller than 0 are presumed to be situated in the left spine, while those with an LR-coordinate equal to or greater than 0 are presumed to be located in the right spine.

B.5.2. Selection of ρ. In this section, we explain why the threshold ρ is set to be 0.4 in the edge detection. For a pair of neurons $(i, j) \in \mathcal{V}$, we denote the correlation trajectory by $X(t)$ and denote its expectation by $\mu(t) \equiv \mathbb{E}X(t)$, where $t \in [0, T]$.

The correlation trajectories among all pairs of neurons are shown in Figure B.12. Prior to synchronization, neurons show no or irregular spontaneous activities, where the correlations mostly fall below 0.2. Hence, we hypothesize that truly connected neurons should have a correlation curve that exceeds and stays above 0.2. Formally, we characterize the detection of edges using the

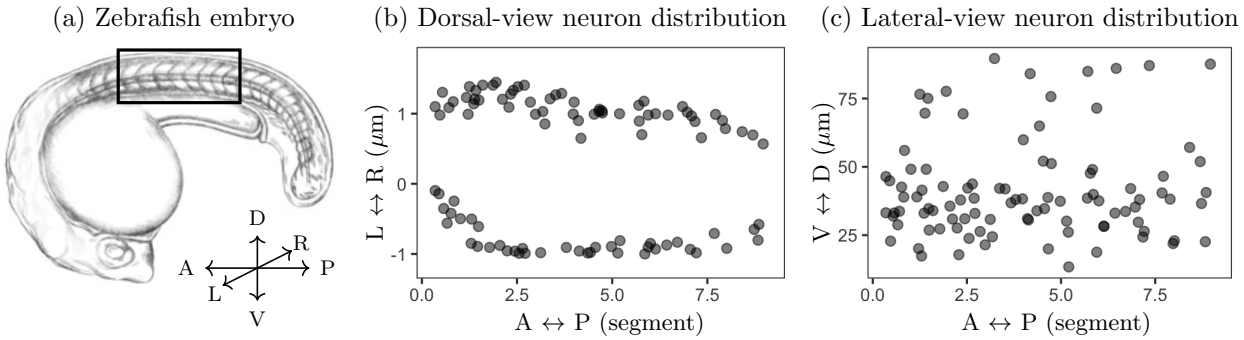

FIGURE B.11. The spatial location of recorded neurons in zebrafish. Panel (a) shows a schematic zebrafish embryo, with the black rectangle highlighting the spinal cord where neural activities are recorded. Additionally, a coordinate system is displayed, with "A" representing anterior, "P" representing posterior, "L" representing left, "R" representing right, "D" representing dorsal, and "V" representing ventral. Panel (b) shows the spatial distribution of neurons from a dorsal-view perspective. The AP-coordinate is measured in *segment*, which is defined by motor nerve roots according to Wan et al. [2019]. Panel (c) shows the lateral-view spatial distribution of neurons.

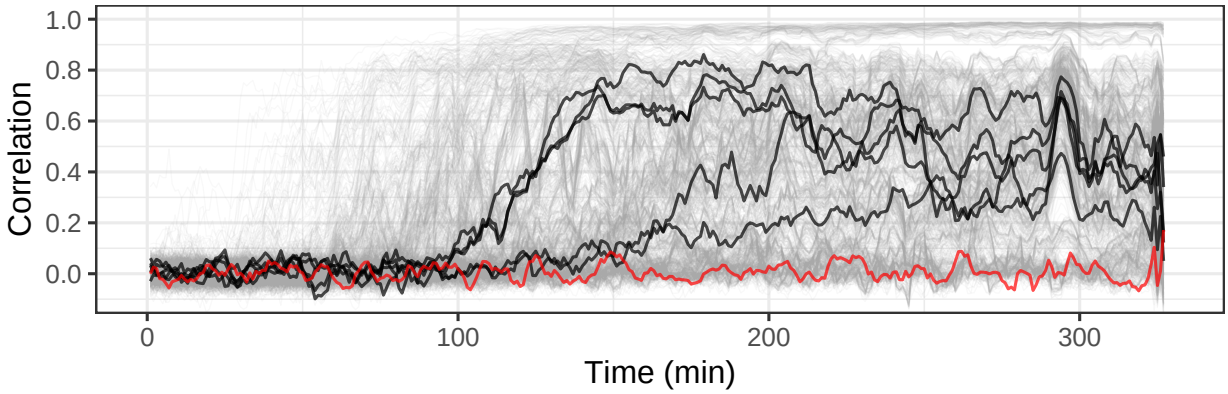

FIGURE B.12. Correlation trajectories (i.e., gray curves) between all pairs of neurons in the left spine. Black curves are (randomly selected) examples of correlation trajectories that exceed 0.2 at some time point. The red curve is an example of correlation trajectories that stays less than 0.2 during the experiment.

following null and alternative hypotheses:

$$H_0 : \mu(t) = 0, t \in [0, T] \quad \text{v.s.} \quad H_a : \mu(t) > 0.2, t \in [t_0, T] \text{ for some } t_0 \in [0, T).$$

To conduct hypothesis test, we use the test statistic $X_{\max} \equiv \max_{t\in[0,T]} X(t)$ with rejection region $\{X : X_{\max} > \rho\}$, where ρ is the threshold.

TABLE B.2. Empirical type I error rate under various values of the threshold ρ.

	Threshold (ρ)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Type I error rate (%)	94	0	0	0	0	0	0	0	0	0

We choose the value of ρ by examining the empirical type I error rate and the empirical power by simulating data based on the observed correlation curves in Figure B.12.

(1) To evaluate the type I error rate, we simulate data based on the red curve in Figure B.12, where the null hypothesis seems to hold. To be specific, we fit an ARIMA model to the red correlation trajectory, and use the fitted model to simulate 5000 stochastic processes. With simulated processes, we test the hypothesis with $\rho \in \{0.1, 0.2, \cdots, 1.0\}$ and compute the empirical type I error rates, as shown in Table B.2. From the simulation results, we can choose $\rho \geq 0.2$ to achieve a near zero type I error rate.

(2) To evaluate the power, we need to simulate data under different alternative hypotheses from the real data. To this end, we apply k-means with 5 clusters to all correlation trajectories and obtain 5 mean correlation trajectories as shown in Figure B.13(a). We can see that the gray curve might correspond to the non-edges (i.e., the null), hence we remove this curve in this power analysis. Setting each mean correlation trajectory as $\mu(\cdot)$, we simulate 5000 stochastic processes by introducing noise generated from the fitted ARIMA model. We then evaluate the empirical power of hypothesis test with $\rho \in \{0.1, 0.2, \cdots, 1.0\}$ as shown in Figure B.13(b). For $\rho \leq 0.5$, the power is close to 1 under all alternatives.

In conclusion, the simulation shows that a threshold ρ between 0.2 to 0.5 yields close-to-zero type I error rate and close-to-one power. Therefore, we set ρ to be 0.4 in our analysis.

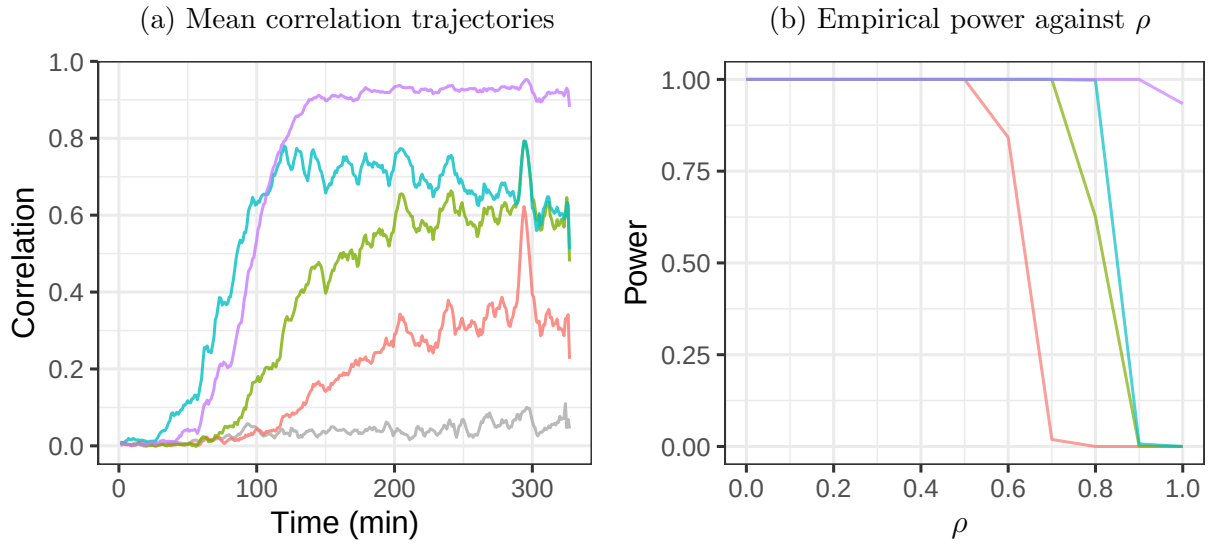

FIGURE B.13. Mean correlation trajectories and empirical power under various ρ. Panel (a) shows mean correlation trajectories in different colors estimated using k-means with 5 clusters on all correlation trajectories. Panel (b) shows empirical powers for the alternatives using the four mean trajectories in panel (a). Panels (a) and (b) use the same color coding system.

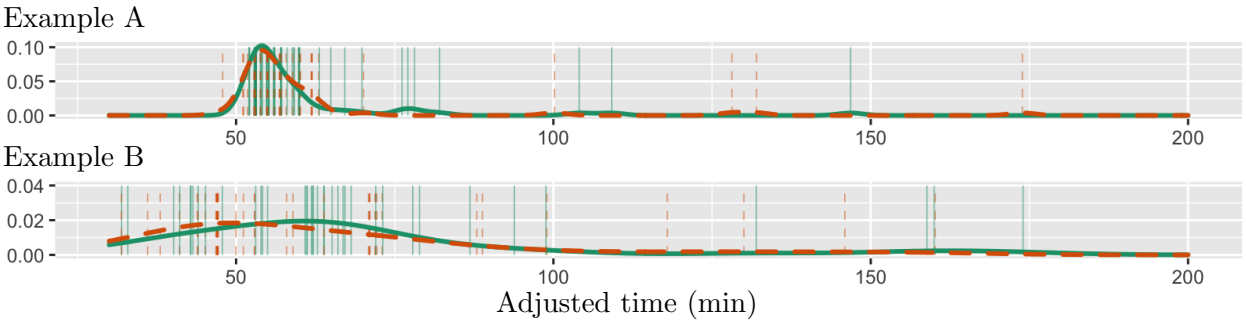

FIGURE B.14. Examples of neuron pairs exhibiting similar connecting behaviour. Example A showcases two neurons from estimated Cluster 2, whereas Example B involves two neurons from estimated Cluster 3. Vertical bars represent the adjusted connecting times, and the smooth curves represent the densities of connecting times. Different colors and linetypes correspond to different neurons.

B.6. Additional Data Analysis Results

B.6.1. Examples of neuron pairs with similar connecting behavior. Figure B.14 showcases examples of neuron pairs that display similar connecting behavior. The figure displays connecting times that have been adjusted based on estimated time shifts. The overlapping density curves indicate the similarity in connecting behavior between the neurons.

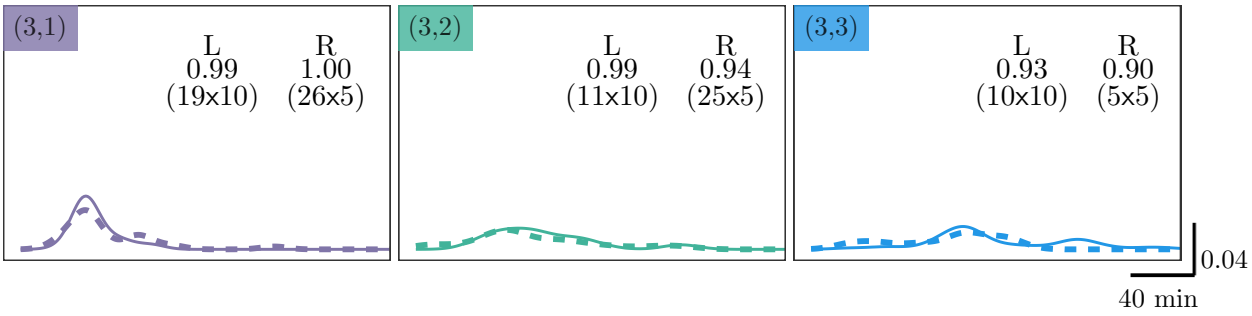

FIGURE B.15. Aligned estimated representative connecting intensities of the third clusters in the left and right spines. Colors and line types are the same as in Figure 3.7 in main text. The estimated connecting intensities are manually shifted to match each other to demonstrate the similarity of their shapes. The estimated intensities $\hat{f}_{L,3,k}$ and $\hat{f}_{R,3,k}$ have almost identical shapes for $k = 1, 2, 3$.

B.6.2. Aligned estimated representative connecting intensities. We match the estimated representative connecting intensities of the third clusters in the left and right spines in Figure B.15. Here the connecting intensities $\{\hat{f}_{R,3,k}, k = 1, 2, 3\}$ are shifted individually to align with $\{\hat{f}_{L,3,k}, k = 1, 2, 3\}$. After the additional alignment, we observe that the estimated intensities $\hat{f}_{L,3,k}$ and $\hat{f}_{R,3,k}$ for $k = 1, 2, 3$ are highly consistent. We suspect that the errors in the estimated shifts might be a result of estimation error given the small sizes of the third clusters.

B.6.3. Fulfillment of Condition C3.2 in the real data. In order to assess the fulfillment of Condition C3.2 in the data set, we examine the distributions of estimated time shifts, which are depicted in Figure B.16. From the figure, it becomes evident that there are overlaps in the ranges of time shifts among various clusters. This observation suggests that Condition C3.2 is met in the real data.

B.6.4. Sensitivity of the analysis results to the value of ρ. To assess the sensitivity of the analysis results to the value of ρ, we conduct analyses using two different values: $\rho = 0.3$ and $\rho = 0.5$. We then compare these results to those obtained from $\rho = 0.4$. In both comparisons, 80% of neurons exhibit consistent memberships. In addition, Figure B.17 presents the estimated connecting intensities from $\rho = 0.3$ and $\rho = 0.5$. The representative connecting intensities display a high level of consistency across different values of ρ.

B.6.5. Real data analysis results using the PPSBM. For comparison, we apply PPSBM [Matias et al., 2018] to the neural data. Recall that PPSBM estimates the cluster structure without

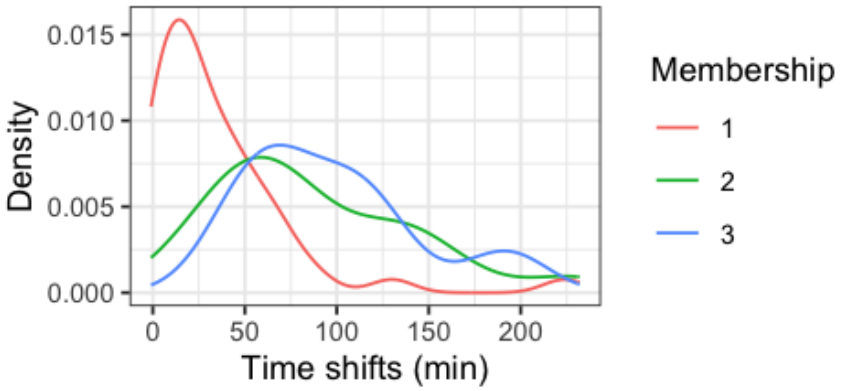

FIGURE B.16. Distributions of estimated time shifts of neurons in the three estimated clusters. Each curve represents the density of estimated time shifts within a particular cluster, with different colors denoting distinct clusters. There exist overlaps in the ranges of time shifts across the three clusters.

(a) Estimated connecting intensities when $\rho = 0.3$ (b) Estimated connecting intensities when $\rho = 0.5$

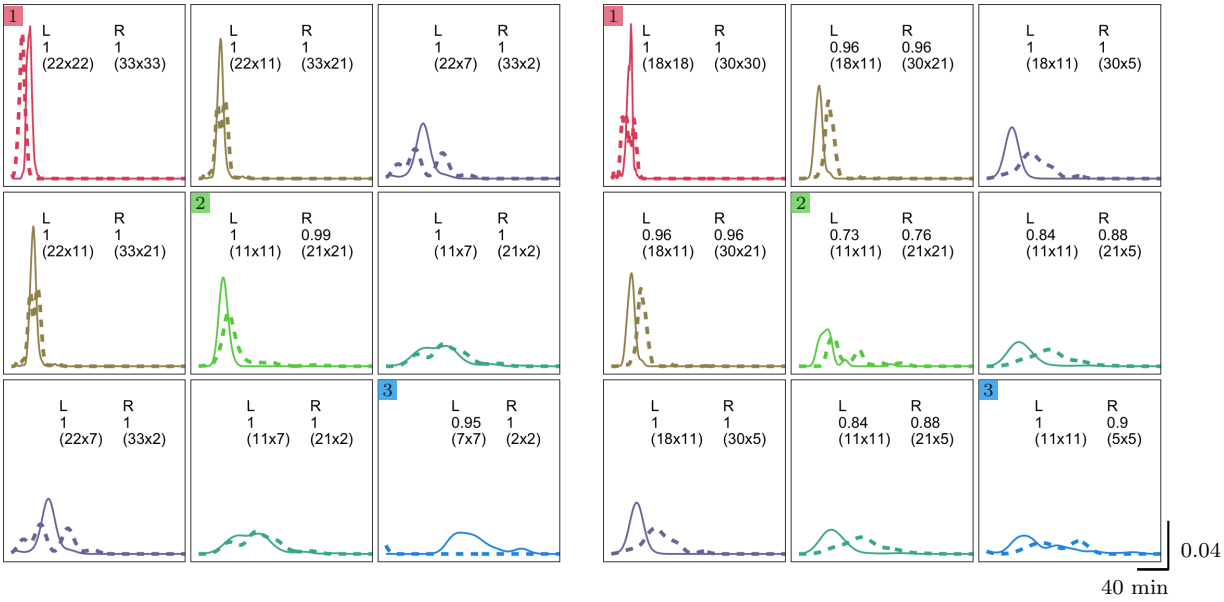

FIGURE B.17. Estimated connecting intensities of networks obtained from $\rho = 0.3$ or $\rho = 0.5$. The solid and dashed curves represent estimated representative connecting intensities for the left and right spines, respectively. Each subfigure contains a tabular legend, where the first row shows "L" for "left" and "R" for "right", the second row shows connecting probabilities, the third row shows the sizes of pairs of clusters. As the value of ρ varies, the representative connecting intensities remain highly consistent.

accounting for the unknown time shifts. Estimated connecting intensities are shown in Figure B.18.

We observe that estimated intensities between the left and right spines are less consistent than the

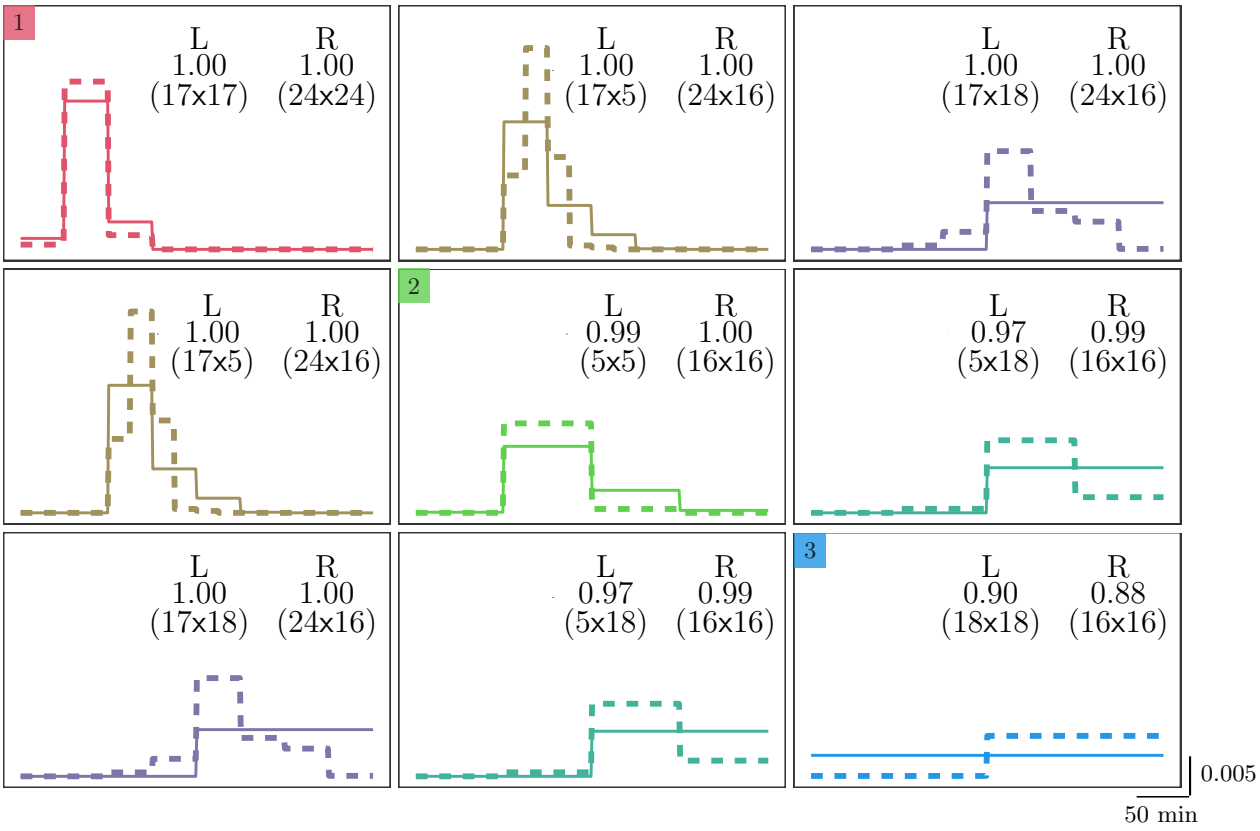

FIGURE B.18. Estimated connecting intensities by PPSBM [Matias et al., 2018] on the neural data set from Wan et al. [2019]. The rows and columns are indexed by clusters as shown in the top left squares on diagonal. For $q, k = 1, 2, 3$, solid and dashed curves at position (q, k) represent estimated representative connecting intensity $\hat{f}_{L,q,k}$ and $\hat{f}_{R,q,k}$ from the left and right spines, respectively. Each subfigure contains a tabular legend, where the first row shows "L" for "left" and "R" for "right", the second row shows connecting probabilities, the third row shows the sizes of pairs of clusters. The estimates from the left and right spines are not consistent.

estimates in Figure 3.7 in main text, although the estimated intensities of Cluster 1 and 2 still bear some resemblance between the two sides of the spine. From Figure B.19, we can see that estimated cluster memberships are highly correlated with connecting time of first edges of neurons. In other words, the cluster structure from PPSBM is determined by the activation time of neurons rather than the actual connecting patterns.

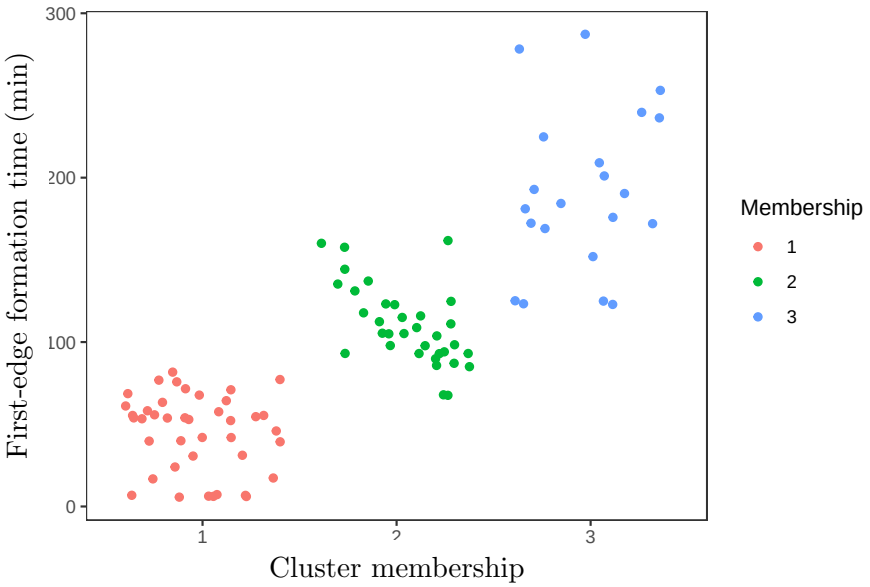

FIGURE B.19. Clusters estimated by PPSBM are highly correlated with the connecting time of the first edge of each neuron. The x-coordinate is the memberships of neurons (jittered for visualization), and the y-coordinate is the time of the first edge of each neuron.

B.7. Derivation of Gradients in Section B.3

Recall that the loss function in (B.27) is

$$R(\mathbf{v}) = \sum_{i\in\mathcal{V}} \sum_{k\in[K]} \left(|\hat{J}_{i,k}| \cdot \overline{N}_{i,k}(T) \cdot \sum_{|\ell|\le \ell_0} \left| \overline{c}_{\ell,i,k} e^{\mathbf{i}2\pi\ell(v_i/T)} - \phi_{\ell,\hat{z}_i,k} \right|_2^2 \right), \tag{B.29}$$

where

$$\overline{c}_{l,i,k} = \frac{\sum_{j\in\hat{J}_{i,k}:N_{i,j}(T)=1} c_{\ell,i,j}}{\#\{j \in \hat{J}_{i,k} : N_{i,j}(T) = 1\}} \tag{B.30}$$

and

$$\begin{aligned} \phi_{\ell,q,k} =& \frac{\sum_{i\in\hat{\mathcal{C}}_q, j\in\hat{\mathcal{C}}_k, N_{i,j}(T)=1, i\neq j} c_{\ell,i,j} \cdot \exp(\mathbf{i}2\pi\ell \max(v_i, v_j)/T)}{\#\{(i,j) : i \in \hat{\mathcal{C}}_q, j \in \hat{\mathcal{C}}_k, N_{i,j}(T) = 1, i \neq j\}} \\ =& \frac{\sum_{i\in\hat{\mathcal{C}}_q} \#\{j \in \hat{J}_{i,k} : N_{i,j}(T) = 1\} \overline{c}_{\ell,i,k} \cdot e^{\mathbf{i}2\pi\ell v_i/T}}{\#\{(i,j) : i \in \hat{\mathcal{C}}_q, j \in \hat{\mathcal{C}}_k, N_{i,j}(T) = 1, i \neq j\}} \\ &+ \frac{\sum_{j\in\hat{\mathcal{C}}_k} \#\{j' \in \hat{J}_{j,q} : N_{j,j'}(T) = 1\} \overline{c}_{\ell,j,q} \cdot e^{\mathbf{i}2\pi\ell v_j/T}}{\#\{(i,j) : i \in \hat{\mathcal{C}}_q, j \in \hat{\mathcal{C}}_k, N_{i,j}(T) = 1, i \neq j\}}. \end{aligned} \tag{B.31}$$

Let $\hat{J}_{1,i,k} = \{j \in \hat{J}_{i,k} : N_{i,j}(T) = 1\}$ and $\hat{D}_{q,k} = \{(i,j) : i \in \hat{\mathcal{C}}_q, j \in \hat{\mathcal{C}}_k, N_{i,j}(T) = 1, i \neq j\}$. By definition, $\hat{D}_{q,k}$ is the collection of pairs of connected nodes in clusters $\hat{\mathcal{C}}_q$ and $\hat{\mathcal{C}}_k$, and $\hat{J}_{1,i,k}$ is the collection of nodes in cluster $\hat{\mathcal{C}}_k$ that are connected with node i and has a smaller time shift than node i. Thus we know that $|\hat{D}_{q,k}| = \sum_{i\in\hat{\mathcal{C}}_q} |\hat{J}_{1,i,k}| + \sum_{j'\in\hat{\mathcal{C}}_k} |\hat{J}_{1,j',q}|$. For any $j \in \mathcal{V}$, we separate the terms including $\exp(\mathbf{i}2\pi\ell v_j/T)$ from $\phi_{\ell,q,k}$ by writing

$$\phi_{\ell,q,k} = Q_{\ell,q,k,j} e^{\mathbf{i}2\pi\ell v_j/T} + (\phi_{\ell,q,k} - Q_{\ell,q,k,j} e^{\mathbf{i}2\pi\ell v_j/T}), \tag{B.32}$$

where we introduce $Q_{\ell,q,k,j}$ for ease of bookkeeping

$$Q_{\ell,q,k,j} = \begin{cases} |\hat{J}_{1,j,k}| \cdot \overline{c}_{\ell,j,k} \cdot |\hat{D}_{q,k}|^{-1}, & \text{if } j \in \hat{\mathcal{C}}_q, q \neq k, \\ |\hat{J}_{1,j,q}| \cdot \overline{c}_{\ell,j,q} \cdot |\hat{D}_{q,k}|^{-1}, & \text{if } j \in \hat{\mathcal{C}}_k, q \neq k, \\ 2|\hat{J}_{1,j,q}| \cdot \overline{c}_{\ell,j,q} \cdot |\hat{D}_{q,q}|^{-1}, & \text{if } j \in \hat{\mathcal{C}}_k, q = k, \\ 0, & \text{if } j \notin \hat{\mathcal{C}}_q \cup \hat{\mathcal{C}}_k. \end{cases} \tag{B.33}$$

Then for $i \neq j$, we have the following gradient

$$\begin{aligned} &\frac{\partial}{\partial v_j} \left| \overline{c}_{\ell,i,k} e^{\mathbf{i}2\pi\ell v_i/T} - \phi_{\ell,\hat{z}_i,k} \right|_2^2 \\ &= \frac{\partial}{\partial v_j} \left| \overline{c}_{\ell,i,k} e^{\mathbf{i}2\pi\ell v_i/T} - Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T} - (\phi_{\ell,\hat{z}_i,k} - Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T}) \right|_2^2 \\ &= 2\frac{\partial}{\partial v_j} \mathrm{Re}\left[\overline{\Big(- Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T} \Big)} \Big(\overline{c}_{\ell,i,k} e^{\mathbf{i}2\pi\ell v_i/T} - (\phi_{\ell,\hat{z}_i,k} - Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T}) \Big) \right] \\ &= 2\mathrm{Re}\left[\Big((\mathbf{i}2\pi\ell/T) \overline{Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T}} \Big) \Big(\overline{c}_{\ell,i,k} e^{\mathbf{i}2\pi\ell v_i/T} - (\phi_{\ell,\hat{z}_i,k} - Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T}) \Big) \right] \\ &= 2\mathrm{Re}\left[\Big((\mathbf{i}2\pi\ell/T) \overline{Q_{\ell,\hat{z}_i,k,j} e^{\mathbf{i}2\pi\ell v_j/T}} \Big) \Big(\overline{c}_{\ell,i,k} e^{\mathbf{i}2\pi\ell v_i/T} - \phi_{\ell,\hat{z}_i,k} \Big) \right]. \end{aligned} \tag{B.34}$$

For $i = j$, we have the following gradient

$$
\begin{aligned}
&\frac{\partial}{\partial v_j}\left|\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell v_j/T} - \phi_{\ell,\hat{z}_j,k}\right|_2^2 \\
&= \frac{\partial}{\partial v_j}\left|(\overline{c}_{\ell,j,k} - Q_{\ell,\hat{z}_j,k,j})e^{\mathbf{i}2\pi\ell v_j/T} - (\phi_{\ell,\hat{z}_j,k} - Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T})\right|_2^2 \\
&= 2\frac{\partial}{\partial v_j}\mathrm{Re}\left[\overline{(\overline{c}_{\ell,j,k} - Q_{\ell,\hat{z}_j,k,j})e^{\mathbf{i}2\pi\ell v_j/T}}(-\phi_{\ell,\hat{z}_j,k} + Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T})\right] \\
&= 2\mathrm{Re}\left[(\mathbf{i}2\pi\ell/T)\overline{(\overline{c}_{\ell,j,k} - Q_{\ell,\hat{z}_j,k,j})e^{\mathbf{i}2\pi\ell v_j/T}}(\phi_{\ell,\hat{z}_j,k} - Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T})\right] \\
&= \begin{cases} 2\mathrm{Re}\left[(\mathbf{i}2\pi\ell/T)\Big(|\hat{D}_{\hat{z}_j,k}|\cdot|\hat{J}_{1,j,k}|^{-1} - 1\Big)\overline{Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\right], & \text{if } \hat{z}_j \neq k, \\ 2\mathrm{Re}\left[(\mathbf{i}2\pi\ell/T)\Big(|\hat{D}_{k,k}|\cdot(2|\hat{J}_{1,j,k}|)^{-1} - 1\Big)\overline{Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\right], & \text{if } \hat{z}_j = k. \end{cases}
\end{aligned} \tag{B.35}
$$

Therefore, the gradient of $R(\mathbf{v})$ with respect to v_j is as follows

$$
\begin{aligned}
&\frac{\partial}{\partial v_j}R(\mathbf{v}) \\
=&\frac{\partial}{\partial v_j}\sum_{i\in\mathcal{V}}\sum_{k\in[K]}\left(|\hat{J}_{i,k}|\cdot\overline{N}_{i,k}(T)\cdot\sum_{|\ell|\le\ell_0}\left|\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell(v_i/T)} - \phi_{\ell,\hat{z}_i,k}\right|_2^2\right) \\
=&\frac{\partial}{\partial v_j}\sum_{i\in\mathcal{V}}\sum_{k\in[K]}\left(|\hat{J}_{1,i,k}|\cdot\sum_{|\ell|\le\ell_0}\left|\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell(v_i/T)} - \phi_{\ell,\hat{z}_i,k}\right|_2^2\right) \\
=&\sum_{i\in\mathcal{V},i\neq j}\sum_{k\in[K]}\left(|\hat{J}_{1,i,k}|\cdot\sum_{|\ell|\le\ell_0}\frac{\partial}{\partial v_j}\left|\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell(v_i/T)} - \phi_{\ell,\hat{z}_i,k}\right|_2^2\right) \\
&+\sum_{k\in[K]}\left(|\hat{J}_{1,i,k}|\cdot\sum_{|\ell|\le\ell_0}\frac{\partial}{\partial v_j}\left|\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell(v_j/T)} - \phi_{\ell,\hat{z}_j,k}\right|_2^2\right).
\end{aligned} \tag{B.36}
$$

Plugging in the form of the derivatives yields

$$
\begin{aligned}
&\frac{\partial}{\partial v_j} R(\mathbf{v}) \\
=&\sum_{i\in\mathcal{V}, i\neq j}\sum_{k\in[K]}\Bigg(|\hat{J}_{1,i,k}|\sum_{|\ell|\leq \ell_0} 2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\overline{Q_{\ell,\hat{z}_i,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\Big)\Big(\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,\hat{z}_i,k}\Big)\Big]\Bigg) \\
&+\sum_{k\in[K],k\neq\hat{z}_j}\Bigg(|\hat{J}_{1,j,k}|\sum_{|\ell|\leq\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\Big(|\hat{D}_{\hat{z}_j,k}||\hat{J}_{1,j,k}|^{-1}-1\Big)\overline{Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big]\Bigg) \\
&+\sum_{k=\hat{z}_j}\Bigg(|\hat{J}_{1,j,k}|\sum_{|\ell|\leq\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\Big(|\hat{D}_{k,k}|(2|\hat{J}_{1,j,k}|)^{-1}-1\Big)\overline{Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big]\Bigg) \\
\equiv&A_1+A_2+A_3.
\end{aligned}
\tag{B.37}
$$

We can further simplify the three terms A_1, A_2 and A_3. The first term follows that

(B.38)

$$
\begin{aligned}
A_1=&\sum_{i\in\mathcal{V},i\neq j}\sum_{k\in[K]}\Bigg(|\hat{J}_{1,i,k}|\cdot\sum_{|\ell|\leq\ell_0}2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\overline{Q_{\ell,\hat{z}_i,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\Big)\Big(\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,\hat{z}_i,k}\Big)\Big]\Bigg) \\
=&\sum_{q\in[K],q\neq\hat{z}_j}\sum_{k=\hat{z}_j}\sum_{|\ell|\leq\ell_0}\Bigg(2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\overline{Q_{\ell,q,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\Big)\Big\{\sum_{i\in\hat{\mathcal{C}}_q}|\hat{J}_{1,i,k}|\cdot\Big(\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,q,k}\Big) \\
&\qquad+\sum_{j'\in\hat{\mathcal{C}}_k,j'\neq j}|\hat{J}_{1,j',q}|\cdot\Big(\overline{c}_{\ell,j',q}e^{\mathbf{i}2\pi\ell v_{j'}/T}-\phi_{\ell,k,q}\Big)\Big\}\Big]\Bigg) \\
&+\sum_{q=k=\hat{z}_j}\sum_{|\ell|\leq\ell_0}\Bigg(2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\overline{Q_{\ell,q,q,j}e^{\mathbf{i}2\pi\ell v_j/T}}\Big)\sum_{i\in\hat{\mathcal{C}}_q,i\neq j}|\hat{J}_{1,i,q}|\cdot\Big(\overline{c}_{\ell,i,q}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,q,q}\Big)\Big]\Bigg),
\end{aligned}
$$

where we have used the property that $Q_{\ell,\hat{z}_i,k,j}=0$ if $j\notin\hat{\mathcal{C}}_{\hat{z}_i}\cup\hat{\mathcal{C}}_k$, and break down the summation according to $\{(i,k):i\in\mathcal{V},i\neq j,k\in[K]\}=\{(i,k):i\in\hat{\mathcal{C}}_{\hat{z}_j},i\neq j,k=\hat{z}_j\}\cup\{(i,k):i\in\hat{\mathcal{C}}_{\hat{z}_j},i\neq j,k\neq\hat{z}_j\}\cup\{(i,k):i\notin\hat{\mathcal{C}}_{\hat{z}_j},k=\hat{z}_j\}$.

Note that

$$
\begin{aligned}
&\sum_{i\in\hat{\mathcal{C}}_q}|\hat{J}_{1,i,k}|\cdot\Big(\overline{c}_{\ell,i,k}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,q,k}\Big)+\sum_{j'\in\hat{\mathcal{C}}_k}|\hat{J}_{1,j',q}|\cdot\Big(\overline{c}_{\ell,j',q}e^{\mathbf{i}2\pi\ell v_{j'}/T}-\phi_{\ell,k,q}\Big) \\
=&|\hat{D}_{q,k}|\cdot\phi_{\ell,q,k}-\Big(\sum_{i\in\hat{\mathcal{C}}_q}|\hat{J}_{1,i,k}|+\sum_{j'\in\hat{\mathcal{C}}_k}|\hat{J}_{1,j',q}|\Big)\cdot\phi_{\ell,q,k} \\
=&0,
\end{aligned}
\tag{B.39}
$$

and

$$\sum_{i\in\hat{\mathcal{C}}_q}|\hat{J}_{1,i,q}|\cdot\Big(\overline{c}_{\ell,i,q}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,q,q}\Big)=\frac{1}{2}|\hat{D}_{q,q}|\cdot\phi_{\ell,q,q}-\sum_{i\in\hat{\mathcal{C}}_q}|\hat{J}_{1,i,q}|\cdot\phi_{\ell,q,q}=0. \tag{B.40}$$

Therefore, we have

$$\begin{aligned}
A_1 =& -\sum_{q\in[K],q\neq\hat{z}_j}\sum_{k=\hat{z}_j}\sum_{|\ell|\le\ell_0}\left(2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\overline{Q_{\ell,q,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\Big)\sum_{j'=j}|\hat{J}_{1,j',q}|\cdot\Big(\overline{c}_{\ell,j',q}e^{\mathbf{i}2\pi\ell v_{j'}/T}-\phi_{\ell,k,q}\Big)\Big]\right)\\
&-\sum_{q=k=\hat{z}_j}\sum_{|\ell|\le\ell_0}\left(2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\overline{Q_{\ell,q,q,j}e^{\mathbf{i}2\pi\ell v_j/T}}\Big)\sum_{i=j}|\hat{J}_{1,i,q}|\cdot\Big(\overline{c}_{\ell,i,q}e^{\mathbf{i}2\pi\ell v_i/T}-\phi_{\ell,q,q}\Big)\Big]\right)\\
=&\sum_{q\in[K],q\neq\hat{z}_j}\sum_{|\ell|\le\ell_0}\left(2\mathrm{Re}\Big[\Big((\mathbf{i}2\pi\ell/T)\ |\hat{J}_{1,j,q}|^2\ |\hat{D}_{q,\hat{z}_j}|^{-1}\ \overline{c}_{\ell,j,q}e^{\mathbf{i}2\pi\ell v_j/T}\Big)\phi_{\ell,\hat{z}_j,q}\Big]\right)\\
&+\sum_{|\ell|\le\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\cdot 2|\hat{J}_{1,j,\hat{z}_j}|^2\ |\hat{D}_{\hat{z}_j,\hat{z}_j}|^{-1}\overline{c}_{\ell,j,z_j}e^{\mathbf{i}2\pi\ell v_j/T}\phi_{\ell,z_j,z_j}\Big]\\
\equiv& I_1+I_2,
\end{aligned} \tag{B.41}$$

where in the second equality we use the definition of $Q_{\ell,q,q,j}$ and the fact that

$$\overline{\overline{c}_{\ell,j,q}e^{\mathbf{i}2\pi\ell v_j/T}}\overline{c}_{\ell,j,q}e^{\mathbf{i}2\pi\ell v_j/T}=\ \text{Const.}$$

Plugging in the definition of $Q_{\ell,\hat{z}_j,k,j}$ into A_2 and A_3 yields

$$\begin{aligned}
A_2 =& \sum_{k\in[K],k\neq\hat{z}_j}\left(|\hat{J}_{1,j,k}|\cdot\sum_{|\ell|\le\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\Big(|\hat{D}_{\hat{z}_j,k}|\cdot|\hat{J}_{1,j,k}|^{-1}-1\Big)\overline{Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big]\right)\\
=&\sum_{k\in[K],k\neq\hat{z}_j}\sum_{|\ell|\le\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\Big(|\hat{J}_{1,j,k}|-|\hat{J}_{1,j,k}|^2\cdot|\hat{D}_{\hat{z}_j,k}|^{-1}\Big)\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell v_j/T}\phi_{\ell,\hat{z}_j,k}\Big],
\end{aligned} \tag{B.42}$$

and

$$\begin{aligned}
A_3 =& \sum_{k=\hat{z}_j}\left(|\hat{J}_{1,j,k}|\cdot\sum_{|\ell|\le\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\Big(|\hat{D}_{k,k}|\cdot(2|\hat{J}_{1,j,k}|)^{-1}-1\Big)\overline{Q_{\ell,\hat{z}_j,k,j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big]\right)\\
=&\sum_{|\ell|\le\ell_0}2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)\Big(|\hat{J}_{1,j,\hat{z}_j}|-(2|\hat{J}_{1,j,\hat{z}_j}|^2)\cdot|\hat{D}_{\hat{z}_j,\hat{z}_j}|^{-1}\Big)\overline{c}_{\ell,j,\hat{z}_j}e^{\mathbf{i}2\pi\ell v_j/T}\phi_{\ell,\hat{z}_j,\hat{z}_j}\Big].
\end{aligned} \tag{B.43}$$

We can see that

$$
\begin{aligned}
I_1 + A_2 &= \sum_{k\in[K],k\neq\hat{z}_j} \sum_{|\ell|\leq\ell_0} 2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)|\hat{J}_{1,j,k}|\overline{\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big], \\
I_2 + A_3 &= \sum_{|\ell|\leq\ell_0} 2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)|\hat{J}_{1,j,\hat{z}_j}|\overline{\overline{c}_{\ell,j,\hat{z}_j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,\hat{z}_j}\Big].
\end{aligned} \tag{B.44}
$$

Finally, we can derive that

$$
\begin{aligned}
\frac{\partial}{\partial v_j}R(\mathbf{v}) &= I_1 + I_2 + A_2 + A_3 \\
&= \sum_{k\in[K],k\neq\hat{z}_j} \sum_{|\ell|\leq\ell_0} 2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)|\hat{J}_{1,j,k}|\overline{\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big] \\
&\quad + \sum_{|\ell|\leq\ell_0} 2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)|\hat{J}_{1,j,\hat{z}_j}|\overline{\overline{c}_{\ell,j,\hat{z}_j}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,\hat{z}_j}\Big] \\
&= \sum_{k\in[K]} \sum_{|\ell|\leq\ell_0} 2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)|\hat{J}_{1,j,k}|\overline{\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell v_j/T}}\phi_{\ell,\hat{z}_j,k}\Big] \\
&= \sum_{k\in[K]} \sum_{|\ell|\leq\ell_0} 2\mathrm{Re}\Big[(\mathbf{i}2\pi\ell/T)|\hat{J}_{j,k}|\cdot\overline{N}_{j,k}(T)\cdot\overline{\overline{c}_{\ell,j,k}e^{\mathbf{i}2\pi\ell(v_j/T)}}\phi_{\ell,\hat{z}_j,k}\Big],
\end{aligned} \tag{B.45}
$$

where in the first equality we merge the two summations into one, and in the second equality we plug in the definition of $\hat{J}_{1,j,k}$.

Bibliography

C. Abraham, P.-A. Cornillon, E. Matzner-Løber, and N. Molinari. Unsupervised curve clustering using b-splines. *Scandinavian journal of statistics*, 30(3):581–595, 2003.

F. Barone-Adesi, L. Vizzini, F. Merletti, and L. Richiardi. Short-term effects of italian smoking regulation on rates of hospital admission for acute myocardial infarction. *European heart journal*, 27(20):2468–2472, 2006.

K. J. Beath. Infant growth modelling using a shape invariant model with random effects. *Statistics in medicine*, 26(12):2547–2564, 2007.

M. Benedicks. On fourier transforms of functions supported on sets of finite lebesgue measure. *Journal of Mathematical Analysis and Applications*, 106(1):180–183, 1985. ISSN 0022-247X. doi: https://doi.org/10.1016/0022-247X(85)90140-4. URL https://www.sciencedirect.com/science/article/pii/0022247X85901404.

A. Benucci, D. L. Ringach, and M. Carandini. Coding of stimulus sequences by population responses in visual cortex. *Nature neuroscience*, 12(10):1317–1324, 2009.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, July 2000. ISSN 1939-3539. doi: 10.1109/34.865189.

J. Bigot and S. Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *The Annals of Statistics*, 38(4):2422–2464, 2010. doi: 10.1214/10-AOS800.

J. Bigot and X. Gendre. Minimax properties of Fréchet means of discretely sampled curves. *The Annals of Statistics*, 41(2):923–956, 2013. doi: 10.1214/13-AOS1104.

J. Bigot, S. Gadat, T. Klein, and C. Marteau. Intensity estimation of non-homogeneous Poisson processes from shifted trajectories. *Electronic Journal of Statistics*, 7:881–931, 2013. ISSN 1935-7524. doi: 10.1214/13-EJS794.

A. G. Blankenship and M. B. Feller. Mechanisms underlying spontaneous patterned activity in developing neural circuits. *Nature Reviews Neuroscience*, 11(1):18–29, 2010.

H. Boecker, J. Jankowski, P. Ditter, and L. Scheef. A role of the basal ganglia and midbrain nuclei for initiation of motor sequences. *Neuroimage*, 39(3):1356–1369, 2008.

D. Bontemps and S. Gadat. Bayesian methods for the Shape Invariant Model. *Electronic Journal of Statistics*, 8:1522–1568, 2014. ISSN 1935-7524. doi: 10.1214/14-EJS933.

C. Bouveyron and J. Jacques. Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5:281–300, 2011.

M. Bues, M. Steiner, M. Stafflage, and M. Krafft. How mobile in-store advertising influences purchase intention: Value drivers and mediating effects from a consumer perspective. *Psychology & Marketing*, 34(2):157–174, 2017.

A. Capilla, P. Pazo-Alvarez, A. Darriba, P. Campo, and J. Gross. Steady-state visual evoked potentials can be explained by temporal superposition of transient event-related responses. *PloS one*, 6(1):e14543, 2011.

M. S. Cembrowski and N. Spruston. Heterogeneity within classical cell types is the rule: lessons from hippocampal pyramidal neurons. *Nature Reviews Neuroscience*, 20(4):193–204, 2019.

J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *2012 IEEE conference on visual analytics science and technology (VAST)*, pages 143–152. IEEE, 2012.

W. Cheng, I. L. Dryden, and X. Huang. Bayesian registration of functions and curves. 2016.

J.-M. Chiou and P.-L. Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):679–699, 2007. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2007.00605.x.

J.-M. Chiou and P.-L. Li. Correlation-based functional clustering via subspace projection. *Journal of the American Statistical Association*, 103(484):1684–1692, 2008.

D. Chudova, S. Gaffney, E. Mjolsness, and P. Smyth. Translation-invariant mixture models for curve clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 79–88, New York, NY, USA, Aug. 2003. Association for Computing Machinery. ISBN 978-1-58113-737-8. doi: 10.1145/956750.956763.

L. T. Coddington and J. T. Dudman. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nature neuroscience*, 21(11):1563–1573, 2018.

L. T. Coddington and J. T. Dudman. Learning from action: reconsidering movement signaling in midbrain dopamine neuron activity. *Neuron*, 104(1):63–77, 2019.

B. W. Connors and M. J. Gutnick. Intrinsic firing patterns of diverse neocortical neurons. *Trends in neurosciences*, 13(3):99–104, 1990.

D. R. Cox and V. Isham. *Point processes*, volume 12. CRC Press, 1980.

C. M. Crainiceanu, A.-M. Staicu, and C.-Z. Di. Generalized multilevel functional regression. *Journal of the American Statistical Association*, 104(488):1550–1561, 2009.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer, New York, second edition, 2003. ISBN 978-0-387-95541-4. doi: 10.1007/b97277.

H. Dana, Y. Sun, B. Mohar, B. K. Hulse, A. M. Kerlin, J. P. Hasseman, G. Tsegaye, A. Tsang, A. Wong, R. Patel, et al. High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature methods*, 16(7):649–657, 2019.

J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-007-9046-7.

C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi. Multilevel functional principal component analysis. *The annals of applied statistics*, 3(1):458, 2009.

J. Diesner and K. M. Carley. Exploration of communication networks from the enron email corpus. In *SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA*, pages 3–14, 2005.

S. Dini, N. M. Douglas, J. R. Poespoprodjo, E. Kenangalem, P. Sugiarto, I. D. Plumb, R. N. Price, and J. A. Simpson. The risk of morbidity and mortality following recurrent malaria in papua, indonesia: a retrospective cohort study. *BMC medicine*, 18:1–12, 2020.

S. Escola, A. Fontanini, D. Katz, and L. Paninski. Hidden markov models for the stimulus-response relationships of multistate neural systems. *Neural computation*, 23(5):1071–1132, 2011.

S. Evans, E. Agnew, E. Vynnycky, J. Stimson, A. Bhattacharya, C. Rooney, B. Warne, and J. Robotham. The impact of testing and infection prevention and control strategies on within-hospital transmission dynamics of covid-19 in english hospitals. *Philosophical Transactions of the Royal Society B*, 376(1829):20200268, 2021.

S. Gaffney and P. Smyth. Joint probabilistic curve clustering and alignment. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

A. Gajardo, S. Bhattacharjee, C. Carroll, Y. Chen, X. Dai, J. Fan, P. Z. Hadjipantelis, K. Han, H. Ji, C. Zhu, H.-G. Müller, and J.-L. Wang. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2021. URL https://CRAN.R-project.org/package=fdapace. R package version 0.5.8.

M. Galvani, A. Torti, A. Menafoglio, and S. Vantini. Funcc: A new bi-clustering algorithm for functional data with misalignment. *Computational Statistics & Data Analysis*, 160:107219, 2021.

C. Gao, Y. Lu, and H. H. Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6): 2624–2652, Dec. 2015. ISSN 00905364. doi: 10.1214/15-AOS1354.

L. A. Garcia-Escudero and A. Gordaliza. A proposal for robust curve clustering. *Journal of classification*, 22(2):185–201, 2005.

D. Giorgi, C. Matias, T. Rebafka, and F. Villers. *ppsbm: Clustering in longitudinal networks*, 2018. URL https://CRAN.R-project.org/package=ppsbm. R package version 0.2.2.

C. Grienberger and A. Konnerth. Imaging calcium in neurons. *Neuron*, 73(5):862–885, 2012.

C. Haythornthwaite. Social network analysis: An approach and technique for the study of information exchange. *Library & information science research*, 18(4):323–342, 1996.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec. 1985. ISSN 01764268. doi: 10.1007/BF01908075.

H. K. Inagaki, S. Chen, M. C. Ridder, P. Sah, N. Li, Z. Yang, H. Hasanbegovic, Z. Gao, C. R. Gerfen, and K. Svoboda. A midbrain-thalamus-cortex circuit reorganizes cortical dynamics to initiate movement. *Cell*, 185(6):1065–1081, 2022.

J. Jacques and C. Preda. Funclust: A curves clustering method using functional random variables density approximation. *Neurocomputing*, 112:164–171, 2013.

J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8:231–255, 2014a.

J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106, 2014b.

G. M. James. Curve alignment by moments. 2007.

L.-E. Jao, B. Appel, and S. R. Wente. A zebrafish model of lethal congenital contracture syndrome 1 reveals gle1 function in spinal neural precursor survival and motor axon arborization. *Development*, 139(7):1316–1326, 2012.

J. J. Jun, N. A. Steinmetz, J. H. Siegle, D. J. Denman, M. Bauza, B. Barbarits, A. K. Lee, C. A. Anastassiou, A. Andrei, Ç. Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

A. Karr. *Point processes and their statistical inference*. Routledge, 2017.

M. Kayano, K. Dozono, and S. Konishi. Functional cluster analysis via orthonormalized gaussian basis expansions and its application. *Journal of classification*, 27:211–230, 2010.

B. Kim, K. H. Lee, L. Xue, and X. Niu. A review of dynamic network models with latent variables. *Statistics surveys*, 12:105, 2018.

D. G. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer New York, New York, NY, 2012. ISBN 978-1-4419-6645-2 978-1-4419-6646-9. doi: 10.1007/978-1-4419-6646-9.

J. Klevens, F. Luo, L. Xu, C. Peterson, and N. E. Latzman. Paid family leave's effect on hospital admissions for pediatric abusive head trauma. *Injury prevention*, 2016.

A. Kneip and J. Engel. Model Estimation in Nonlinear Regression Under Shape Invariance. *The Annals of Statistics*, 23(2):551–570, Apr. 1995. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176324535.

A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, pages 1266–1305, 1992.

D. C. Knill and A. Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719, 2004.

A. O. Komendantov, S. Venkadesh, C. L. Rees, D. W. Wheeler, D. J. Hamilton, and G. A. Ascoli. Quantitative firing pattern phenotyping of hippocampal neuron types. *Scientific reports*, 9(1): 17915, 2019.

A. Kreiß, E. Mammen, and W. Polonik. Nonparametric inference for continuous-time event counting and link-based dynamic network models. *Electronic Journal of Statistics*, 13(2):2764–2829, 2019. ISSN 1935-7524. doi: 10.1214/19-EJS1588.

P. N. Krivitsky and M. S. Handcock. A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12014.

J. M. Krueger, D. M. Rector, S. Roy, H. P. Van Dongen, G. Belenky, and J. Panksepp. Sleep as a fundamental property of neuronal assemblies. *Nature Reviews Neuroscience*, 9(12):910–919, 2008.

B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, et al. Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. *Science*, 352(6293):1586–1590, 2016.

C. Lee and D. J. Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

J. Lee, T. R. Darlington, and S. G. Lisberger. The neural basis for response latency in a sensory-motor behavior. *Cerebral Cortex*, 30(5):3055–3073, 2020.

J. Lei, K. Chen, and B. Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, Mar. 2020. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asz068.

M. Levakova, M. Tamborrino, S. Ditlevsen, and P. Lansky. A review of the methods for neuronal response latency estimation. *Biosystems*, 136:23–34, 2015.

L. Lin, Z. He, and S. Peeta. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, 97:258–276, 2018.

X. Liu and M. C. Yang. Simultaneous curve registration and clustering for functional data. *Computational Statistics & Data Analysis*, 53(4):1361–1376, Feb. 2009. ISSN 01679473. doi: 10.1016/j.csda.2008.11.019.

L. Longepierre and C. Matias. Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model. *Electronic Journal of Statistics*, 13(2):4157–4223, 2019. ISSN 19357524. doi: 10.1214/19-EJS1624.

P. Loupos, A. Nathan, and M. Cerf. Starting cold: The power of social networks in predicting non-contractual customer behavior. *Available at SSRN 3001978*, 2019.

Z. Lu and W. Lou. Shape invariant mixture model for clustering non-linear longitudinal growth trajectories. *Statistical Methods in Medical Research*, 28(12):3769–3784, Dec. 2019. ISSN 0962-2802. doi: 10.1177/0962280218815301.

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM transactions on computational biology and bioinformatics*, 1(1):24–45, 2004.

H. Markram, M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu. Interneurons of the neocortical inhibitory system. *Nature reviews neuroscience*, 5(10):793–807, 2004.

C. Matias and V. Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4): 1119–1141, Sept. 2017. ISSN 13697412. doi: 10.1111/rssb.12200.

C. Matias, T. Rebafka, and F. Villers. A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680, 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy016.

M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019.

J. P. McIntire, L. K. McIntire, and P. R. Havig. Methods for chatbot detection in distributed text-based communications. In *2010 International Symposium on Collaborative Technologies and Systems*, pages 463–472. IEEE, 2010.

E. Menelaou and D. L. McLean. Hierarchical control of locomotion by distinct types of spinal v2a interneurons in zebrafish. *Nature communications*, 10(1):4197, 2019.

S. Mews, B. Surmann, L. Hasemann, and S. Elkenkamp. Markov-modulated marked poisson processes for modeling disease dynamics based on medical claims data. *Statistics in Medicine*, 2023.

B. J. Molyneaux, P. Arlotta, J. R. Menezes, and J. D. Macklis. Neuronal subtype specification in the cerebral cortex. *Nature reviews neuroscience*, 8(6):427–437, 2007.

J. S. Morris and R. J. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2):179–199, 2006.

A. M. Mudge, K. Kasper, A. Clair, H. Redfern, J. J. Bell, M. A. Barras, G. Dip, and N. A. Pachana. Recurrent readmissions in medical patients: a prospective study. *Journal of Hospital Medicine*, 6(2):61–67, 2011.

C. R. Nicholas, J. Chen, Y. Tang, D. G. Southwell, N. Chalmers, D. Vogt, C. M. Arnold, Y.-J. J. Chen, E. G. Stanley, A. G. Elefanty, et al. Functional maturation of hpsc-derived forebrain interneurons requires an extended timeline and mimics human neural development. *Cell stem cell*, 12(5):573–586, 2013.

H. Nishimaru, C. E. Restrepo, J. Ryge, Y. Yanagawa, and O. Kiehn. Mammalian motor neurons corelease glutamate and acetylcholine at central synapses. *Proceedings of the National Academy of Sciences*, 102(14):5245–5249, 2005.

M. W. Oram, D. Xiao, B. Dritschel, and K. R. Payne. The temporal resolution of neural codes: does response latency have a unique role? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424):987–1001, 2002.

A. E. Orhan and W. J. Ma. Neural population coding of multiple stimuli. *Journal of Neuroscience*, 35(9):3825–3841, 2015.

B. Osatuyi. Information sharing on social media sites. *Computers in Human Behavior*, 29(6): 2622–2631, 2013.

L. S. Overstreet-Wadiche and G. L. Westbrook. Functional maturation of adult-generated granule cells. *Hippocampus*, 16(3):208–215, 2006.

S. Paul and Y. Chen. A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *Annals of Applied Statistics*, 14(2): 993–1029, 2020.

D. M. Pavlović, B. R. Guillaume, E. K. Towlson, N. M. Kuek, S. Afyouni, P. E. Vértes, B. T. Yeo, E. T. Bullmore, and T. E. Nichols. Multi-subject stochastic blockmodels for adaptive analysis of individual differences in human brain network cluster structure. *NeuroImage*, 220:116611, Oct. 2020. ISSN 10959572. doi: 10.1016/j.neuroimage.2020.116611.

M. Pensky. Dynamic network models and graphon estimation. *The Annals of Statistics*, 47(4): 2378–2403, Aug. 2019. ISSN 0090-5364. doi: 10.1214/18-aos1751.

D. H. Perkel, G. L. Gerstein, and G. P. Moore. Neuronal spike trains and stochastic point processes: I. the single spike train. *Biophysical journal*, 7(4):391–418, 1967.

D. Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140, 1981.

J. O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(2):351–363, 1998.

J. K. Rogers, S. J. Pocock, J. J. McMurray, C. B. Granger, E. L. Michelson, J. Östergren, M. A. Pfeffer, S. D. Solomon, K. Swedberg, and S. Yusuf. Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to charm-preserved. *European journal of heart failure*, 16(1):33–40, 2014.

E. T. Rolls and A. Treves. The neuronal encoding of information in the brain. *Progress in neurobiology*, 95(3):448–490, 2011.

L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. K-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233, May 2010. ISSN 01679473. doi: 10.1016/j.csda.2009.12.008.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, New York, Oct. 2017. ISBN 978-1-315-14091-9. doi: 10.1201/9781315140919.

M. Sims, R. Maxwell, L. Bauld, and A. Gilmore. Short term impact of smoke-free legislation in england: retrospective analysis of hospital admissions for myocardial infarction. *Bmj*, 340, 2010.

Y. B. Slimen, S. Allio, and J. Jacques. Model-based co-clustering for functional data. *Neurocomputing*, 291:97–108, 2018.

J. Song, I. Pallucchi, J. Ausborn, K. Ampatzis, M. Bertuzzi, P. Fontanel, L. D. Picton, and A. El Manira. Multiple rhythm-generating circuits act in tandem with pacemaker properties to control the start and speed of locomotion. *Neuron*, 105(6):1048–1061, 2020.

N. A. Steinmetz, C. Koch, K. D. Harris, and M. Carandini. Challenges and opportunities for large-scale electrophysiology with neuropixels probes. *Current opinion in neurobiology*, 50:92–100, 2018.

N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, Dec. 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1787-x.

N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun, M. Pachitariu, M. Bauza, M. Beau, J. Bhagat, C. Böhm, M. Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, 2021.

W. Sun, J. Wang, and Y. Fang. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. 2012.

Y. Tanaka, T. Kurashima, Y. Fujiwara, T. Iwata, and H. Sawada. Inferring latent triggers of purchases with consideration of social effects and media advertisements. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 543–552, 2016.

L. Tang, P. Zeng, J. Qing Shi, and W.-S. Kim. Model-based joint curve registration and classification. *Journal of Applied Statistics*, 50(5):1178–1198, 2023.

D. Telesca and L. Y. T. Inoue. Bayesian hierarchical curve registration. *Journal of the American Statistical Association*, 103(481):328–339, 2008.

R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

S. Tokdar, P. Xi, R. C. Kelly, and R. E. Kass. Detection of bursts in extracellular spike trains using hidden semi-markov point process models. *Journal of computational neuroscience*, 29:203–212, 2010.

J. D. Tucker. *fdasrvf: Elastic Functional Data Analysis*, 2023. URL https://CRAN.R-project.org/package=fdasrvf. R package version 2.0.2.

M. Vimond. Efficient estimation for a subclass of shape invariant models. *The Annals of Statistics*, 38(3):1885–1912, 2010a. doi: 10.1214/07-AOS566.

M. Vimond. Efficient estimation for a subclass of shape invariant models. 2010b.

V. V. Vyazovskiy and K. D. Harris. Sleep and the single neuron: the role of global slow oscillations in individual cell rest. *Nature Reviews Neuroscience*, 14(6):443–451, 2013.

V. V. Vyazovskiy, U. Olcese, E. C. Hanlon, Y. Nir, C. Cirelli, and G. Tononi. Local sleep in awake rats. *Nature*, 472(7344):443–447, 2011.

Y. Wan, Z. Wei, L. L. Looger, M. Koyama, S. Druckmann, and P. J. Keller. Single-cell reconstruction of emerging population activity in an entire developing circuit. *Cell*, 179(2):355–372.e23, 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.08.039.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and its application*, 3:257–295, 2016.

P. Wenner and M. J. O'Donovan. Mechanisms that initiate spontaneous network activity in the developing chick spinal cord. *Journal of neurophysiology*, 86(3):1481–1498, 2001.

S. Wu, H.-G. Müller, and Z. Zhang. Functional data analysis for point processes with rare events. *Statistica Sinica*, pages 1–23, 2013.

E. P. Xing, W. Fu, and L. Song. A state-space mixed membership blockmodel for dynamic network tomography. *The Annals of Applied Statistics*, 4(2):535–566, 2010. ISSN 1932-6157. doi: 10.1214/09-AOAS311.

G. Xu, M. Wang, J. Bian, H. Huang, T. R. Burch, S. C. Andrade, J. Zhang, and Y. Guan. Semi-parametric learning of structured temporal point processes. *The Journal of Machine Learning Research*, 21(1):7851–7889, 2020.

K. S. Xu. Stochastic block transition models for dynamic networks. *Journal of Machine Learning Research*, 38:1079–1087, 2015. ISSN 15337928.

K. S. Xu and A. O. Hero. Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal on Selected Topics in Signal Processing*, 8(4):552–562, 2014. ISSN 19324553. doi: 10.1109/JSTSP.2014.2310294.

L. Xu, J. A. Duan, and A. Whinston. Path to purchase: A mutually exciting point process model for online advertising and conversion. *Management Science*, 60(6):1392–1412, 2014.

T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin. Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. *Machine Learning*, 82(2):157–189, Feb. 2011. ISSN 08856125. doi: 10.1007/s10994-010-5214-7.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100(470):577–590, June 2005. ISSN 0162-1459. doi: 10.1198/016214504000001745.

L. Yin, G. Xu, H. Sang, and Y. Guan. Row-clustering of a point process-valued matrix. *Advances in Neural Information Processing Systems*, 34:20028–20039, 2021.

A. H. Zadeh and R. Sharda. Modeling brand post popularity dynamics in online social networks. *Decision Support Systems*, 65:59–68, 2014.

H. Zeng and J. R. Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9):530–546, 2017.

J. Zhang, W. W. Sun, and L. Li. Mixed-effect time-varying network model and application in brain connectivity analysis. *Journal of the American Statistical Association*, 115(532):2022–2036, Oct. 2020. ISSN 0162-1459. doi: 10.1080/01621459.2019.1677242.

W. Zong, H. A. Obenhaus, E. R. Skytøen, H. Eneqvist, N. L. de Jong, R. Vale, M. R. Jorge, M.-B. Moser, and E. I. Moser. Large-scale two-photon calcium imaging in freely moving mice. *Cell*, 185(7):1240–1256, 2022.