Inference for Neuronal Networks using Temporal and Graphical Models

by

Kohinoor Dasgupta

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Vijay N. Nair, Co-Chair
Assistant Professor Xuanlong Nguyen, Co-Chair
Associate Professor Stilian, A. Stoev, Co-Chair
Associate Professor Edward L. Ionides
Assistant Professor Victoria Booth

To my family

ACKNOWLEDGEMENTS

I have been fortunate to have received the guidance and have worked under the supervision of multiple faculty members over the course of my graduate study. I am grateful to the co-chairs of my committee for their time and generous advice and patience. I would like to thank Professor Vijay Nair for his advice and helping me take informed decisions both academic and non academic over the course of my five years at the Department. He had originally introduced me to the applications of statistics to neuroscience and has guided me all along my dissertation work. Professor Long Nguyen has encouraged me to apply graphical models in my research and introduced me to a branch of statistics I was previously unaware of. Under the guidance of Professor Stilian Stoev, I have gained experience in stochastic modeling and would like to thank him for explaining the theoretical details with great patience. Working with faculty members having varied areas of expertise has helped me gain experience in different facets of statistics and has been instrumental in my growth as a statistician and researcher.

I would like to thank Professors Edward Ionides and Victoria Booth for kindly serving on my committee. I have taken several courses with them and the courses on stochastic processes and computational neuroscience were especially interesting and informative. I am grateful to Dr. K.P. Unnikrishnan for his scientific insights regarding novel approaches of analyzing spike train data. The work of the second chapter was partially supported through a project funded by General Motors R & D Center, Warren for which he was the PI. I would also like to thank Dr. Casey

Diekman for guiding me all along and also for building my interest in computational neuroscience. He is a terrific researcher and a very good friend! I thank my fellow students in Statistics department for making my time at Michigan so enjoyable, specially my office mates Joel Vaughan and Toshiya Hoshikawa for helping me in every way possible and for making the office ambience both productive as well as relaxing.

I would like to thank and acknowledge my family for their love, encouragement and support over the course of my PhD life. My parents and brother Surya have supported me in every way possible my entire life and being so far away from them over the last few years has only reiterated their importance in my life. A special note of appreciation for Samopriyo for cheering me in my gloomy days and being someone I can always rely upon. I feel privileged and lucky to have got the company of friends as well as the support and encouragement from my advisors as well as the company of friends both in and outside the department.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

A brain tissue is composed of many neurons which function and interact with each other by generating a time sequence of characteristic electrical pulses known as action potentials or spikes. These time sequences of spikes generated by a set of neurons are referred to as multi-neuronal spike train data. Such data contain the stochastic firing (or spike generating) events of individual neurons as well as spiking activity due to coordinated functioning of a group of neurons which are functionally interconnected. Analyzing multi-neuronal data to identify the spatio-temporal network structure of the functional connectivity of the neurons underlying a specific brain activity is one of the biggest challenges in neuroscience.

Experimenters may not only characterize neuronal activity in anatomically well-defined regions, but they can also examine dynamics of neuronal response and their relationship to behavior. Although elementary methods of data analysis such as t-tests or visual examination of the peristimulus time histogram (PSTH) remain useful for many purposes, the growing complexity of neuroscientific experiments, often examining subtle changes on a comparatively fine timescale, requires careful attention to statistical methods for data analysis.

In neurophysiological experiments, individual spikes are not directly recorded. This is because when multiple electrodes are implanted, the extracellular voltage

potentials recorded on any electrode represent the simultaneous electrical activity of an unknown number of neurons. To identify the spike train data from these voltage traces is a process called "spike sorting". There are two stages in the spike sorting process. In the first stage the spike events or action potentials must be identified from the voltage traces and in the second stage the number of neurons being recorded must be determined and each spike must be assigned to the neuron that produced it. The spike sorting is the mandatory first step in all multiple spike train data analyses.

The accuracy of the spike sorting critically affects the accuracy of all subsequent analyses. Many algorithms are used for spike sorting and at present, there is no consensus as to which are the best (Harris et al. (2000), Fee et al. (1996)). Different algorithms using a variety of methods such as Gaussian errors for model based parametric algorithms have been proposed (Lewicki (1988)). A Monte Carlo based strategy has also been proposed recently in Harris et al. (2000). Spike sorting is a complex problem and is an open area of research.

In this dissertation, we assume that the data have been spike sorted and that the firings can be attributed to individual neurons. If this is not the case, we can treat the term "neurons" as other appropriate "units". This dissertation makes several statistical contributions to modeling and analyzing multi-neuronal spike train data to characterize the functional connectivity among a set of neurons. Throughout, we assume stationary firing rates. The methods here can be extended to piecewise stationary and slowly time-varying non-stationary models.

The first part deals with time sequences of firings of a group of neurons and models them as multivariate point processes. The research goal here is to develop computationally fast methods of inference based on the frequent episodes and examine their efficiencies in a broad range of models. We examine connections with delays and develop procedures for detecting connections between pairs of neurons that are stronger than a given threshold. Methods for screening out false positives that occur in

chains and problems that arise with multiple testing are also considered. We showed applications of the results to simulated multi-neuronal data and demonstrated the usefulness of the results on a real spike train data set.

The second contribution develops a new class of temporal models for characterizing Markovian dependence between a set of neurons. Likelihood-based inference methods and associated algorithms are developed for estimating the connectivity matrix and baseline firing rates for each neuron. We also develop the asymptotic theory associated with the estimators and demonstrate the usefulness of the models and methods on simulated neuronal networks. The results are also applied to real spike train data from cultures of cortical neurons.

The final part develops a general graphical modeling framework and extends the work in the second contribution to also study delays. An iterative algorithm is developed to estimate the delay and connectivity strength matrices for a group of neurons, and the results are used to select the appropriate model and network structure. The efficiency of the method is studied and the results are applied to several simulated examples as well as the data from cultures of cortical neurons.

There exist many representations of univariate and multivariate point processes in terms of conditional intensity function (Kass and Ventura (2001), Chornoboy et al. (1988)). The general conditional intensity framework allows inclusion of a wide variety of measured effects, including trial-to-trial variation terms and the local field potential (LFP) (Truccolo et al. (2005), Paninski (2011)). Local field potential (LFP) is an electrophysiological signal which is the result of the activity of a group of neurons. LFP records the sum of synaptic activity of multiple neurons and tends to be an oscillation on a slower scale than spiking. It is possible to use LFP data as a supplemental data in improving the efficiency of inference from spike train data.

Nevertheless, we do not consider the use of LFP data here. Finally, the methods developed here are illustrated on only small set of neurons. Some of the methods here

can scale up to thousands of neurons but others only to hundreds. Further work is needed to develop algorithms and techniques that will extend our results to situations involving larger neuronal networks.

CHAPTER II

Statistical Inference of Functional Connectivity in Neuronal Networks using Frequent Episodes

2.1 Introduction

An animal's central communication system is composed primarily of neurons (nerve cells) which are responsible for the accumulation, processing and transmission of information. A neuron is composed of dendrites, which are susceptible to chemical stimuli, the cell body or soma and axon. A "message" received by the dendrites propagates to the soma and is eventually communicated to the junctions of neuronal networks or synapses via the axon. The message is actually an electric pulse called firing of a neuron or spike of a neuron. The sequence of spikes of a neuron is referred to as a "spike train" and it may carry important information processed by the brain and underlying cognitive functions and sensory perception.

Neurons can fire spontaneously with no outside stimulus. The firing can also be influenced by other neurons through synaptic connections. These connections are usually of two types: excitatory and inhibitory, depending on whether the firing of one neuron makes the firing of a second neuron either more likely or less likely. Refractoriness is another phenomenon in a neuron due to which after one firing of a neuron, the chance of its firing immediately after that is reduced (perhaps to zero). The time sequences of spikes generated by a set of neurons is referred to as multi-neuronal spike train data. Analyzing the multi-neuronal data to identify the spatio-

temporal network structure of the functional connectivity of the neurons underlying a specific brain activity is one of the biggest challenges in neuroscience (Brillinger (1988), Brillinger (1992), Brillinger and Villa (1997)).

There have been many technical advances, such as multi-electrode array (MEA), that have led to new methods for collecting and storing huge amount of electrophysiological data with finer time scale. Figure 2.1 shows an example of MEA data. The x-axis in Figure 2.1 shows the time in seconds and the y-axis shows the neurons. Each black bar in the plot shows the time at which a particular neuron has spiked. This has resulted in increased interest in examining the dynamics of neuronal response and their relationship to behavior patterns. Elementary methods of data analysis and visual examination of the cross-correlogram or peristimulus time histogram (PSTH) are adequate in many cases. However, the growing complexity and enormity of neuroscientific experiments and examination of subtle changes on a comparatively fine timescale requires the development and use of more advanced research techniques (Brown et al. (2004), Kass and Ventura (2001)).

This chapter of the dissertation focuses on determining connectivity by developing methods to determine significant interactions among the neurons. The research goal here is to develop fast and robust methods of inference based on the frequent episodes and examine their efficiencies in a broad range of models described in Section 2.2.1. Frequent episodes are statistical summaries that characterize the directional relationships among neurons. They are formally defined in Section 2.2.2, where we also provide additional background for the data types considered in the dissertation. We develop the distribution theory for the frequent episodes under a stochastic framework. In Section 2.2.3 we develop a method for analyzing the significance of sequential firing patterns that extends beyond the currently available techniques by allowing the null hypothesis to include "weak dependence" among neurons and by rank ordering significant patterns according to the "strength of influence" among participating neu-

rons. The results are then used to develop statistical thresholds for determining the important frequent episodes (those that correspond to strong functional connectivity among the neurons) and to estimate the strengths of the connectivity. Section 2.2.4 considers an appropriate notion of strength and develop methods for determining connectivity above the specified measure of strength. We describe the test of hypothesis in Section 2.2.5. Two issues arise in implementing the methods in practice: a) screening out false positives that occur due to chains; and b) controlling the overall error rates due to multiple testing. Both of these topics are also discussed in Section 2.2.6 and 2.2.7. We demonstrate the effectiveness of our method on simulated as well as real neuronal network data. We show how the results can be used to recover the network structure on two simulated networks in Sections 2.3.1. Section 2.3.2 illustrates application of the methods to multi-electrode array data from cortical cultures. The chapter concludes with Section 2.4 discussing how the results can be extended.

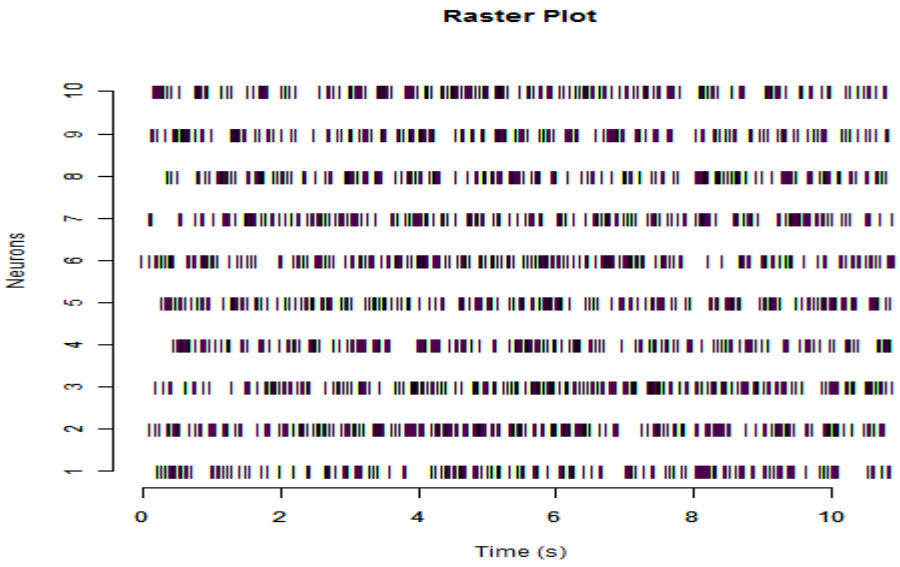

Figure 2.1: Spike recordings of Neurons

2.2 Methods

2.2.1 The Bernoulli Model

The results in Chapter II are developed under a simple stationary model for firing of neurons and their dependence. We start from the simplest possible model, a homogeneous Poisson process, with constant firing rate λ_j for neuron j, $j = 1, ..., J$. In this case the spike generating process is said to be *stationary*. If the neurons fire independently, then the J Poisson processes are independent. It is convenient to work with a discretized version of these processes. Let Δ be a small enough interval so that there is at most one spike in an interval. We observe the processes over a period of time, and suppose there are L intervals during this period. Assume for now that the neurons fire independently. So neuron j either fires or not in interval k, and we have a Bernoulli process $X_j(k)$, $k = 1, ..., L$, $j = 1, ..., J$. Let p_j be the probability of firing in an interval $(= 1 - exp(-\lambda_j \Delta t))$ and denote by N_j the number of firings during this period for neuron $j = 1, ..., J$. The N_j's are independent binomial$(L,\ p_j)$ random variables.

To describe the dependence, consider two neurons A and B with a one-directional structure: A influences B, written as $A \to B$. We say that the *episode* $A[k]B$ occurs if a firing of A is followed by a firing of B after k time units. If A and B are independent, the probability of observing $A[k]B$ is just $P_A \times P_B$. However, if the firing of A leads to an excitation of neuron B and it is more likely to fire after k time units, then $P(A[k]B)$ will be higher. Similarly, if the effect is inhibitory, it will be lower. Throughout this section, we will focus on excitatory effects. Neurons have a particular synaptic delay depending on the neuro-transmitter they have, so one neuron typically affects another neuron to which is connected to it after some delay. For simplicity, we will assume that the time delays between the neurons are fixed at a given value k. This is what is typically done in other papers in the literature (Sastry and Unnikrishnan (2010)).

In practice, one will analyze the data with different values of k and then combine the evidence for dependence across the values of k for a given pair of neurons. We will come back to the issue at the end of this section.

Let $P_k(B|A)$ denote the conditional probability of $A[k]B$, i.e., B fires at time $j+k$ given that A fired at j. From now on we will drop the suffix k from $P_k(B|A)$ for convenience. Define $s = P(B|A)/P_B$. Let s_0 be a user-defined threshold that is larger than one. We say that the (excitatory) influence of A on B is 'strong' if $s > s_0$. Since $A \to B$, the firing rate of A is unaffected and the firings are iid Bernoulli with rate P_A. To get the distribution of the firings of B, note that it can fire on its own or it can fire because it is excited by the firing of A; thus, it can be viewed as a mixture of two processes. Since each of these processes is independent over time, the combined firings in different intervals remain independent. We can compute the mixture probability of B firing in any interval as follows:

$$\begin{aligned} P(X_{B_t}=1) &= P(X_{B_t}=1|X_{A_{t-k}}=1)P(X_{A_{t-k}}=1) \\ &+ P(X_{B_t}=1|X_{A_{t-k}}=0)P(X_{A_{t-k}}=0) \\ &= P_A P_{B|A} + (1-P_A)P_{B|\bar{A}} \\ P(X_{B_t}=0) &= 1 - P(X_{B_t}=1) \end{aligned}$$

Further, X_{B_t} is marginally a Bernoulli process with $P_B^c = P_A\, P_{B|A} + (1-P_A)\, P_{B|\bar{A}}$, where the superscript c stands for combined. However, (X_{A_t}, X_{B_t}) are now dependent. For $j = 1, ..., L-k$, consider the binary data Z_j that equals 1 if $X_{A_j} = 1$ and $X_{B_{j+k}} = 1$. Now, this is an iid Bernoulli process with success probability $P_E = P_{B|A}P_A$. So the number of occurrences of the episode $A[k]B$ during the period is Binomial$(L-k+1,\ P_E)$.

The formulation here using dependent binary processes is simple, but they have been found to be a reasonable approximation for the firing statistics of cortical neu-

rons in other studies (Shadlen and Newsome (1998)). However, neurons do exhibit time-varying behavior, and the validity of a Poisson assumption should be tested empirically (Kass et al. (2005)). If the temporal behavior is piece-wise stationary or slowly varying, the results can be applied to segments of time intervals within which the behavior is stationary. Another complication is that, immediately following a spike, a neuron may exhibit a short (typically less than 1 ms) absolute refractory period during which it cannot spike again (Kass and Ventura (2001), Olson et al. (2000)). Such refractory effects become detectable when neurons have high enough firing rates. This means that the Bernoulli process assumption may not be valid right after a firing. However, the critical assumption in this chapter is that the distribution of the number of episodes follow approximately a Bernoulli process, which can still hold even when the spiking of individual neurons do not (Abeles and Gat (2001)).

This result extends readily for multi-neurons provided there is no cycle, i.e., the network structure connecting the neurons is a directed acyclic graph. It does not appear to hold if we have loops or cycles. Consider a 2 neuron-cycle, i.e., A and B are both connected to each other with different delays k and s and with constant firing rate P_A and P_B respectively. Let $\{X_{A_t|B_{t-s}}\}$ be a mixture Bernoulli process consisting of a finite sequence of independent random variables $X_{A_1|B}, X_{A_2|B}, X_{A_3|B}, ...$ such that for each t, the value of $X_{A_t|B}$ is either 0 or 1 accordingly the neuron A has fired at time t or not given that B has fired at time $t-k$ or not. Similarly, $X_{B_t|A_{t-k}}$ will be a mixture of Bernoulli processes describing the firing pattern of neuron B given that A has fired or not at time point $t-k$. Thus, X_{A_t} can influence $X_{B_{t+k}}$ which in turn can influence $X_{A_{t+k+s}}$. Hence we have 2 neurons effecting each other at different delays and we can characterize the dependence structure of the spiking process of A as:

$$\begin{pmatrix} X_{A_0} & X_{A_{k+s}} & X_{A_{2(k+s)}} & \cdots \\ X_{A_1} & X_{A_{k+s+1}} & X_{A_{2(k+s)+1}} & \cdots \\ \vdots & \vdots & \vdots & \\ X_{A_{k+s-1}} & X_{A_{2(k+s)+1}} & X_{A_{3(k+s)-1}} & \cdots \end{pmatrix}$$

Here, the random variables in each column are distributed as independent Bernoulli but may have different success probabilities. The random variables in each row are distributed as dependent Bernoulli but the random variables in different row are independent of each other. For example X_{A_0} is independent of X_{A_1} as well as $X_{A_{k+s+1}}$ but dependent on $X_{A_{k+s}}$. So, the marginal Bernoulli assumption does not hold.

To demonstrate how these results are used, suppose we observe the number of firings over a time period with L bins, each of size Δ. Let P_A be the probability that neuron A fires in an interval. Define P_B similarly. Suppose we are interested in the event $E = A[k]B$, i.e., B fires k time units after A fires. Let $I_E(t) = 1$ if E occurs at time t, i.e., A fires at time t and B fires at time $t + k$. Let P_E be the probability of this event at any time t. Then, $P_E = P(B|A) \times P_A$. In other words, $P_E = sP_B \times P_A$, where s is the strength of the connection.

Suppose we have spike train data on J neurons and the assumptions of our model hold. Then, we can compute the number of frequent episodes of order 2, say $i[k]j$ for pairs of neurons (i, j), frequent episodes for triplets, etc. and use the binomial distribution theory or the corresponding normal approximations to test for significance. Once we have identified a subset of these connections, we can attempt to identify the underlying network. Examples of this process will be discussed later. One will have to use the screening for important frequent episodes hierarchically as done in apriori and other data mining algorithms. We can look for a higher order episode, say, (r,s,t) only if all pairwise interactions were significant.

2.2.2 Frequent Episodes

Enumerating the occurrences of all possible episodes in multiple spike trains is computationally challenging. Instead, counting only the "non-overlapping" occurrences makes the process computationally efficient (Laxman et al. (2005)). Two occurrences of an episode are said to be non-overlapping if their corresponding time periods do not overlap. For example, the first occurrence of $A[k]B$ overlaps with the second one. (See Figure 2.2 for illustration). Figure 2.2 shows three occurrences of the episode $A[k]B$, however at most two of them are non-overlapping. A set of occurrences of an episode is said to be non-overlapping if every pair is non-overlapping.

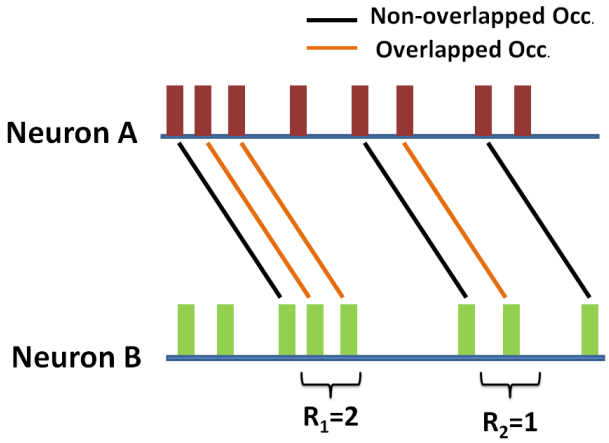

Figure 2.2: Non-overlapping occurrences of episode A[k]B. The total number of occurrences is $N = 6$, consisting of $M = 3$ non-overlapped occurrences and $R_1 + R_2 = 3$ overlapped occurrences.

The restriction to non-overlapping episodes significantly reduces the computational complexity as only N finite state automata are required to count occurrences of an N-node episode (Laxman et al. (2005)). Analysis based on the number of non-overlapping frequent episodes Mannila et al. (1997), has been proposed as a method for characterizing temporal firing patterns from multi-neuronal spike train data (Patnaik et al. (2008)). In this section, we develop statistical inference methods when we have data only on the non-overlapping episodes. Specifically, we develop statistically-

defined thresholds that allow us to distinguish the unimportant ones (random or weak connectivity) from the significant ones, so that only counts above pre-specified thresholds need to be counted. The results here are more general and more efficient than those of Sastry and Unnikrishnan (2010) who use iterative methods to calculate the moments and propose a conservative threshold based on Chebyshev inequality.

In contrast, we can characterize the entire distribution and compute the moment generating function. Further, we estimate the connectivity strength of the neurons as well as the loss of efficiency we incur due to the use of non-overlapping occurrences instead of total occurrences. It is relatively straightforward to compute thresholds under the assumption that the neurons fire independently of each other. It is more interesting (and useful) to compute thresholds under appropriate notions of *weak connectivity*, and look at counts larger than these thresholds, indicating moderate to strong connectivity. We will do this under the notion of strength based on conditional probabilities introduced in Sastry and Unnikrishnan (2010).

To recap, the patterns we consider are ordered firing sequences by groups of neurons with specific time lags or delays between successive neurons in the sequence. Such a pattern (when it repeats many times) may denote a chain of triggering events, and, thus, unearthing such patterns from spike data can help us understand the underlying functional connectivity. Such patterns of ordered firing sequences (with fairly constant delays between successive neuronal firings) have been observed in many experiments, and there is much interest in detecting such patterns and assessing their statistical significance (Gerstein (2004), Abeles and Gat (2001)).

2.2.3 Inferring Connection Strengths

Distribution in the Unrestricted Case

We start with simple case where we are not restricting attention to non-overlapping episodes. For $t = 1, ..., L - k$, consider the binary process where $I_E(t)$ that equals 1 if $X_A(t) = 1$ and $X_B(t + k) = 1$. The data are iid Bernoulli with success probability $P_E = P_A P_{B|A}$. The number of times that $I_E(t)$ equals 1 is equivalent to the number of occurrences of the serial episode $E = A[k]B$, which we denote as N. Note that the possible values of N range from 0 to $(L - k)$; any $A[k]B$ episode where A fires after time $t = L - k$ will not be complete by time $t = L$. So the number of occurrences of the episode E during L intervals has a binomial distribution with parameters $(L - k)$ and P_E. In particular, we have $\hat{P}_E = N/(L - k)$ and $Var(\hat{P}_E) = P_E(1 - P_E)/(L - k)$. This result holds provided there are no cycles among the neurons, i.e., the network structure connecting the neurons is a directed acyclic graph (DAG). If neuron A influences B and B in turn influences A, the model becomes much more complicated. We do not consider such situations in this chapter.

Distribution in the Non-Overlapping Case

Consider now the case where frequent-episode counts are based on non-overlapping occurrences. Consider again two neurons A and B, and let M be the number of *non-overlapping* occurrences of the event $E = A[k]B$. Recall that N is the total number of all occurrences of E discussed in the last sub-section. Sastry and Unnikrishnan (2010) used iterative methods to calculate the mean and variance of M and proposed a conservative testing procedure based on Chebyshev inequality. In contrast, we provide a complete characterization of the distribution of M and compute its moment generating function, which provides expression for moments of all order. This characterization also allows one to simulate the distribution of M easily. We also suggest

a simple approximation to the distribution. We can partition N as follows: consider the j^{th} non-overlapping occurrence of E, and let R_j be the number of episodes of E that occur between the $(j-1)^{th}$ and j^{th} non-overlapping episodes. We illustrate the notion of non-overlapped episodes in Figure 2.2. We have the following.

Proposition II.1. *N has the same distribution as*

$$N = M + \sum_{j=1}^{M} R_j \tag{2.1}$$

Further, the R_j's are iid binomial(k, P_E) random variables and are independent of M. The cumulative distribution function (CDF) or the probability mass function (pmf) of M cannot be obtained analytically. However, we can readily compute the moment-generating function and hence all the moments of M. The mean and variance can be obtained directly. Taking expectations of both sides of Equation 2.1, and using the independence assumption and the distributions of N and R_j's, we have

$$E[M] = \frac{E[N]}{1 + E[X_1]} = \frac{L-k}{1/P_E + k} \tag{2.2}$$

Similarly, the variance of M is

$$Var[M] = \frac{(L-k)P_E(1-P_E)}{(1+kP_E)^3} \tag{2.3}$$

$$= \frac{Var[N]}{(1+kP_E)^3} \tag{2.4}$$

The moment generating function (MGF) of M can be obtained as follows:

$$\begin{aligned} E[\exp(tN)] &= E[\exp(t(M + \sum_{j=1}^{M} R_j))] \\ &= E_M[E_X[\exp(t(M + \sum_{j=1}^{M} R_j))|M]] \\ &= E[\exp(tM)(Q_E \ + \ P_E \ \exp(t))^{Mk}] \\ &= E[\exp(M(t + k\log(Q_E + P_E \exp(t))))] \end{aligned} \tag{2.5}$$

where $Q_E = 1 - P_E$. Let

$$s(t) = t + k\log(Q_E + P_E \exp(t)).$$

Note that this is a one-to-one and strictly increasing function of t, so a unique inverse exists. Denote this inverse function by $g(s)$. (Note that $g(s)$ has to be obtained numerically.) Then, the MGF of M can be expressed as

$$E[\exp(sM)] = E[\exp(g(s)N)] = (Q_E + P_E \exp(g(s))^{L-k}.$$

We can get the k^{th} moment of the distribution of M by differentiating the MGF k times w.r.t. to s and evaluating it at $s = 0$.

The CDF and pmf can be 'estimated' with good precision by Monte Carlo sampling using the characterization of the distribution of M. We examined several other ways of approximating the distribution of M and found a following fourth-order polynomial function of a normal random variable (due to Fleishman, 1978) to be reasonable. The Fleishman technique approximates the distribution of M by $\sum_j a_j Z^j$, where Z is a standard normal random variable and the a_j's are chosen to match the first four moments of M. Table 2.1 provides a comparison of the quantiles of M obtained from this method and quantiles obtained by simulating M directly. Here we use $L = 100s$

and $k = 50ms$ and different values for P_E. Specifically, we simulated 1,000 replications of A and B as independent with 5Hz firing rates, and also with dependence of B on A ($P_{B|A}$=0.2).

Table 2.1: Table showing quantiles of M from different simulations

	From simulated moments of M		From direct simulation of M	
$P_{B\|A}$	0.025 quantile	0.975 quantile	0.025 quantile	0.975 quantile
0.02	27.00	51.03	27.51	51.05
0.10	156.00	204.00	155.28	200.85
0.20	298.96	359.00	297.34	356.75
0.30	423.96	489.00	421.28	488.90
0.40	531.00	596.03	532.90	597.52
0.50	627.00	692.03	627.13	693.06
0.60	710.00	776.00	709.65	777.83
0.70	783.00	850.00	781.18	849.63
0.80	850.96	914.00	849.83	916.57
0.90	912.00	971.03	909.08	975.77

From Equation (2.2), we can get an estimate of P_E based on M as

$$\hat{P}_E = \Big(\frac{L-k}{M} - k\Big)^{-1}, \tag{2.6}$$

a non-linear function of M. We can use Taylor series to approximate its variance as

$$Var[\hat{P}_M] \approx \frac{(1+kP_E)P_{A[k]B}(1-P_E)}{(L-k)}$$

It is of interest to compare the (asymptotic) variances of $\hat{P}_E$ based on N with that from just M. Let $\widehat{P}_E^N$ be the estimator based on N (all the episodes) and $\widehat{P}_E^M$ be based on M (non-overlapping). The ratio of the asymptotic variances is

$$\text{Relative Efficiency} = \frac{1}{1+kP_E}. \tag{2.7}$$

Note that this ratio is independent of L. Further, it decreases as k or P_E get large, suggesting a greater loss in efficiency when one restricts attention to M in these cases.

Of course, these are also the cases where the number of occurrences will be large, and the computational efficiency is greater in using non-overlapping occurrences. Thus, we lose more efficiency as k or P_E or both gets large. Relative efficiency seems to be independent of L since the efficiency of both the estimators increases with L. When k or P_E are large, we lose the efficiency of the estimators but in turn gain computational efficiency by computing non-overlapping occurrences, M instead of total occurrences, N. In Tables 2.2 and 2.3, we calculate the number of overlapped episodes along with the relative loss of efficiency of the estimators for different values of P_E and k.

Table 2.2: Total, non-overlapped, overlapped occurrences and loss of efficiency calculated from the theoretical expressions with L=200s

		Expected # Occurrences (calculated theoretically)			
k	P_E	Total	Non-overlapped	Overlapped	Loss of Efficiency
5	0.0005	99.99	99.75	0.24	0.9975
	0.0010	199.99	199.01	0.98	0.9950
	0.0100	1999.96	1904.72	95.24	0.9524
50	0.0005	99.98	97.54	2.44	0.9756
	0.0010	199.95	190.43	9.52	0.9524
	0.0100	1999.51	1333.01	666.50	0.6667
100	0.0005	99.95	95.19	4.76	0.9524
	0.0010	199.90	181.72	18.18	0.9091
	0.0100	1999.01	999.51	999.50	0.5000
250	0.0005	99.88	88.78	11.10	0.8889
	0.0010	199.75	159.80	39.95	0.8000
	0.0100	1997.51	570.71	1426.80	0.2857

Both the simulation and theoretical results are for $L = 200s$ and all the results are shown for fixed L since both the number of overlapped occurrences and loss of efficiency does not change with L. As we increase k or P_E the number of overlapped episodes increases and thus we gain computational efficiency at the cost of loss of efficiency in estimation. A confidence interval for M given in Table 2.1 can be directly inverted to get a confidence interval or test of hypothesis for P_E and hence the strength s. To demonstrate this, we simulated two neurons A and B with a variety

Table 2.3: Total, non-overlapped, overlapped occurrences and loss of efficiency calculated from 1000 replications of direct simulation with L=200s

k	P_E	Expected # Occurrences (calculated for 1000 replicates) Total	Non-overlapped	Overlapped	Loss of Efficiency
5	0.0005	100.36	100.14	0.22	0.9989
	0.0010	200.76	199.97	0.79	0.9842
	0.0100	2001.29	1924.05	77.24	0.8175
50	0.0005	100.29	97.78	2.51	0.9831
	0.0010	199.68	190.39	9.29	0.9544
	0.0100	2000.78	1341.92	658.86	0.6093
100	0.0005	99.71	94.94	4.77	0.9400
	0.0010	197.76	180.27	17.49	0.8939
	0.0100	2001.68	1005.33	996.35	0.5540
250	0.0005	99.88	88.93	10.95	0.8637
	0.0010	199.59	159.98	39.61	0.8506
	0.0100	1998.27	572.93	1425.34	0.1801

of connections strengths as shown in Table 2.4. For each simulation, we obtained a count of the non-overlapping occurrences of the episode $A[k]B$, which we used to estimate P_E. Table 2.4 shows the empirical coverage probabilities which are very close to the true value of 0.95. Figure 2.3 shows the confidence intervals obtained for 100 replications at two different connections strengths 0.005 and 0.025.

We can also estimate the total number of occurrences, N from the non-overlapping occurrences, M. We know $E(N) = (L - k + 1)P_E$ and we can estimate P_E from the non-overlapping occurrences, M. Then the estimated number of total occurrences $\hat{N}$ will be:

$$\hat{N} = (L - k + 1)\hat{P}_M \tag{2.8}$$

We simulated a spike train data for 6 seconds and calculated the number of total occurrences, N and nonoverlapped occurrences, M. Figure 2.4 shows a quantile-quantile plot (QQ plot) of the actual number of total occurrences, N and estimated number of total occurrences, $\hat{N}$.

Table 2.4: Estimated strength of functional connection $A[k]B$ (1000 replications)

$P_{B\|A}$	$\hat{P}_{B\|A}$	Sample coverage probability of $\hat{P}_{B\|A}$
0.005	0.0052	0.946
0.010	0.0103	0.958
0.025	0.0253	0.941
0.050	0.0501	0.959
0.080	0.0802	0.959
0.100	0.1002	0.956
0.120	0.1204	0.950
0.150	0.1503	0.951
0.180	0.1801	0.935
0.200	0.2003	0.938

2.2.4 Detecting Active Connections

Consider again two neurons A and B. If an excitatory connection is present, $P_{A[k]B} > P_A \times P_B$; for an inhibitory connection, $P_{A[k]B} < P_A \times P_B$. We will restrict attention to the excitatory case as the arguments for the inhibitory case are analogous. The simplest question is to ask if the number of observed episodes (based on N or M) is larger than what one would expect under "randomness". This can be formulated as testing the null hypothesis $P_{A[k]B} = P_A \times P_B$ against the alternative $P_{A[k]B} > P_A \times P_B$. If we let $S = P_{A[k]B}/P_A P_B$, the question becomes testing $H_0 : S = 1$ against $H_1 : S > 1$. The value of S can be viewed as a measure of the connection strength between A and B (Sastry and Unnikrishnan (2010)).

Researchers are typically interested in detecting only connections that are "active" or sufficiently strong, and weak connections may not be of interest. Let S_0 be a user-specified threshold for strength. Let $\tau(S) = P_{A[k]B} - SP_AP_B$. Then, detecting $S \geq S_0$ is equivalent to testing the hypothesis

$$H_0 : \tau(S_0) \leq 0 \quad \text{against} \quad H_A : \tau(S_0) > 0 \tag{2.9}$$

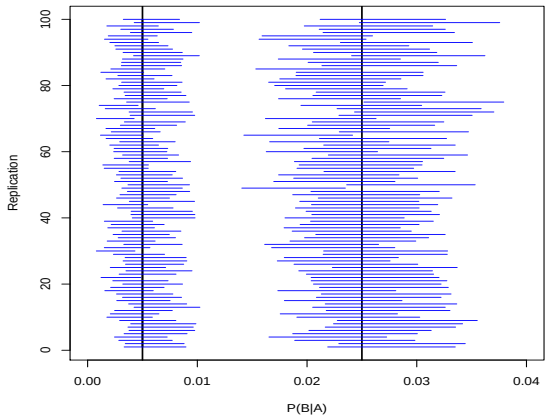

Figure 2.3: 95% confidence intervals for estimate of $P_{B|A}$ based on a single count of M (100 replications) for $P_{B|A}$=0.005 (A and B independent) and $P_{B|A}$=0.025

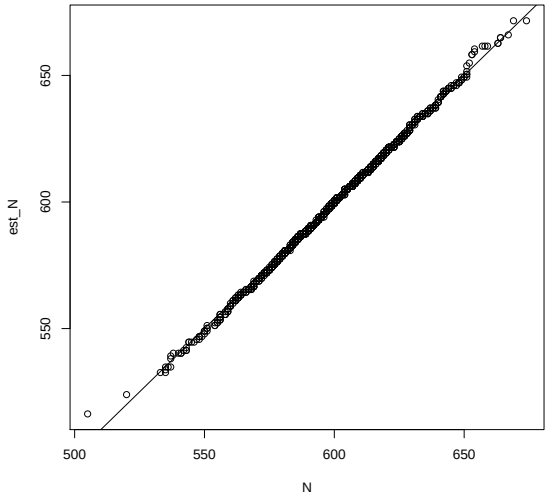

Figure 2.4: QQplot of Simulated N (with $L = 600$, $k = 50$ and $P_{B|A} = 0.2$) and Estimated N from $\hat{P}_M$

In other words, we decide in favor for $S \geq S_0$ only if the null hypothesis is rejected, a decision that requires sufficiently strong evidence.

2.2.5 Test Procedures

Consider first the case where one observes N, the episode count based on all occurrences of $A[k]B$. Then, we can estimate the parameters in $\tau(S)$ based on the underlying binomial data. Denote the corresponding estimator as $\hat{\tau}_N(S)$. The components of its variance are given by:

$$Var(\hat{\tau}_N(S)) = Var(\hat{P}_{A[k]B}) + S^2 Var(\hat{P}_A \hat{P}_B) - 2SCov(\hat{P}_{A[k]B}, \hat{P}_A \hat{P}_B). \qquad (2.10)$$

Expressions for these variances are obtained in the Appendix A. The problem is non-trivial since $\hat{P}_A$ and $\hat{P}_B$ are not independent when $S > 1$. The estimated variance of $\hat{\tau}_N(S)$ can be obtained by substituting the estimates for the unknown parameters in the variance expressions. One can then test the hypothesis of an active connection using normal approximation for the standardized statistic $Z = \hat{\tau}_N(S)/\sqrt{\widehat{Var}(\hat{\tau}_N(S))}$. Note that this is a one-sided test. In practice, one can implement the test procedure by computing the approximate lower confidence bound corresponding to the desired $\alpha-$level for $\tau(S_0)$ and checking if the confidence bound does not include the value zero.

The more technically challenging problem is to develop analogous procedures based on M or equivalently $\hat{\tau}_M(S)$. The formulation of the problem and the structure of the estimator $\hat{\tau}_M(S)$ is exactly the same as before. However, computation of $Var(\hat{\tau}_M(S))$ is now much more involved. Knowledge of the third and fourth order moments of M, discussed in the last section, is needed for this purpose. The details of the variance calculations are developed in Appendix C. An illustration of these procedures is discussed in the next section. We have examined the adequacy of the normal approximations to the test-statistics through simulation studies, and they were adequate for the most part. See Diekman et al. (2012) for details.

2.2.6 Multiple Testing

Recall that the proposed procedure involves testing all pairs of nodes and at all possible values of the delay k. If either J or k is large, we have a large number of tests, and the problem of multiple testing becomes a concern. The researcher will typically not view this as a formal hypothesis testing problem, and so one is not interested in controlling the overall error rate or any of the related quantities discussed in the multiple comparisons literature. Nevertheless, preforming hundreds of thousands of tests will lead to the detection of many false positives. Several approaches have been proposed in the literature, such as the use of Bonferroni's method to control the overall error or the more recent techniques of false detection rates, etc. We used simulation studies to examine the performance of these methods, but there was no single approach that worked better than others. So we propose a heuristic method in the context of specific examples in later sections and use the examples to illustrate the idea.

2.2.7 Eliminating False Edges

When testing for connectivity among all pairs of J neurons, we are likely to detect false edges or connections. Figure 2.5 shows the two kinds of false positives that can occur (the dashed lines). The left panel shows the situation where we have a pair of active connections $A \xrightarrow{k_1} B$ and $B \xrightarrow{k_2} C$. This may lead to the detection of $A \xrightarrow{k_1+k_2} C$ as significant even if it is not an active connection. The right panel shows a different scenario where false edges can be detected. When there are active connections $A \xrightarrow{k_1} B$ and $A \xrightarrow{k_1+k_2} C$, then $B \xrightarrow{k_2} C$ may be detected as active even if it is not. So, we need to do a second-pass through the connections that are detected as active to determine if these are not false edges due to the above phenomenon.

Consider first the case in the left panel. We propose the following approach to detect if $A \xrightarrow{k_1+k_2} C$ is a false edge, given that $A \xrightarrow{k_1} B$ and $B \xrightarrow{k_2} C$ are identified

as active connections. Consider the two quantities: $P_{A\bar{B}C}$, the probability of the episode $A \overset{k_1+k_2}{\longrightarrow} C$ in the absence of B, and $P_A P_{\bar{B}} P_C$, the probability of the same episode under the hypothesis of independence. The difference in these two quantities provides a good measure of whether $A \overset{k_1+k_2}{\longrightarrow} C$ is a true edge. If the probability $P_{A\bar{B}C}$ is significantly higher than $P_A P_{\bar{B}} P_C$, $A \overset{k_1+k_2}{\longrightarrow} C$ is a true edge and is not a byproduct of the episodes $A \overset{k_1}{\longrightarrow} B$ and $B \overset{k_2}{\longrightarrow} C$ occurring sequentially. Thus, consider the following hypothesis to screen out such false edges, $A \overset{k_1+k_2}{\longrightarrow} C$:

$$\begin{aligned} H_0 &: \xi = P_{A\bar{B}C} - P_A P_{\bar{B}} P_C \leq 0 \\ H_1 &: \xi > 0 \end{aligned} \tag{2.11}$$

There are other ways to detect the false edge, but this is a simple and easily implementable procedure that is especially useful when dealing with M, test statistics based on non-overlapping occurrences. The variances of this test statistic are developed in Appendices B and D. As before, one can use the standardized test statistic (test statistic divided by its estimated standard error) and the normal approximation to implement the procedure.

Consider now the second case of false positive in the right panel of Figure 2.5 that is, we have the following true edges: $A \overset{k_1}{\longrightarrow} B$ and $A \overset{k_1+k_2}{\longrightarrow} C$, and we also detected $B \overset{k_2}{\longrightarrow} C$ as active. Our goal is to test if the latter is a false edge. We now use the difference between $P_{\bar{A}BC}$, the probability of the episode $B \overset{k_2}{\longrightarrow} C$ in the absence of A, and the probability of the same episode under independence, namely $P_{\bar{A}} P_B P_C$. The hypothesis test to eliminate this type of false edge is as follows:

$$\begin{aligned} H_0 &: \eta = P_{\bar{A}BC} - P_{\bar{A}} P_B P_C \leq 0 \\ H_1 &: \eta > 0 \end{aligned} \tag{2.12}$$

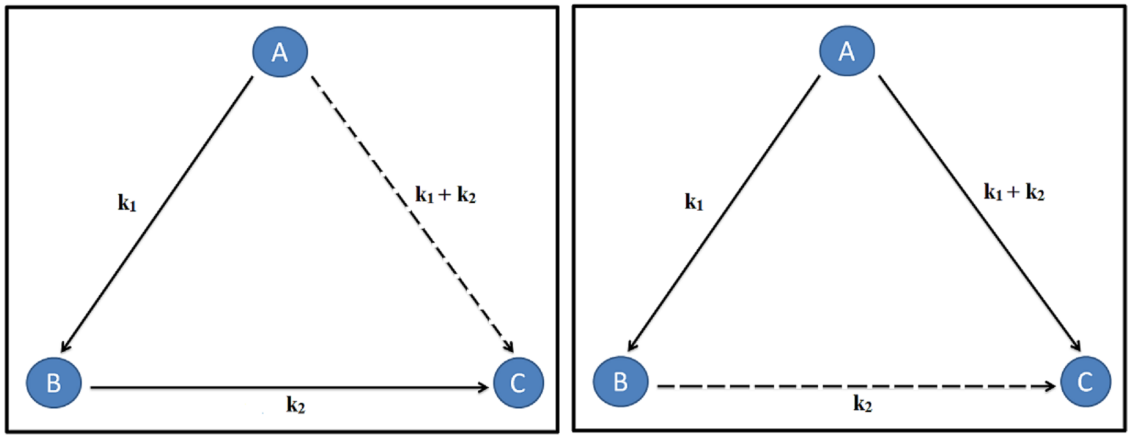

Figure 2.5: A network with 3 neurons and embedded connections illustrating two kinds of false positives (dashed lines) that can occur

2.3 Results

2.3.1 Simulation Study

We use the example with 9 neurons and connections in Figure 2.6 and simulated data to examine the usefulness of our results. In this example, each neuron has an baseline firing rate of 10Hz, with three different 3-neuron connections with delays as shown in Figure 2.6. We varied the connectivity strengths of the 7 edges (connections) at values of S ranging from 1 to 40. The value of $S = 1$ effectively has no connectivity as it implies the neurons spike according to the baseline firing rate.

For each value of S, we simulated 100 data sets from the network. To analyze the data, we first count the total number of spikes from the 9 neurons; these can be viewed as 1-node episodes. We then obtain the count of all possible 2-node serial episodes with delays ranging from 1 to 200 ms. There are 81 pairwise episodes, so this is a total of $81 \times 200 = 16,200$ cases. We can then test for active 2-node connections. Since we are performing $n = 81 \times 200 = 16,200$ tests, we must consider the issue of multiple testing as described in Section 2.2.6.

We implemented our 2-node detection tests as described in Section 2.2.4. More

specifically, we used the procedure in Equation (2.9) with different values of S_0 and examined the number of significant 2-node connections. The mean number of such connections for the 100 replicates are shown in Table 2.5. The row corresponds to different values of the true connection S for the 7 active edges with 2 false positives and the columns correspond to the different values of S_0 used in the test procedure. Consider first the cell corresponding to $S = S_0 = 1$. This is the null case where there are no connections and we test the hypothesis of independence. At significance level is $\alpha = 0.05$, we would expect about $16,200 \times 0.05$ or approximately 800 of these tests to be declared as significant. We see that about 200 connections were identified as significant, so the test procedure is conservative. If we used $S_0 = 2$ for the case with $S = 1$ (zero connections), we see that, on the average, 1 edge is falsely detected as significant; when $S_0 = 3$, we detect zero which is the correct value. Looking at moderate sized values of S (ranging from 2 to 4), we can see that the tests with $S_0 = 2$ miss some of the 7 active edges. When the connection strength S is stronger (above 5), the tests with $S_0 = 2$ detect at least 7 connections. A detailed analysis showed that all the active edges were indeed identified. For high values of S, we also detect some false edges. An inspection of Figure 2.6 shows that we may detect two false edges: $A[100]C$ and $G[10]D$ for the reasons discussed in Section 2.2.7. This was indeed the case here. $A[100]C$ will occur whenever an occurrence of $A[50]B$ is followed by an occurrence of $B[50]C$. Similarly, $G[5]D$ is triggered by the same occurrence of $H[5]G$ and $H[10]D$.

As an example, we discuss the case with $S = 30$ for one particular simulated data set. We applied all the test procedures, including the false-edge detection tests in (2.11) and (2.12). Table 2.6 shows the Z-statistics for all 9 edges. We see that Z_ξ for $A[100]B$ is negative, indicating that the relatively large value of $Z_\tau = 7.85$ was triggered by the chain $A[50]B[50]C$. Consider now the group of neurons D, H and G. Similarly, the value of Z_η for $G[10]D$ is negative, but Z_ξ for $H[30]D$ is large.

This indicates that that the $G[10]D$ edge was falsely detected, caused by the pairs of connections $H[20]G$ and $H[30]D$ (see Figure 2.6). On the other hand, the Z-statistics for all the pairs $E[5]F$, $F[10]I$, and $E[15]I$ are large, indicating that all three pairs of connections are active and there are no false edges here.

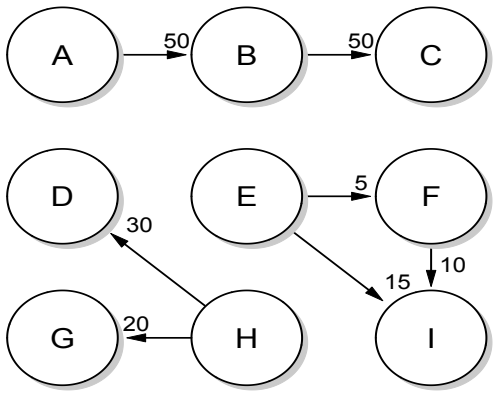

Figure 2.6: A network with 9 neurons and embedded connections. Each neuron had an intrinsic firing rate of 20Hz. The connections had varying strengths and time delays from 5 to 60 ms.

Table 2.5: Showing significant 2 node connections for different S_0 with alpha α= 0.05

S/S_0	1	2	3	4	5	10	15	20	25
1	198.51	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	208.16	1.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	209.75	3.27	0.25	0.00	0.00	0.00	0.00	0.00	0.00
4	217.25	5.07	1.42	0.33	0.00	0.00	0.00	0.00	0.00
5	230.42	7.34	3.44	1.11	0.44	0.00	0.00	0.00	0.00
10	239.33	8.25	7.00	6.88	6.63	0.38	0.00	0.00	0.00
15	241.39	7.67	7.00	7.00	7.00	6.25	0.28	0.00	0.00
20	245.27	7.61	7.00	7.00	7.00	7.00	5.38	0.35	0.00
30	249.71	9.76	8.86	7.14	7.14	7.11	7.02	7.00	4.91
40	265.47	9.14	9.00	8.57	8.57	7.45	7.13	7.00	7.00

Table 2.6: Test statistics for the initial detection test and false edge elimination test for the 9 neuron network

Neuron1[k]Neuron2	Z_τ	Z_ξ	Z_η
A[50]B	15.76	-	-
A[100]C	7.85	-3.70	-
B[50]C	18.04	-	15.72
E[5]F	16.15	-	-
E[15]I	17.85	13.09	-
F[10]I	20.37	-	16.34
G[10]D	8.70	-	-1.38
H[30]D	16.36	13.13	-
H[20]G	17.32	-	-

2.3.2 Analysis of a Real Data Set

To understand and model the interactions of neurons in a network, neurobiologists conduct experiments in which they record the activity of several neurons from a region in the brain, sometimes simultaneously, in response to external stimuli. These experiments are done both in-vivo (on live test subjects) and in-vitro (on cell cultures). In-vitro experiments are conducted on networks of neurons cultured on, for example, a petri-dish. A sophisticated tool called the multielectrode array (MEA) is used to measure the activities of groups of neurons. The neuron cells are cultured on top of a grid of micro-electrodes. Each electrode in the grid is capable of recording activity of one or more neurons around it. A typical MEA setup consists of 8×8 grid of 64 electrodes. These electrodes, besides recording cell activity, can also inject electric charge or external stimulus into the neurons. Some of the special areas of interest for such experiments have been the primary motor cortex which initiate and co-ordinates muscle activity; hippocampus or pre-frontal cortex responsible for cognitive tasks, long term memory and retinal ganglion cells to understand vision encoding.

The output of a MEA setup is in the form of spike trains, a spike train coming from each electrode. The analysis of these recordings involves finding regularities in the spike train data so as to correspond these regularities with neuronal firings

chains of an activity. The nature of the MEA data recordings makes it possible to be expressed as an event stream after suitable pre-processing of the data. We used our methods to analyze spike train data available in the literature on cultures of cortical neurons (Wagenaar et al. (2006)). Their work focused on characterizing "population bursts" in the spiking activity of the cultures. Such bursts, defined as brief periods of time during which the firing rate of several cells or electrodes greatly exceeds the baseline rate, are a common feature observed in cultures of many different types of neurons. Our focus is not to characterize the bursts but rather to detect precisely timed spiking patterns involving multiple neurons in order to estimate the strength of functional connectivity between different neurons in the culture. Since our methods assume that the firing rates of individual neurons are relatively stationary in the analysis window, we selected a segment of the data that did not contain bursts.

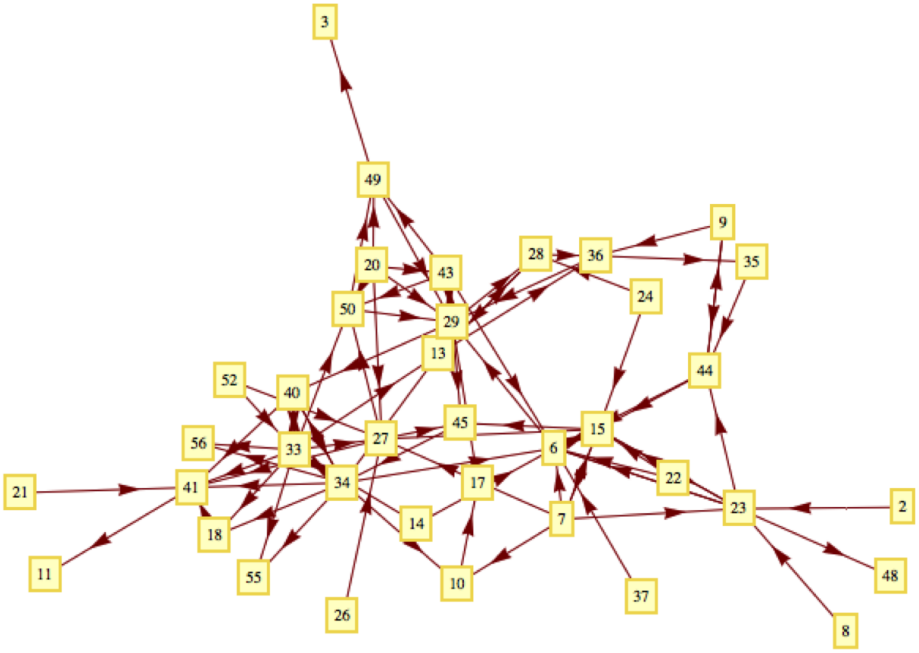

Figure 2.7: Inferred functional connectivity in cortical culture 2-1-34 with $S_0 = 2$.

We analyzed 120 seconds of data from culture 2-1-34. To visualize the functional connectivity present in the culture, we show the network graphs based on our pruned list of significant 2-node episodes. If there is a significant episode with any delay for

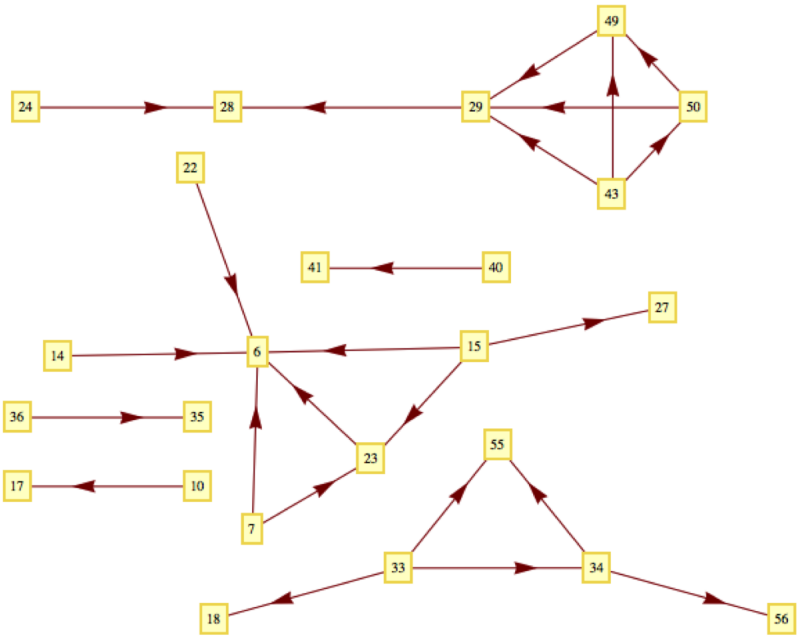

Figure 2.8: Inferred functional connectivity in cortical culture 2-1-34 with $S_0 = 10$.

a pair of neurons (i, j), we draw a directed edge $i \to j$ connecting them. Figures 2.7 and 2.8 show the inferred network structure for two different S_0 thresholds. As S_0 is increased from $S_0 = 2$ to $S_0 = 10$ (Figures 2.7 and 2.8), the weaker connections are no longer significant and the inferred network becomes more sparse. In this way, our methods can be used to identify the strongest microcircuits present in the culture. Although we considered episodes with delays of up to 200ms, all the significant episodes had delays of less than 10ms. Fast delays such as this are consistent with the timescale of the action for AMPA, a common excitatory neurotransmitter in the cortex.

2.4 Discussion

In this chapter, we have presented a statistical framework for discovering functional connectivity of neuronal networks from spike trains based on repeating occurrences of temporally precise patterns of spikes. Most methods for detecting repeated occurrences of precise spike patterns and assessing their significance are based on cor-

relations between time-shifted spike trains (Abeles and Gerstein (1988), Tetko and Villa (2001), Diekman et al. (2009)). In contrast, the framework developed here is based on computationally efficient data mining algorithms for frequent episode discovery. Our framework also allows one to infer episodes that represent strong interactions among neurons. We have demonstrated the effectiveness of our methods both on simulated neuronal spike data as well as on data collected from in vitro neuronal cultures of cortical neurons.

The strength of connection between a pair of neurons is characterized using the conditional probability that one neuron fires after a specified delay given that the other neuron has fired. This delay is usually caused by the axonal delay of spike propagation and the delay at synapses due to chemical diffusion. In our analysis, we have assumed that the inter-event time constraints which represent the time delays in neuronal connections are constants. In general there would be variation in these delays. Also, since we do not have complete control over which specific neurons are recorded from, we may be seeing some connections mediated by more than one synapse. Hence, one very useful extension for the method proposed here is to incorporate variable delays as inter-event time constraints.

Currently, our statistical framework handles only serial episodes and the case of directed acyclic graphs. Another interesting problem would be to extend the statistical theory for the case of more general episodes which may be able to represent additional types of connectivity graphs. For example, Viswanathan et al. (2010) has developed a framework for determining the statistical significance of parallel episodes. Parallel episodes can capture the co-occurrence of spikes from many different neurons within a time-window, a type of synchronous activity of interest to many researchers (Grun et al. (2001), Pipa et al. (2008)). We will be addressing such extensions in our future work.

CHAPTER III

A Class of Models and Likelihood-Based Inference for Assessing Functional Connectivity in Neuronal Networks

3.1 Introduction

There are several points process models characterizing the spiking activity of single neurons and the interactions among small numbers of neurons over time (Brillinger (1988), Brillinger (1992), Brillinger and Villa (1997)). Kass and Ventura (2001) introduced a non-homogeneous Markov interval point process in which memory structure is determined by the inter-arrival times of successive spikes. Kass et al. (2005), Koyama and Kass (2008) proposed the BARS method as a model-based smoother of the instantaneous firing rate function. Brown et al. (2004) and Kass et al. (2005) present their perspectives on the state-of-the-art research in the area of multiple spike trains analysis, concerned with the development of statistical models for the joint firing activity of many neurons over time. These papers indicate two key challenges: estimating the mutual dependence of the firing activity of several neurons over time and modeling the noise induced by spike detection problems. In order to address the former issue, Brillinger and Villa (1997) proposed a discrete time random threshold model where the interactions among neurons are captured via their membrane potential and threshold functions. Recently, Okatan et al. (2005) developed maximum likelihood methods for estimating the functional connectivity of stochastic neuronal

networks based on a discretization of the approach of Chornoboy et al. (1988). Truccolo et al. (2005) proposed a point process framework to relate the spiking probability of neuronal ensembles to the neurons' own spiking history, to the concurrent ensemble activity and to extrinsic covariates, such as external stimuli and behavior. Martignon et al. (2000) modeled the high order interactions among the measured spike trains using log-linear models. So we see that there are many representations of univariate and multivariate point processes in terms of conditional intensity function (see Kass and Ventura (2001), Truccolo et al. (2005), Chornoboy et al. (1988)).

In this chapter we develop a new class of temporal (Markovian) models to characterize the connections between a group of neurons and develop likelihood-based methods of inference for the connectivity matrix and baseline firing rate. We are able to estimate both excitatory and inhibitory connections – a challenge for other methods in the literature (Sastry and Unnikrishnan (2010)). We also derive the asymptotic theory for estimators of the connectivity matrix and baseline firing rates. We illustrate our method on simulated data as well as data from cultures of cortical neurons.

The chapter is organized as follows. In Section 3.2.1, we develop the framework and the model. We discuss the estimation method in Section 3.3. The corresponding asymptotic properties of the parameters are described in Section 3.3.2. In Section 3.4.1 and 3.4.2 we use a simulated network of neurons and illustrate the usefulness of the methods. We also describe the application of the results to multi-electrode array data from cortical cultures in Section 3.4.3.

3.2 A Class of Models

The formulation of the model in discrete time follows the set up in the last chapter. We assume the intervals between discrete time points to be a small enough so that there is at most one spike in an interval. We observe the processes over a period of

time, and suppose there are L intervals during this period. Let J be the number of neurons in the network. Let $X_j(t)$ be the binary random variable with $X_j(t) = 1$ if neuron j fires at time t and $X_j(t) = 0$ otherwise, for $t = 1, ..., L$ and $j = 1, ..., J$.

3.2.1 Conditional Probability and Joint Likelihood

Let $\mathcal{H}_t$ denote the spike train sequence up to time t, namely $\{X_j(s), s \leq t, j = 1, \cdots, J\}$. Let $p_j(t+1|\mathcal{H}_t)$ denote the conditional probability of the firing of the jth neuron at time $t+1$ given the history of firing of all the neurons up to time t. This conditional probability depends on a $J \times 1$ vector b, the set of baseline firing coefficients, and a $J \times J$ connectivity matrix W. The (i, j)th entry $w_{i,j}$ of the connectivity strength matrix W represents the connectivity or the influence of neuron i on neuron j.

Let $\tau_j(t)$ denote the last firing time of neuron j before time t, defined formally as

$$\tau_j(t) = \max\{1 \leq s < t \, : \, X_j(s) = 1\}.$$

Then, our class of models are defines by the conditional probability

$$p_j(t+1|\mathcal{H}_t) = P\{X_j(t+1) = 1|\mathcal{H}_t\} = F\Big(\sum_{i=1}^{J} w_{ij} g(X_i(\tau_j(t) : t]) + b_j\Big), \qquad (3.1)$$

where $F(x)$ is any continuous, increasing cumulative distribution function (CDF) with $x \in (-\infty, +\infty)$ and $X_i(\tau_j(t) : t]$ denotes the set of random variables $(X_i(\tau_j(t), ..., X_i(t))$ for $i, j = 1, ..., J$, and $g(\cdot)$ is some function that maps this set of random variables to the real line. We discuss some specific examples of $g(\cdot)$ below. We assume that the spike train for the jth neuron at time $t+1$, $X_j(t+1)$ is dependent on the spiking history only until time t that is on $\mathcal{H}_t$. In other words, $X_j(t+1)|\mathcal{H}_t$ is conditionally independent of $X_i(t+1)|\mathcal{H}_t$ for all $j \neq i$.

The above defines a class of models for different choices of $F(\cdot)$ and $g(\cdot)$. Each

model specifies how the spiking probability of the j^{th} neuron depends on the past spiking history of itself and of the other neurons to which connect it is connected, as dictated by the connectivity matrix W. The influence is mediated through the function $g(\cdot)$. If the j^{th} neuron is not connected to other neurons in the network, i.e., $w_{ij} = 0$ for all i, then neuron j will have a constant baseline firing rate of b_j. If $w_{ij} > 0$, then a spike of i leads to an increase in the spiking rate of j; $w_{ij} < 0$, then the spikes of i have an inhibitory effect on j. Turning to the diagonal terms, if $w_{jj} < 0$, the neuron j is refractory while $w_{jj} > 0$, neuron j is self-excitatory and can display a bursting behavior. We note that discrete stochastic models similar to Equation (3.1) has been proposed in the literature (Rigat et al. (2006), Truccolo et al. (2005) and Chornoboy et al. (1988)).

In the rest of this paper, we take $F(x) = \exp(x)/(1 + \exp(x))$, the logistic distribution function. The results can be readily extended to other choices. Examples of $g(\cdot)$ include:

1. $g(X_i(\tau_j(t) : t]) = \min(X_i(\tau_j(t)), ..., X_i(t))$ which equals 1 if there is a firing of neuron i in the interval $(\tau_j(t) : t]$ and zero otherwise. This corresponds to the case in Chapter II without accounting for delays.

2. $g(X_i(\tau_j(t) : t]) = \sum_{k=\tau_j(t)+1}^{t} X_i(k)/(t - \tau_j(t) + 1)$, the average number of firings of neuron i within the relevant interval. This model was considered in Rigat et al. (2006) who used Bayesian methods of inference to estimate the model parameters. In this paper, we use maximum likelihood methods.

3. $g(X_i(\tau_j(t) : t]) = \sum_{k=\tau_j(t)+1}^{t} \lambda^{t-k} X_i(k)$ with $\lambda \in (0, 1]$, a model that discounts the influence of the firing of neuron i at time k by the discount factor λ^{t-k}. So recent firings have more weight than those in remote past. The value of λ can also be treated as a parameter.

It turns out that the parameters w_{jj} and b_j are not identifiable in Example 1 above.

General identifiability conditions are discussed in Lemma III.3. Example 2, or Rigat's model, can be viewed as a normalized version of Example 3 with $\lambda = 1$.

We will restrict attention to the choice of $g(\cdot)$ in Example 3, although the results readily extend to other choices also. We will also discuss estimation of λ using likelihood methods. We illustrate this model with the help of a simple network of 2 neurons A and B, where A influences B with a excitatory connection. Figure 3.1 shows the probability of firing of neuron B which is influenced by the spikes of neuron A. The spikes of neurons A (brown stars) and B (green solid blocks) are also plotted at the bottom. We see that the firing probability of neuron B increases as neuron A spikes. The probability of firing only depends on the number of spikes of the neurons connected to that neuron. There is a exponential drop in the firing rate after each peak and the rate of this decrease is controlled by the value of λ.

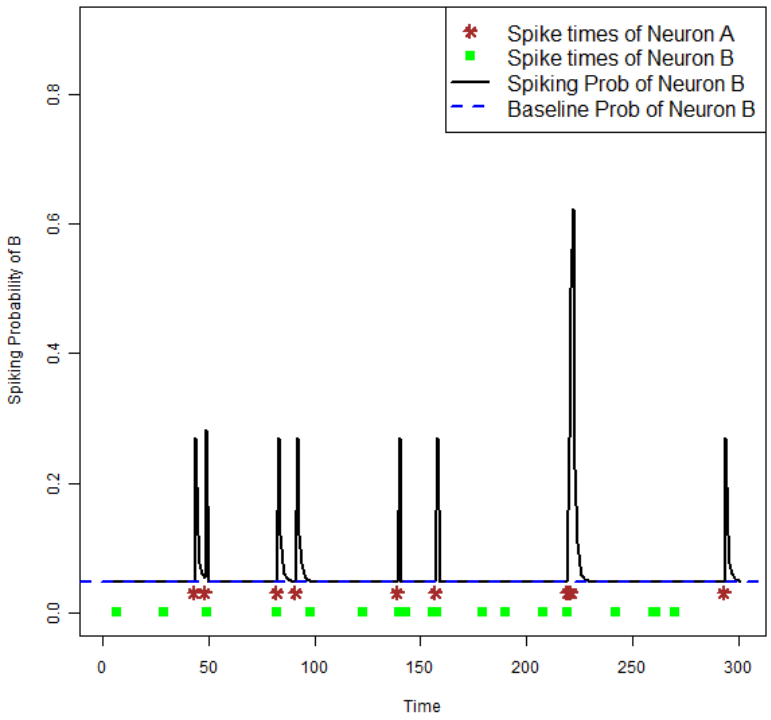

Figure 3.1: Probability of firing of Neuron B given the spikes of Neuron A (star) and Neuron B (solid blocks) for the Discounted Model

3.3 Likelihood-based Inference

The joint likelihood of the data for all neurons can be written as the product of the conditional likelihood for each neuron. The first few spikes for all the neurons in the network are generated from a Bernoulli distribution with the baseline firing rates as the success probability. We define t' as time when all the neurons in the network fire at least once for the first time. $t' = \min_t \left\{ \sum_{s=1}^{t} X_j(s) > 0 \ \text{ for all } \ j \right\}$

Thus, the joint likelihood of the model becomes

$$L = \prod_{k=1}^{J} \prod_{t=t'}^{T-1} \left[p_k(t+1|\mathcal{H}_t)^{X_k(t+1)} (1 - p_k(t+1|\mathcal{H}_t))^{1-X_k(t+1)} \right]$$

where $p_k(t+1|\mathcal{H}_t) = P(X_k(t+1) = 1|\mathcal{H}_t)$ is the conditional probability of spiking for neuron k at time $t+1$ given the history of spiking until time t. We get the expression of $p_k(t+1|\mathcal{H}_t)$ from Equation (3.1).

The log-likelihood then becomes

$$\begin{aligned} \ell_T = \log(L) &= \sum_{k=1}^{J} \Big[\sum_{t=t'}^{T-1} X_k(t+1) \log(p_k(t+1|\mathcal{H}_t)) \\ &+ (1 - X_k(t+1)) \log(1 - p_k(t+1|\mathcal{H}_t)) \Big] \end{aligned} \tag{3.2}$$

Lemma III.1. *The score functions are given by*

$$\begin{aligned} \frac{\partial \ell_T}{\partial w_{ij}} &= \sum_{t=t'}^{T-1} (X_j(t+1) - p_j(t+1|\mathcal{H}_t)) g(X_i(\tau_j(t) : t]) \\ \frac{\partial \ell_T}{\partial b_j} &= \sum_{t=t'}^{T-1} (X_j(t+1) - p_j(t+1|\mathcal{H}_t)) \end{aligned} \tag{3.3}$$

Proof. The proof is sketched in Appendix E. □

Remark: Our derivations apply to general forms of $g(\cdot)$ that are not necessarily as in Equation (3.1).

The score functions are non-linear equations, so the maximum likelihood estimates (MLEs) that have to be solved using numerical methods. We used the iteratively reweighted least squares (IRWLS) algorithm (Agresti (1990)) to obtain the MLEs. Let $u_j^{J\times 1}$ be the derivatives of the log likelihood with respect to the parameters w, for the jth neuron with connectivity parameters $w_j^{J\times 1}$. Then, the ith element of u_j equals (3.4).

$$u_{ij} = \sum_{t=t'}^{T-1} \Big(X_j(t+1) - \frac{\exp(\eta_j)}{1+\exp(\eta_j)} \Big) g(X_i(\tau_j(t) : t])$$

where $\eta_j = \sum_{i=1}^{J} w_{ij} g(X_i(\tau_j(t) : t]) + b_j$. Let $W^{(s)}$ be the estimate of the connectivity matrix at the sth iterative step. Then

$$W_{J\times 1}^{(s+1)} = W_{J\times 1}^{(s)} + (\mathfrak{I}_{J\times J}^{(s)})^{(-1)} u_{J\times 1}^{(s)}$$

where $\mathfrak{I}_{i,j}^{(s)}$ is the sth iterative value of the (i,j)th element of the information matrix, $E\Big[-\frac{\partial^2 \ell_T}{\partial w_{ij} \partial w_{i'j'}} \Big]$. The detailed derivation and theoretical expression of $\mathfrak{I}_{ii'}$ is given in Section 3.3.2. The estimated asymptotic covariance matrix can be obtained by inverting the observed information matrix. The information matrix for our model is of the form

$$\begin{aligned} \mathfrak{I}_{ii'} &= E\Big[-\frac{\partial^2 \ell_T}{\partial w_{ij} \partial w_{i'j'}} \Big] \quad \text{if} \quad j = j' \\ &= 0 \quad \text{if} \quad j \neq j' \end{aligned}$$

The discounting rate λ is estimated by numerically maximizing the profile likelihood for a given value of W and b.

3.3.1 Identifiability Issues

In this section, we consider two special cases of $g(\cdot)$ to illustrate the broad scope of the estimation procedure and to discuss the identifiability issue of the model (3.1). As noted earlier, the model (3.1) is not fully identifiable if we choose $g(\cdot)$ to be an indicator function $g(X_i(\tau_j(t) : t]) = \mathbb{I}\Big\{ \sum_{k=\tau_j(t)+1}^{t} X_i(k) > 0 \Big\}$. In this setting, the baseline parameters b_i are confounded with the the connectivity strength w_{ii} and we will not be able to estimate both separately. We refer to Lemma III.3 in Section 3.3.2 for more details of the theoretical justification with necessary and sufficient conditions for identifiability.

The discounted model, however, is identifiable for all $\lambda \in (0, 1)$. We illustrate this next over an example network which includes three neurons with the configurations shown in Figure 3.2. We embedded both excitatory and inhibitory connections of equal strength affecting neuron A in the network with the connectivity strength between the neurons B and A, $w_{BA} = -1$, and same between the neurons C and A, $w_{CA} = 1$, in the network. We simulate 500 replicates of ten thousand data points or in other words 10 seconds worth of data for each neuron. The baseline firing rates were set to suitable values in order to ensure that the simulated firing rates are comparable to those observed for in-vitro multi-electrode recordings. The baseline coefficient b of each neuron is taken -2.5 so our baseline firing rates become 7.5Hz. We apply our estimation method to analyze the simulated data. We extensively study the joint influence of neurons B and C on neuron A. The histogram of the estimated values of the connectivity strengths and baseline firing rate of the neuron A are shown in Figure 3.3. The vertical black lines indicate the true value of the parameters and the histograms show no discernible bias in estimation. Hence, the study confirms the theoretical finding that there is no identifiability issue or in other words the influences of B and C, even though they are of "opposite" types, are not nullifying the intrinsic properties of A when we use the form of $g(\cdot)$ as shown in Example 3 of Section 3.2.1.

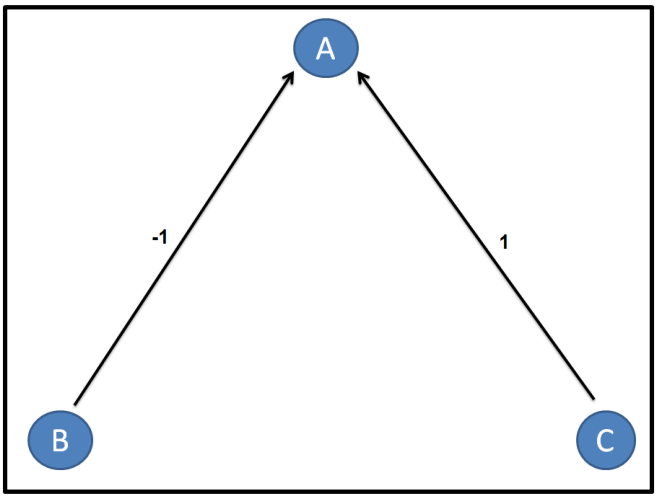

Figure 3.2: A network with 3 neurons with an intrinsic firing rate of 5Hz for each. Neuron B has an inhibitory influence on A and Neuron C has an excitatory influence on A with equal and opposite connectivity strengths

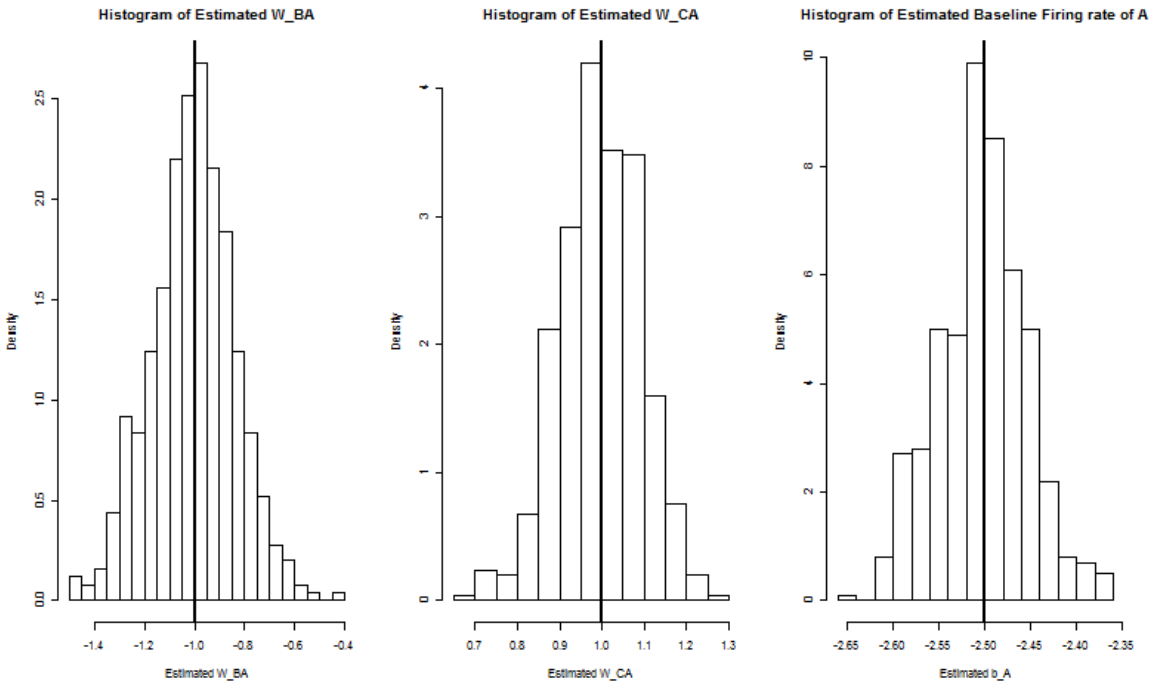

Figure 3.3: Histograms of the Estimated connectivity strengths w_{BA} and w_{CA} and baseline firing rate b_A. The vertical black line shows the true value of the parameters

3.3.2 Asymptotic Properties of the Estimator

In order to prove the asymptotic properties of the parameters, we redefine our parameters as

$$\theta = (W, \vec{b}) \tag{3.4}$$

with $W = (w_{ij})_{J\times J}$, the matrix of connectivity strength, and $\vec{b} = (b_j)_{j=1\cdots,J}$, the vector of baseline coefficients. We also introduce a parameter, $\Delta \in \mathbb{N}$, which plays an important technical role. It shows that the remote past before time $t-\Delta$ has no effect on the conditional probability of spiking at time $t+1$. This ultimately turns $\{\vec{X}(t)\}_{t\in\mathbb{N}}$ into a $\Delta-$dependent process. Note that introducing Δ is not very restrictive. In fact, it does not seem physically possible that spikes in the remote past (of any neuron i) would have any effect on the present state of a given neuron j. Consider the following model:

$$p_j(t+1|\mathcal{H}_t) = F\Big(b_j + \sum_{i=1}^{J} w_{ij} g(X_i(\tau_j(t) \vee (t-\Delta) : t])\Big). \tag{3.5}$$

The descriptions of the parameters of model are same as the model shown in Equation (3.1) in Section 3.2.1. Let $D(t) := (D_{j,s}(t))_{\Delta\times J}$ be a $J\times\Delta$ matrix, where $D_{j,s}(t) := X_j(t-s+1),\ s=1,\cdots,\Delta,\ j=1,\cdots,J$. One can show by using Equation (3.1) that $\{D(t),\ t\geq\Delta\}$ is a homogeneous Markov chain on the state space $S := \{0,1\}^{J\times\Delta}$. Observe that $X_j(t) = h_j(D(t))$, where $h_j(D) = d_{j,1}, D = (d_{i,j})_{J\times\Delta}$, and where $\{D(t)\}_{t\geq 0}$ is the underlying Markov chain associated with the spike train model. We introduce the notation:

$$p_j(t+1|\mathcal{H}_t) \equiv p_j(D(t);\theta),\quad j=1,\cdots,J, \tag{3.6}$$

Suppose that $\theta_* = (W_*, \vec{b}_*)$ is the true unknown parameter.

Proposition III.2. *If $|g| \leq K$, then the Markov chain $D = \{D(t)\}_{t\geq 0}$ is irreducible and aperiodic.*

Proof. Since g is bounded and $0 < F(x) < 1$ for all $x \in \mathbb{R}$, we have that for some $\epsilon_0 > 0$, $p_j(t+1|\mathcal{H}_t) \in [\epsilon_0, 1-\epsilon_0] \subset (0,1)$, with probability one. Therefore, since the state space is finite, it is easy to see that $D \leftrightarrow \widetilde{D}$, for all $D, \widetilde{D} \in S$ and hence the Markov chain is irreducible. The fact that $p_j(t+1|\mathcal{H}_t) \in (0,1)$ also implies that the chain is aperiodic.

Note: even if one has $\Delta = \infty$ (i.e. infinite horizon) the chain should be irreducible and positive (recurrent). □

In our setting, the function g is trivially bounded. Therefore, the Markov chain D is ergodic and irreducible and by the strong law of large numbers for such Markov chains following Theorem 17.0.1 in Chapter 17 of Meyn and Tweedie (2005), we have:

$$\frac{1}{T}\sum_{t=1}^{T} f(D(t)) \xrightarrow{a.s.} \mathbb{E}f(D^*), \quad \text{as } T \to \infty, \tag{3.7}$$

for all $f \in L^1(\mathbb{J})$, where D^* has the (unique) stationary distribution π of the Markov chain.

Lemma III.3. *The model (3.1) is identifiable if and only if for each fixed $j = 1, \cdots, J$, the functions $\vec{1}$, $g_{i,j}(.)$, $1 \leq i \leq J$ are linearly independent, where $g_{ij}(D) := g(D_{i,(\tau_j:\Delta]})$ where $\tau_j = \max\{0 \leq t \leq \Delta : D_{j,t} = 1\}$ with $\max\{\varnothing\} = 0$, by convention.*

Proof. The identifiability of (3.1) is equivalent to the fact that

$$p_j(D;\theta) = p_j(D;\theta'), \text{ for all } j = 1, \cdots, J, \text{ and } D \in \{0,1\}^{J\times\Delta} \tag{3.8}$$

if and only if $\theta = \theta'$, where $\theta = ((b_j)_{j=1}^J, (w_{i,j})_{J\times J})$.

If $\theta = \theta'$ we trivially have (3.8). Now suppose that (3.8) holds for some $\theta \neq \theta'$, then there exists $j_0 \in \{1, \cdots, J\}$ such that $(b_{j_0}, w_{1j_0}, \cdots, w_{Jj_0}) \neq (b'_{j_0}, w'_{1j_0}, \cdots, w'_{Jj_0})$,

but

$$p_{j_0}(D;\theta) = p_{j_0}(D;\theta'), \text{ for all } D \in \{0,1\}^{\Delta\times J} \tag{3.9}$$

By (3.1) and the fact that F is strictly monotone (3.9) is equivalent to

$$b_{j_0} + \sum_{i=1}^{J} w_{ij_0} g_{ij_0}(D) = b'_{j_0} + \sum_{i=1}^{J} w'_{ij_0} g_{ij_0}(D), \text{ for all } D \in \{0,1\}^{J\times\Delta}.$$

Since $(b_{j_0}, w_{1j_0}, \cdots, w_{Jj_0}) \neq (b'_{j_0}, w'_{1j_0}, \cdots, w'_{Jj_0})$, the last equality can hold for all $D \in \{0,1\}^{J\times\Delta}$ if and only if the functions $\vec{1}$, $g_{1j_0}(.)$,..., $g_{Jj_0}(.)$ are linearly dependent (as functions from $\{0,1\}^{J\times\Delta}$ to $\mathbb{R}$). This is a contradiction showing that $\theta = \theta'$. $\square$

In addition to the lemma we have shown in Section 3.3.1 that the model (3.1) is not fully identifiable in certain special cases particularly if we choose $g(\cdot)$ to be an indicator function as defined below:

$$g(X_i(\tau_j(t):t]) = \mathbb{I}\{\sum_{k=\tau_j(t)+1}^{t} X_i(k) > 0\}$$

In this setting it is obvious that the baseline parameters b_i are confounded with the the connectivity strength w_{ii} and we will not be able to estimate both separately.

Following Chapter 12 of Cappé et al. (2005) we have the following standard criterion for the asymptotic normality of the maximum likelihood estimate. The standard asymptotic properties of the maximum likelihood estimate follows from three basic results: a law of large numbers for the log-likelihood, a central limit theorem for the score function u_{ij}, defined in Section 3.3, and a law of large of numbers for the observed information, $\hat{I}_{ii'}$.

Theorem III.4. *Suppose that:*

(i) For all $\theta \in \Theta$, $T^{-1}\ell_T(\theta) \to \ell(\theta)$ $\mathbb{P}_{\theta_*}$ *- a.s. uniformly over compact subsets*

of Θ, where $\ell_T(\theta)$ is the log-likelihood of the parameter θ given the first T observations and $\ell(\theta)$ is a continuous deterministic function with a unique global maximum at θ_*.

(ii)

$$\frac{1}{\sqrt{T}}\nabla\ell_T(\theta_*) \to \mathcal{N}(0, \mathfrak{I}(\theta_*)) \quad \mathbb{P}_{\theta_*} \; - \; \textit{weakly}, \tag{3.10}$$

where $\mathfrak{I}(\theta_*)$ is the Fisher information matrix at θ_*.

(iii) For all $1 \leq i, j \leq d$,

$$\lim_{\delta\to 0}\lim_{T\to\infty}\sup_{|\theta-\theta_*|\leq\delta}\Big| - T^{-1}\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_T(\theta) - \mathfrak{I}_{i,j}(\theta_*)\Big| = 0 \quad \mathbb{P}_{\theta_*} \; - \; \textit{a.s.} \tag{3.11}$$

If in addition to (i), (ii) and (iii) if θ_* is an interior point of θ and $\mathfrak{I}(\theta_*)$ is non-singular, then

$$\sqrt{T}(\hat{\theta}_{MLE} - \theta_*) \xrightarrow{d} \mathcal{N}\Big(\vec{0}, \mathfrak{I}^{-1}(\theta_*)\Big). \tag{3.12}$$

Remark: *One may be tempted to use Theorem 12.5.7 of Cappé et al. (2005) to obtain the the asymptotic normality of the maximum likelihood estimate. Our case, however, involves a discrete state space where the densities (w.r.t. the counting measure) of the one-step transition probabilities are not bounded away from zero as required in assumption (12.2.1) of Cappé et al. (2005). Therefore, we provide an independent proof of the desired asymptotic normality of the maximum likelihood estimate.*

Proof of Theorem III.4: See page 442-443 of Cappé et al. (2005). The proof of the conditions of Theorem III.4 are given in propositions F.1, F.2 and F.4 reported in the Appendix F. □

3.4 Results

3.4.1 Simulated Data Analysis: Network I

In this section, we illustrate the analysis on a simulated neuronal network by employing our implementation of model (3.1). We show a simple example network which includes five neurons with the configurations shown in Figure 3.4. 10,000 data points or in other words 10 seconds worth of data were generated for each neuron. The baseline firing rates were set to suitable values in order to ensure that the simulated firing rates are comparable to those observed for in vitro multi-electrode recordings. The baseline coefficient b of each neuron is taken -3 so our baseline firing rates become approximately 5Hz. We embedded both excitatory and inhibitory connections in the network with the connectivity strengths ranging between $(-2, 2)$ in the network. The red arrows in Figure 3.4 shows the inhibitory connections and the black arrows shows the excitatory connections. We apply our estimation method to analyze the simulated data. The true and estimated value of the connectivity matrix and of the network parameters are shown in Figure 3.5. Figures 3.7 show the correlation matrix calculated theoretically from Lemma G.1 and from simulated data. As we can see it is block diagonal in nature. It means that the connectivity strengths of the neurons influenced by the same neuron could be correlated.

We estimate the baseline coefficients, b as well as the connectivity strength matrix, w from the simulated network data. We ran 100 replications of the simulated network. Figure 3.6 shows the results of these 100 replicates. In this figure we show 25 boxplots of the estimated connectivity strength matrix W of simulated network with red lines corresponding to actual values of W. We observe that the width of the boxplots for inhibitory connections are much wider than the width of the boxplots with excitatory connections indicating that estimation of the inhibitory connections are much harder. We empirically calculated the coverage probabilities of the corre-

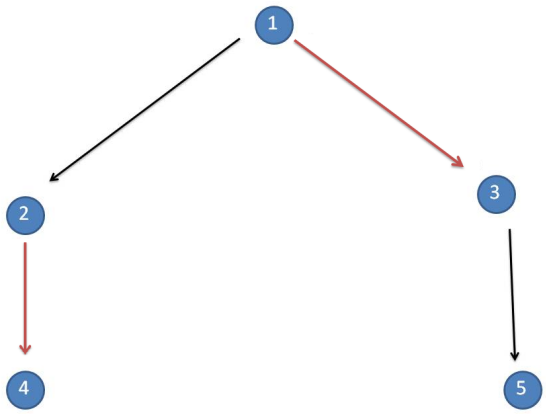

Figure 3.4: Simulated Network I: A network with 5 neurons with an intrinsic firing rate of 5Hz for each

sponding 95% confidence intervals based on these 100 replications. We calculated the 95% confidence intervals for the baseline coefficients for each neuron. Figure 3.8 reports the 95% confidence intervals of three values of our parameters corresponding to $b = -3$ (baseline firing coefficient) and connectivity strength of $w = 0$ and $w = 2$ respectively. We observe that the standard errors of $w = 2$ are much smaller than those of $w = 0$ and $w = -3$. This is because the strength of the signal is much stronger for an excitatory connection and it is more difficult to estimate an inhibitory connection as well as detecting a false edge. The estimated baseline firing rates for the Discounted Model are reported in Table 3.1. We report the empirical coverage probability of each parameter of the W matrix in Table 3.2. These results are generated using 100 replicates of data each with $L = 10^4$ or 10 seconds sample points from the network shown in Figure 3.4.

We apply the profile likelihood methods to estimate the discounting rate λ. We initialize λ at a value within the range of (0,1). We iteratively estimate λ and the connectivity strength matrix, w and choose that value of λ where the profile likelihood is maximized. Figure 3.9 shows the profile likelihood for the Discounted model with different λ. We simulated data from model (3.1) with discounting rate $\lambda = 0.5$. We

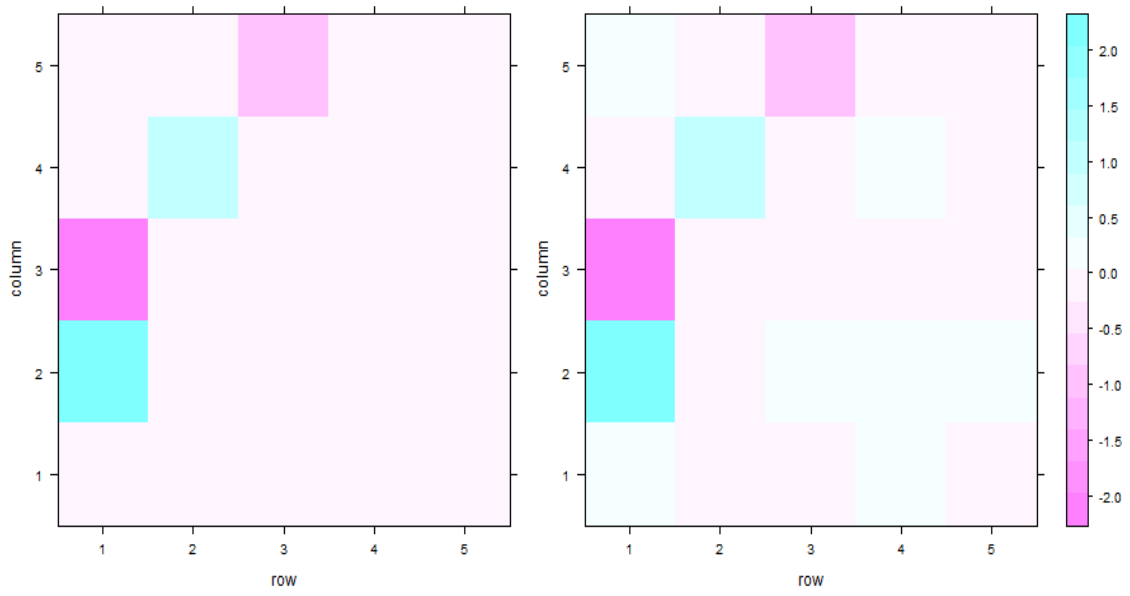

Figure 3.5: Left: The actual connectivity matrix of network B Right: Estimated connectivity matrix from 10s worth of data

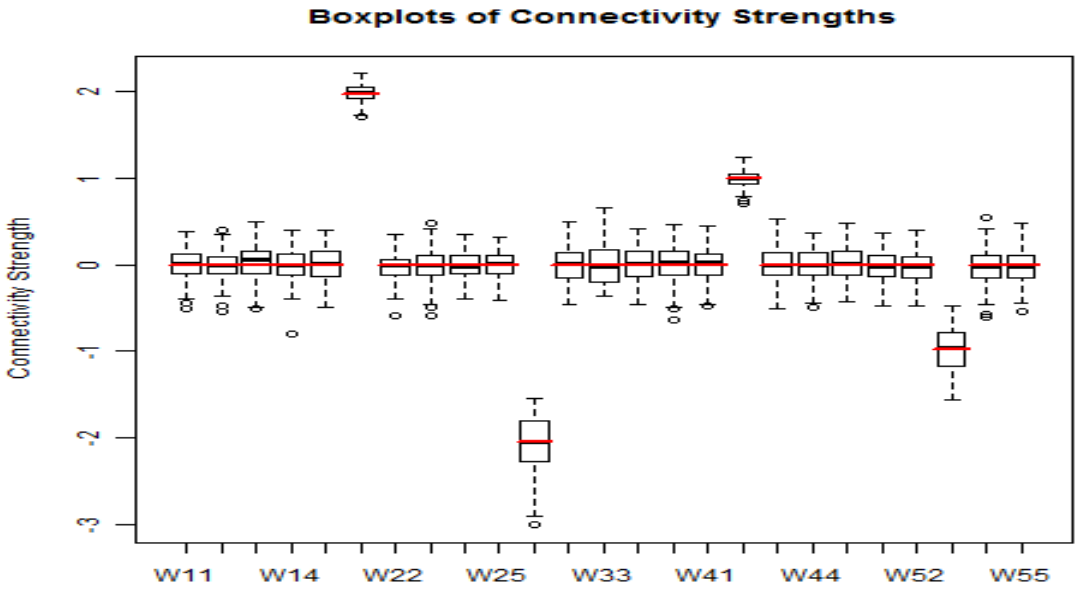

Figure 3.6: Boxplots of the estimated connectivity strengths W of simulated network with red lines corresponding to actual values of W

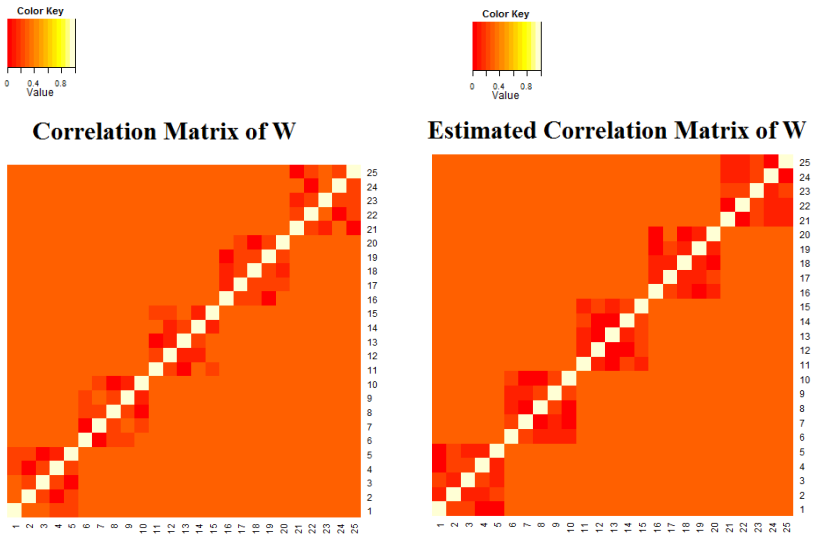

Figure 3.7: Correlation Matrix from theoretical calculation (left) and Simulated Data (right) for the Discounted Model

estimate the profile likelihood at each value of λ chosen and we get $\hat{\lambda}$=0.5 where the likelihood was maximum.

Table 3.1: Estimated baseline firing rates with their standard error for the Discounted Model

Neuron	1	2	3	4	5
Baseline Parameter	-3.00	-3.00	-3.00	-3.00	-3.00
Estimated Baseline Parameter	-3.08	-2.94	-3.03	-3.04	-2.97
Standard Error	0.07	0.06	0.06	0.06	0.06

Table 3.2: Coverage probability of each parameter of the W matrix for 100 replicates for the Discounted Model with $L = 10^4$

(i,j)	1	2	3	4	5
1	0.93	0.97	0.96	0.94	0.93
2	0.92	0.93	0.97	0.96	0.94
3	0.96	0.96	0.95	0.90	0.95
4	0.93	0.93	0.95	0.95	0.94
5	0.91	0.95	0.93	0.95	0.97

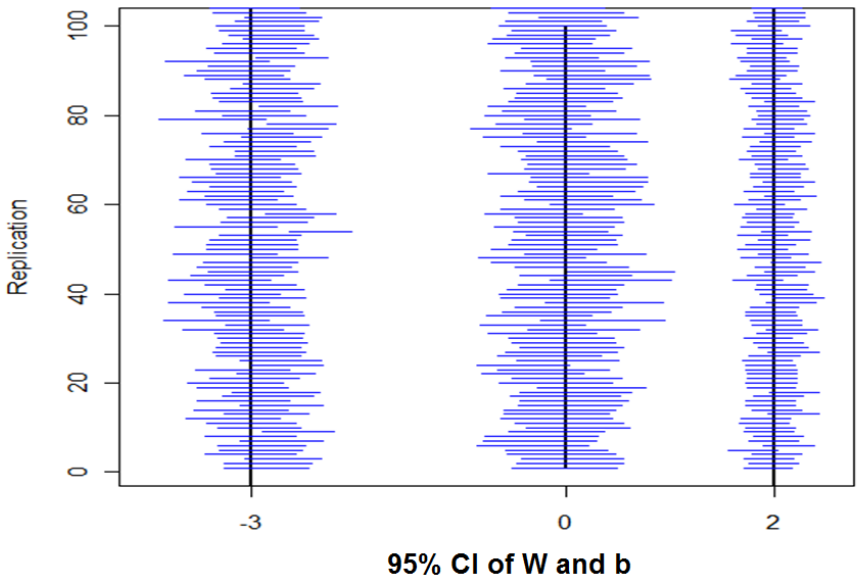

Figure 3.8: Confidence Intervals of W = −3, 0, 2 respectively for 100 replications with L=10^4

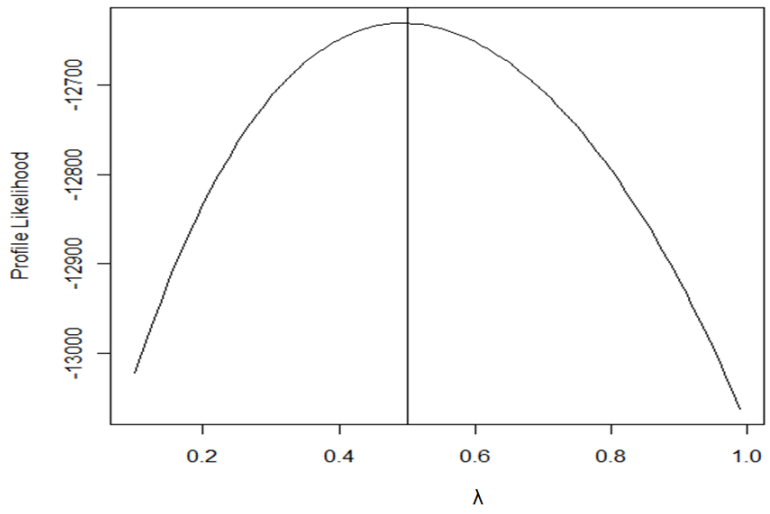

Figure 3.9: Profile likelihood of λ for the Discounted model with the maximum at $\hat{\lambda}$=0.5

3.4.2 Simulated Data Analysis: Example Network II

In this section, we illustrate that our method works for more complicated networks successfully. We simulated a network having different embedded connectivity strengths and functional relationships. We illustrate an example of a 25 neuron network with number of possible configurations shown in Figure 3.10. Moreover, twenty thousand data points or 20 seconds worth of data was generated for each neuron by implementing Equation (3.1). The baseline firing rates, similarly, were set to suitable values in order to ensure that the simulated firing rates are comparable to those observed for in vitro multi-electrode recordings. The baseline coefficient b of each neuron is taken -2.5 so our baseline firing rates become 7.5Hz. We embedded both excitatory and inhibitory connections in the network. The connectivity strengths between the neurons, w, ranges from $(-2, 2)$ in the network. We estimate the baseline coefficients, b as well as the connectivity strength matrix, w from the simulated network data. We observe that both the excitatory and inhibitory connections are estimated accurately. We also calculated 95% confidence intervals for the baseline coefficients for each neuron. The estimate and standard error of the baseline coefficient of a neuron, arbitrarily chosen from the network, are reported as -2.49 and 0.07 respectively.

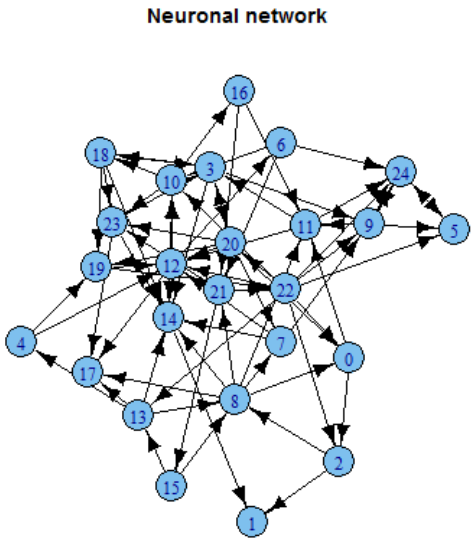

Figure 3.10: Simulated Network II: A network with 25 neurons and embedded connections. Each neuron had an intrinsic firing rate of 7Hz.

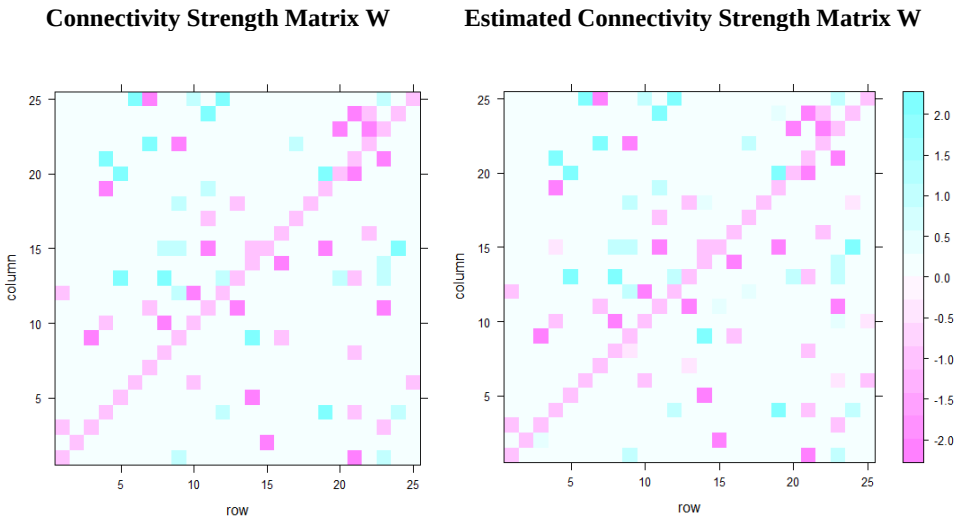

Figure 3.11: Left: The actual connectivity matrix of network B Right: Estimated connectivity matrix from 20s worth of data

3.4.3 Application to Cultured cortical Neurons

Wagenaar et al. (2006) made available to the public an extensive set of multielectrode array (MEA) recordings from cultured cortical neurons. Half-hour recording sessions were performed across 58 different cultures during their first five weeks of development. Wagenaar et al. (2006) focused on characterizing "population bursts" in the spiking activity of the cultures. Such bursts, defined as brief periods of time during which the firing rate of several cells or electrodes greatly exceeds the baseline rate, are a common feature observed in cultures of many different types of neurons. Here, our focus is not to characterize the bursts but rather to detect precisely timed spiking patterns involving multiple neurons. From these patterns we can estimate the strength of functional connectivity between different neurons in the culture.

3.4.3.1 Data Pre-Processing

We began our analysis with culture 2-1-34, meaning the first culture from the second batch after 34 days *in vitro* (DIV). This culture was "densely" plated with approximately 50,000 cells, and on DIV 34 was characterized by Wagenaar et al. (2006) as having bursts of fixed size, with a frequency of between 2 and 10 bursts per minute. The input data for our analysis were the timestamps of the spikes recorded on 56 different electrodes (we did not work with the raw voltage waveforms themselves).

For our analysis we discretized the time axis with a bin size of Δ=1 ms. To identify bursts we used the array-wide spike detection rate (ASDR) measure, defined in Wagenaar et al. (2006) as the number of spikes per unit time summed over all the electrodes in the array. Averaged over the entire 30 minute recording, the ASDR for a 100 ms window in culture 2-1-34 was about 20. We considered the culture to be bursting anytime the ASDR in a 100 ms window exceeded 50, corresponding to a 2.5-fold increase over the average. For example, we will consider the spikes in the first 120 seconds of recording from culture 2-1-34. In the raster plot shown in Figure

3.12, bursts are visible at around 10 and 100 seconds. In Figure 3.13, we see that the ASDR greatly exceeds the threshold of 50 during these bursts.

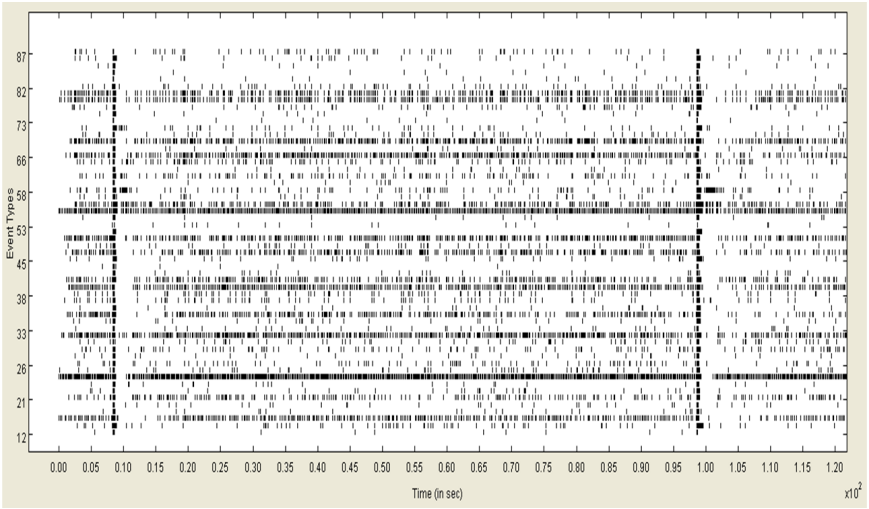

Figure 3.12: Spike raster of first 120 seconds from culture 2-1-34.

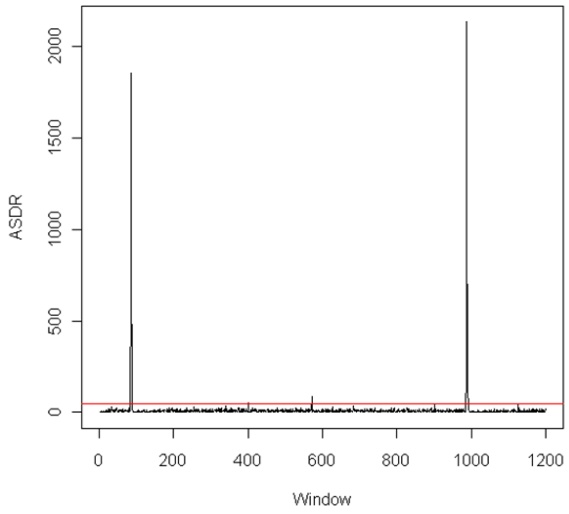

Figure 3.13: ASDR (window size 100 ms) of first 120 seconds from culture 2-1-34. Burst threshold of 50 shown in red.

To obtain sections of the data suitable for our analysis, we looked for 20 second stretches which did not contain any windows that exceeded the threshold. Additionally, we required these 20 second segments to begin and end at least 2 seconds away from any window that exceeded the threshold (this was to avoid including any windows that happened to catch the very beginning of a burst, as well as the brief

periods of extremely low spiking activity that appear to occur immediately following a burst). The first 120 seconds contained one such segment (60 to 80 seconds). In the 30 minutes of data from culture 2-1-34, we found 63 suitable segments. We analyzed each of these segments individually.

3.4.3.2 Detecting Functional Connectivity over Time

In each segment, we fit our discounted model and estimate the connectivity strengths as well as the baseline firing rates of the neurons in the culture. We use 16 neurons from the culture (out of 56). These 16 neurons were chosen based on their consistent firing rates in all the 63 time segments. We also calculated 95% confidence intervals for each strength $\hat{w}$ in each segment. If the confidence limits of $\hat{w}$ do not cover 0, then we considered that strength to be significant in that segment. We repeated this for each strength of pair of neurons and each segment. Based on the set of significant connections, we plotted the estimates of the connectivity strength with their corresponding 95% confidence interval in Figure 3.15.

We study the networks for 6 equally spaced time points in Figure 3.16 to visualize the functional connectivity present in culture 2-1-34. This indicates that the connectivity in the culture is still evolving. Overall, our results are consistent with the patterns reported in a Diekman et al. (2012) and also in Rolston et al. (2007). Specifically, there are thirteen connections mentioned in Rolston et al. (2007), six of them were also found to be significant in our analysis. We show the connections which were significant in at least 60 out of 63 (95%) of the time segments in Figure 3.14. We detected 16 connections which were consistently significant in most (95%) of the 63 time segments and most of these connections are reported in Diekman et al. (2012) and Rolston et al. (2007). But the connections shown in red arrow in Figure 3.14 were not reported in any of the previous studies of the data. We know that the data have a lot of bursts and thus the new connections found in our method are

mostly self exciting properties that are intrinsic to the data.

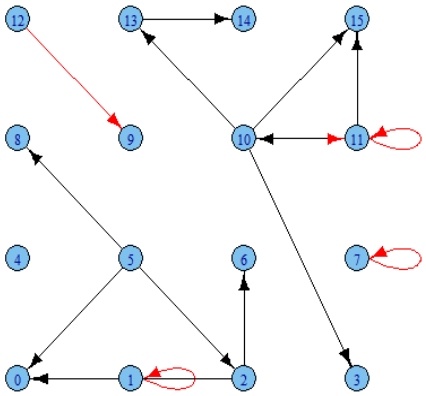

Figure 3.14: Estimated network from our discounted model with black edges previously reported in literature and the red edges are new findings

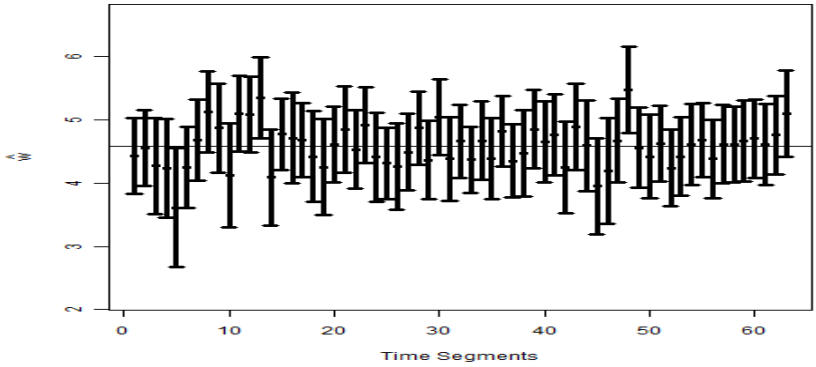

Figure 3.15: Estimates of connectivity strength parameter (which was 95% times significant among the 63 segments) with confidence intervals superimposed with the mean connectivity strength.

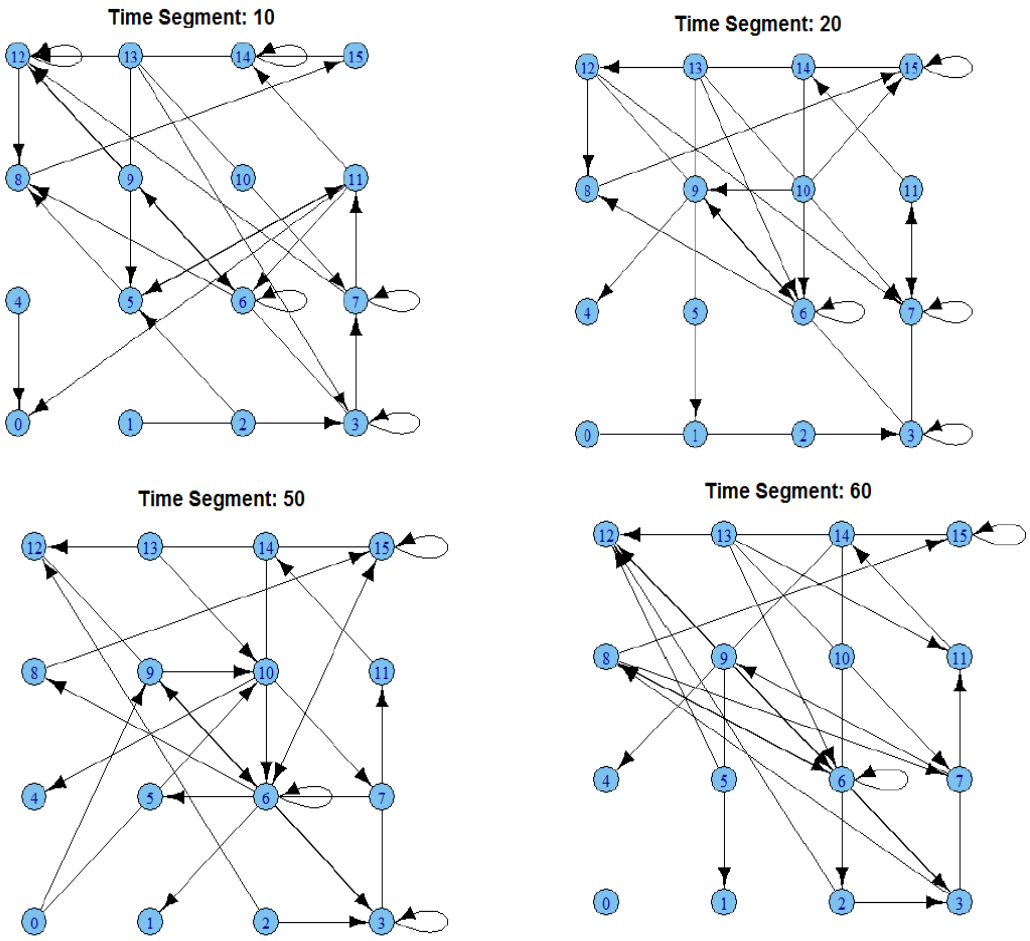

Figure 3.16: The evolving functional connectivity in culture 2-1-34.

3.5 Discussion

In this chapter we introduced a new class of temporal models for the analysis of multiple spike trains recordings. We adopted a Markovian dependence structure over time resembling some of the main features of the spiking process following the lines of Kass and Ventura (2001) and Rigat et al. (2006). A distinctive feature of our approach is that the network structure is modeled explicitly as one of the unknown parameters for the network effects. The network connectivity is explained by a regression term including the available fixed-time covariates.

We model the relationships between the network coefficients, the past history of the spiking process and the firing probabilities of the neurons. The firings recorded

within each inter-spike time of a given neuron are weighted using their arrival times which means that the spikes in recent past has more weightage than the spikes in remote past. Also, multiple spikes of a neuron has more weightage than a single spike of a neuron.

Although the exponent of the conditional probability function (3.1) is linear in the model parameters, the logit link implies a symmetric saturation of the spiking probability with respect to the input process. In fact, since the logistic density is a symmetric bell-shaped curve, the fluctuations of the network activity produce small changes of a neuron's firing probability when the the conditional probability function (3.1) is far from zero and larger changes when it is close to zero.

We also derived the asymptotic theory for the estimators under some assumptions. Our inference methods are shown to be considerably less complex than other comparable methods. Currently, our model has no concept of delay between the neurons and a interesting problem would be to extend the theory for the case of more general model with delays which may be able to represent additional types of connectivity graphs. We will be addressing such extensions in our next chapter and also in future work.

CHAPTER IV

Estimating Functional Connectivity and Delays in Neuronal Networks: A Graphical Model Approach

4.1 Introduction

Neuroscience can be broadly described as the study of the physiology, structure and function of the nervous system, consisting of the brain and nerve pathways running throughout the body. The nervous system is responsible for actuating response to external stimulus. The macro effects of the nervous system depend on individual cells called neurons. Their anatomical structure has long tails (axons) for sending signals and tree like proliferations (dendrites) for receiving signals. Using the neuronal signals, we can study how the signals travel down the axon, how they are integrated in dendrites, etc. The interaction of neurons in a network followed by interaction of networks is another problem of potential interest.

Existing literature in neuroscience has focused on studying the electrophysiological and anatomical features of the cells constituting the nervous system. There are also detailed studies on the micro-level understanding of each cell in a neuronal network. However there is little literature on how these cells interact to form a neuronal circuit and are thus capable of producing different behaviors. Studying the neuronal network is primarily important in order to discover the interactive principles governing the organization of neuronal systems. Using graphical models theory, we intend to provide a quantitative approach for the detection of associations between the neurons and thus

deciphering the neuronal network.

Our problem can be expressed in the graphical theory framework exploiting the correspondence between a directed graphical model and neuronal network. A directed graphical model is used in a variety of domains including statistical physics, natural language processing, image analysis and spatial statistics (Lauritzen (1996), Whittaker (2009), Cressie (1993), Jordan (2004)). Each graph represents a class of graphical models and thus learning a graph is a model selection problem.

We define an association graph and a directed graphical model in the following manner. An association graph has a set of vertices, each vertex representing a variable. An edge connecting two variables represents a conditional association between them. A directed graph is denoted as $G = (V, E)$, with vertex set $V = \{1, 2, \cdots, p\}$ and edge set $E \subset V \times V$. The structure of this graph encodes certain conditional independence assumptions among subsets of the p-dimensional discrete random variable $X = (X_1, X_2, \cdots, X_p)$, where variable X_i is associated with vertex $i \in V$. A basic feature of the notion of a graph is that it is a visual object. It is conveniently represented by a picture with circles used to describe a node and edges to describe connectivity between the nodes.

In our scenario we treat the neurons as vertices and there will be an edge between them if there is functional connectivity between them. Our problem for such a model is to estimate the underlying graph from n samples $\{x_1, x_2, \cdots, x_n\}$ drawn from the distribution specified by the likelihood of our model. The discrete random variables X_i will denote the spike train of the ith neuron. Due to both its importance and difficulty, the problem of structure learning for discrete graphical models has attracted considerable attention. The absence of an edge in a graphical model encodes a conditional independence assumption.

The chapter is organized as follows. In Section 4.2.1, we focus on a simple problem of estimating the weights of the edges or in other words connectivity strength matrix

given the knowledge about the graphical model along with the delay. In Section 4.2.3 we describe our algorithm to find the delay matrix with a grid search method along with estimating the connectivity strength matrix for a known structure of the graphical model. In Section 4.2.4, we focus on a discrete time graphical model with unknown structure and the description of the parameters of the model. We illustrate the method of estimation in Section 4.2.5. We simulate few examples of neuronal networks in Section 4.3.1 and estimate the underlying graphical structure. We also illustrate an application of multi-electrode array data from cortical cultures in Section 4.3.3.

4.2 Methods

We consider the problem of estimating the graph associated with a discrete stochastic model in which the edges of any given node, or neuron in our case, is estimated by performing logistic regression on each neuron. We start with a simple problem of estimating the weights of the edges or in other words connectivity strength matrix given the knowledge about the graphical model along with the delay. Then we describe our algorithm to find the delay matrix with a grid search method along with estimating the connectivity strength matrix for a known structure of the graphical model. Finally we focus on a discrete time graphical model with unknown structure and the description of the parameters of the model and illustrate the method of estimation of the edges and the delay matrix iteratively.

4.2.1 Likelihood and Estimation with Known Structure

The goal of this section is to estimate the delay given known structures, while we do not have any parameterizations for the firing conditional probability model. The conditional probabilities are simply taken to follow Bernoulli distribution. We derive the likelihood of the network and estimate the parameters assuming that we have

information about the edges of the full network. We define X_t as a binary random variable which can take values 0 or 1 according to if the neuron X has fired at time t or not with probability of firing P_X for all t. We assume a simple Bernoulli model for each neuron in the network and assume that if a neuron A is influenced by another neuron it could increase the firing probability of neuron A after a delay of d time units.

To describe the dependence, consider two neurons A and B with a one-directional structure: A influences B. We say that the episode $E = A[d]B$ occurs if a firing of A is followed by a firing of B after d time units. If A and B are independent, the probability of observing $A[d]B$ is just $P_A \times P_B$. However, if the firing of A leads to an excitation of neuron B and it is more likely to fire after d time units, then $P(A[d]B)$ will be higher. Similarly, if the effect is inhibitory, it will be lower. Neurons have a particular synaptic delay depending on the neuro-transmitter they have, so one neuron typically affects another neuron which is connected to it after some delay. For simplicity, we will assume that the time delays between the neurons are fixed at a given value d. This is what is typically done in existing literature (Sastry and Unnikrishnan (2010)). In this section, we assume that we know the delay matrix D. In practice, one will analyze the data by iteratively estimating the delay matrix and the connectivity strength matrix and then combine the evidence for dependence across the values of d for a given pair of neurons. These results are shown in later sections of this chapter.

The likelihood for a generic network will be of the following form:

$$\mathbf{L}(data|\theta) = \sum_{(i,j)=1}^{n} \sum_{v=1}^{V} c_{v_{ij}} (\theta_{v_{ij}} log(\theta_{v_{ij}}) + (1 - \theta_{v_{ij}}) log(1 - \theta_{v_{ij}})) \tag{4.1}$$

Here, $(i, j) = 0$ or 1 and V is the set of nodes i.e. $\{A, B, C\}$. $c_{v_{ij}}$'s are functions of the binary variables x_{v_t} which can take values 1 or 0 according to neuron v has fired

in time bin t or not. The $\theta_{v_{ij}}$'s and $c_{v_{ij}}$'s are function of d here. In Section 4.2.3 we show that we estimate the time delay d by maximizing the given likelihood.

4.2.2 An Illustrative Example

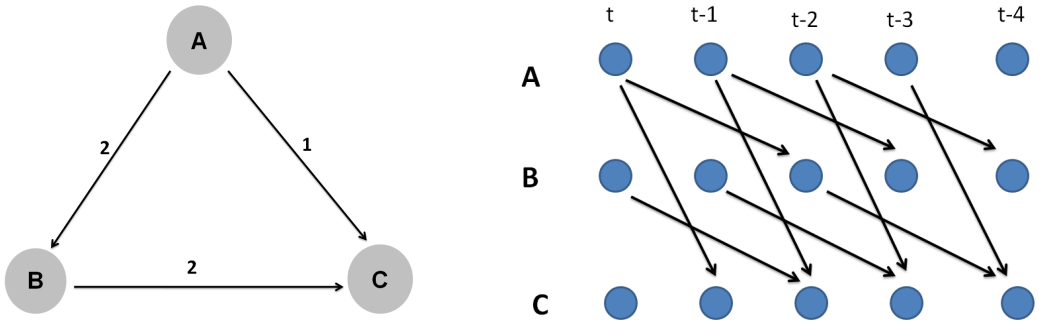

Figure 4.1: A network with 3 neurons and embedded connections. Each neuron had an intrinsic firing rate of 5Hz. The connections had varying strengths and time delays from 1 to 2 ms. The delays are indicated by the numbers next to the arrows.

We show a simple example network of three neurons to explain our model. Figure 4.1 shows the network of three neurons, A, B and C with delays of 2ms, 2ms and 1ms for edges $A \to B$, $B \to C$ and $A \to C$ respectively. The figure on the right panel shows the network from the temporal point of view. For this example network we denote x_A, x_B and x_C as realizations of A_t, B_t and C_t respectively which could take values 0 or 1. We define $\theta_A = P(A_t = 1)$ and θ_B and θ_C are defined similarly. Moreover, we define $\theta_{B_i} = P(B_t = 1 | X_{A_{t-2}} = i)$ and $\theta_{C_{ij}} = P(C_t = 1 | X_{A_{t-1}} = i, X_{B_{t-2}} = j)$ where (i, j) could be 0 or 1. We can compute the joint likelihood of the example network at any interval as follows:

$$L(\theta) \quad = \quad \prod_{t=1}^{n} P(C_t | B_{t-2} A_{t-1}) P(B_t | A_{t-2}) P(A_t)$$

In general, we can write

$$L(\theta) = \prod_{i=1}^{n} f_{\theta_C}(x_{C_i}, x_{B_{i-d_1}}, x_{A_{i-d_2}}) g_{\theta_B}(x_B, x_{A_{i-d_3}}) h_{\theta_A}(x_{A_i})$$

where

Thus, the conditional likelihood can be written as:

$$\begin{aligned} f_{\theta_C}(x_{C_i}, x_{B_{i-d_1}}, x_{A_{i-d_2}}) &= [{\theta_{C_{11}}}^{x_{C_i}}(1-\theta_{C_{11}})^{1-x_{C_i}}]^{x_{A_{i-d_2}} x_{B_{i-d_1}}} \\ &\times [{\theta_{C_{10}}}^{x_{C_i}}(1-\theta_{C_{10}})^{1-x_{C_i}}]^{x_{A_{i-d_2}}(1-x_{B_{i-d_1}})} \\ &\times [{\theta_{C_{01}}}^{x_{C_i}}(1-\theta_{C_{01}})^{1-x_{C_i}}]^{(1-x_{A_{i-d_2}})x_{B_{i-d_1}}} \\ &\times [{\theta_{C_{00}}}^{x_{C_i}}(1-\theta_{C_{00}})^{1-x_{C_i}}]^{(1-x_{A_{i-d_2}})(1-x_{B_{i-d_1}})} \end{aligned}$$

Similarly we get the conditional likelihood of neuron B and A as follows:

$$g_{\theta_B}(x_B, x_{A_{i-d_3}}) = \prod_{i=1}^{n} [{\theta_{B_1}}^{x_{B_i}}(1-\theta_{B_1})^{(1-x_{B_i})}]^{x_{A_{i-d_3}}} [{\theta_{B_0}}^{x_{B_i}}(1-\theta_{B_0})^{1-x_{B_i}}]^{(1-x_{A_{i-d_3}})}$$

and

$$h_{\theta_A}(x_{A_i}) = \prod_{i=1}^{n} [{\theta_A}^{x_{A_i}}(1-\theta_A)^{(1-x_{A_i})}]$$

Thus we get the joint log-likelihood of the network as:

$$\mathbf{l}(\theta) = \sum_{i=1}^{n} [log(f_{\theta_C}(x_{A_{i-d_2}}, x_{B_{i-d_1}}, x_{C_i})) + log(g_{\theta_B}(x_{A_{i-d_3}}, x_{B_i})) + log(h_{\theta_A}(x_{A_i}))]$$

We get the MLE of the parameters by differentiating:

$$\begin{aligned} \frac{\delta \mathbf{l}(\theta)}{\delta \theta} &= \frac{\sum_{i=1}^{n} x_{A_{i-d_2}} x_{B_{i-d_1}} x_{C_i}}{\theta_{C_{11}}} + \frac{\sum_{i=1}^{n} x_{A_{i-d_2}} x_{B_{i-d_1}} (1-x_{C_i})}{1-\theta_{C_{11}}} \\ &\Rightarrow \hat{\theta}_{C_{11}} = \frac{\sum_{i=1}^{n} x_{A_{i-d_2}} x_{B_{i-d_1}} x_{C_i}}{\sum_{i=1}^{n} x_{A_{i-d_2}} x_{B_{i-d_1}}} \end{aligned}$$

Similarly by differentiating the log likelihood we can estimate all the parameters of the network.

$$
\begin{aligned}
\hat{\theta}_{C_{11}} &= \frac{\sum_{i=1}^{n} x_{A_{i-d_2}} x_{B_{i-d_1}} x_{C_i}}{\sum_{i=1}^{n} x_{A_{i-d_2}} x_{B_{i-d_1}}} \\
\hat{\theta}_{C_{10}} &= \frac{\sum_{i=1}^{n} x_{A_{i-d_2}} (1 - x_{B_{i-d_1}}) x_{C_i}}{\sum_{i=1}^{n} x_{A_{i-d_2}} (1 - x_{B_{i-d_1}})} \\
\hat{\theta}_{C_{01}} &= \frac{\sum_{i=1}^{n} (1 - x_{A_{i-d_2}}) x_{B_{i-d_1}} x_{C_i}}{\sum_{i=1}^{n} (1 - x_{A_{i-d_2}}) x_{B_{i-d_1}}} \\
\hat{\theta}_{C_{00}} &= \frac{\sum_{i=1}^{n} (1 - x_{A_{i-d_2}})(1 - x_{B_{i-d_1}}) x_{C_i}}{\sum_{i=1}^{n} (1 - x_{A_{i-d_2}})(1 - x_{B_{i-d_1}})} \\
\hat{\theta}_{B_1} &= \frac{\sum_{i=1}^{n} x_{A_{i-d_3}} x_{B_i}}{\sum_{i=1}^{n} x_{A_{i-d_3}}} \\
\hat{\theta}_{B_0} &= \frac{\sum_{i=1}^{n} (1 - x_{A_{i-d_3}}) x_{B_i}}{\sum_{i=1}^{n} (1 - x_{A_{i-d_3}})} \\
\hat{\theta}_{A} &= \frac{\sum_{i=1}^{n} x_{A_i}}{n}
\end{aligned}
$$

Thus we see that the maximum likelihood estimators for the Bernoulli model (4.1) can be obtained in closed form solutions. However, the delays are not an explicit function of the likelihood so we cannot differentiate to get the maximum likelihood estimators of the delays even if we assume them to be continuous. In this case if the delays are not known we can iteratively estimate the delays and the parameter estimates by maximizing the profile likelihood of the model. We show the details of this procedure in the next section.

4.2.3 Algorithm to Estimate the Delays

In this section, we assume the network structure is known but we do not know the delay between the firings so we estimate the delays jointly by maximizing the likelihood. We have to estimate θ_d's and d's iteratively and we choose d using a local grid search method. The profile likelihood for a network will be of the following form:

$$\mathbf{L}(data|\hat{\theta}_d) = \sum_{(i,j)=1}^{n} \sum_{v=1}^{V} c_{v_{ij}}(\hat{\theta}_{v_{ij}} log(\hat{\theta}_{v_{ij}}) + (1 - \hat{\theta}_{v_{ij}}) log(1 - \hat{\theta}_{v_{ij}}))$$

Thus we get,

$$\begin{aligned} \hat{d} &= \text{argmax}_d \, \mathbf{L}(data|\hat{\theta}_d) \\ &= \sum_{(i,j)=1}^{n} \sum_{v=1}^{V} c_{v_{ij}}(\hat{\theta}_{v_{ij}} log(\hat{\theta}_{v_{ij}}) + (1 - \hat{\theta}_{v_{ij}}) log(1 - \hat{\theta}_{v_{ij}})) \end{aligned} \tag{4.2}$$

Here, $(i, j) = 0$ or 1 and V is the set of nodes i.e. $\{A, B, C\}$. $c_{v_{ij}}$'s are functions of the binary variables x_{v_t} which can take values 1 or 0 according to if neuron v has fired in time bin t or not. The $\hat{\theta}_{v_{ij}}$s and $c_{v_{ij}}$s are functions of d here. We choose the d which maximizes the given likelihood. The profile likelihood here has the form of weighted sum of negative entropy. We can compute d from the following equation:

$$\hat{d} = \text{argmax}_d \, L(data|\hat{\theta}_d)$$

However, the problem of estimating the delays becomes harder and computationally intensive as the number of graphical structures grow. The time delays of the neurons can vary from 1 ms to 200 ms depending on the neurotransmitter. We can have at most J^2 edges with J neurons in the network and that gives us $J^2 \times 200$ possibilities. Thus we implement the following hill-climbing "search" algorithm in a discrete space of all possible delay combinations. The algorithm finds the local

maxima in the large pool of possible delays.

The algorithm estimates the delay matrix D of a network with known structure without evaluating the likelihood at all possible combinations of d. We first count the number of occurrences of the episode $E = A[d]B$ for all pairs of neurons A and B with different values of d. We construct set S with all possible delays arranged in order of their corresponding 2-node counts. We initialize our algorithm to that value of delay matrix D which has the highest counts, say $d^{(0)}$. At the next step of our algorithm we choose the delays which has the highest likelihood among the neighboring points. The delay matrix, D, of a network with J neurons will be of order $J \times J$ so we select an arbitrary (i, j) th value at a time and compare the likelihood of the neighboring values. The neighborhood of delay $d^{(0)}$ is defined as points $d^{(0)} + 1$ and $d^{(0)} - 1$. Our next updated value of the delay d, say $d^{(1)}$ will be the delay with the highest likelihood among the neighborhood points. If we get stuck in a local maximum then we jump out of the neighborhood to the second highest count from the set S and similarly our algorithm repeats the same steps until our results converge. The algorithm is described next.

Algorithm 1: Algorithm to estimate delay D

Construct set S with all possible delays arranged in order of their corresponding 2-node counts
Initialize delay d at $d_{initial}$, say which has the highest counts in set S and value = 1
while value $\neq 0$ **do**
  Randomly pick a direction i
  Calculate the point d_i having maximum likelihood within the neighboring points of k in direction i
  if $d_i = d$ **then**
    Jump to $d = d_{i'}$
  else
    Stay at $d = d_i$
  end if
  if $d_i = d$ for all i **then**
    Draw a random sample d_r outside the neighborhood of d with probability proportional to the 2-node counts from set S
    Jump to $d = d_r$
  else
    Stay at $d = d_i$
  end if
  if Stopping Rule **then**
    Set value = 0
  end if
end while

We use the same simulated network in Figure 4.1 to illustrate the algorithm. We simulated data from the network with three neurons and the given embedded connections. Each neuron had an intrinsic firing rate of 5Hz. The connections had varying strengths and time delays from one to two milliseconds as shown in Figure 4.1. We apply Algorithm 1 to the simulated data and run it for 2000 iterations. Table 4.1 shows the log-likelihood values of the delays where the algorithm found local maximums. The maximum likelihood is the one having the correct delay $k = (1, 2, 2)$. We see that sample path of the algorithm in Figure 4.2. The algorithm visits the local maximums often and Table 4.1 shows the number of times each local maximum has been visited by the algorithm. Also, Figure 4.2 shows that the true delay (denoted by a red box) is visited multiple times along with other local maximums. We are able to select the true delay here since it corresponds to the point with maximum likelihood.

Table 4.1: Table showing log-likelihoods of the delays of the local maximum of the algorithm

	k		log-likelihood	# times repeated
4	2	2	-15905.60	46
1	2	4	-15942.40	39
1	5	2	-15876.19	52
4	5	2	-16167.76	34
1	2	2	-15614.02	57
4	4	5	-16487.53	9
1	4	5	16195.95	20
4	2	4	-16233.98	25

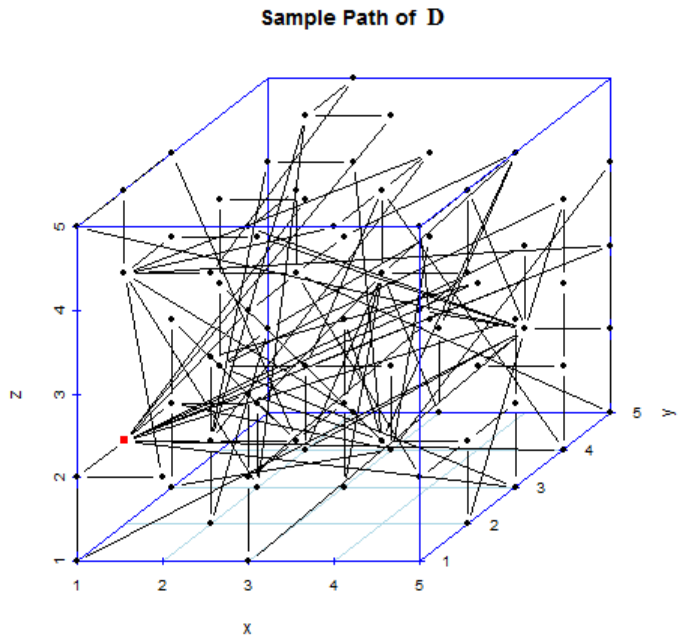

Figure 4.2: Sample Path of D for different iteration of the algorithm. The red square point is the true delay d = (1,2,2) used in our simulations

4.2.4 A Discrete Time Graphical Model

We have a similar setting as in Chapter III. We consider the spike from a neuron as instantaneous and we use discrete time. Let the intervals between the time points be small enough so that there is at most one spike in an interval. We observe the processes over a period of time, and suppose there are L intervals during this period. We consider there are J neurons in the network. Let $X_i(t)$, be the binary random variable of random spiking states, so that $X_i(t) = 1$ if neuron i is firing at time t and $X_i(t) = 0$ otherwise. Let $p_i(t+1|\mathcal{H}_t)$ denote the conditional probability for our model which denotes the probability of firing of the ith neuron at time $t+1$ given the history of firing of all the neurons till time t. The conditional probability can be expressed as a function of the parameter vectors b, denoting the the baseline firing coefficient and a parameter matrix W with dimensions $J \times J$, denoting the connectivity strength matrix. The baseline coefficient vector, b, will be of length J with b_i

denoting the baseline firing coefficient of neuron i. The connectivity strength matrix W and the delay matrix D will be of dimensions $J \times J$, with w_{ij} and d_{ij} denoting the connectivity strength and delay time between neurons i and j, respectively. We define the conditional probability for our model as:

$$\begin{aligned} p_i(t+1|\mathcal{H}_t) &= \mathbb{P}\{X_i(t+1)=1|\mathcal{H}_t\} \\ &= F\Big(\sum_{j=1}^{J} w_{ij} \sum_{k=k_u}^{k_l} \kappa_\sigma(k,(t-d_{ij}))X_j(k) + b_i\Big) \end{aligned} \tag{4.3}$$

where F is the c.d.f. of logistic distribution that is, $F(x) = \exp(x)/1+\exp(x)$ and κ_σ is a Gaussian kernel of the form $\exp\{-\frac{1}{\sigma^2}(k-(t+1-d_{ij}))^2\}$. Thus we see that the spiking probability of the ith neuron depends on the past spiking history of itself and other neurons in the network to which it is connected. If there is no connection between any neuron in the network then also there is active firing due to the baseline firing parameter. Indeed, by (4.3) if $w_{ij} > 0$, then a spike of j leads to an increase in the spiking rate of i. In this chapter we only consider excitatory connections.

We have a Gaussian function which has the highest probability of spiking at d_{ij} and it decreases with normal probability both sides as we move further away from the true delay d_{ij}. We truncated this Gaussian function at k_l and k_u, the lower and upper bounds respectively. In practice we take $k_l = (t+1-d_{ij}-3\sigma)$ and $k_u = t$. In our scenario the spikes at and around the true delay, d_{ij}, have more weight than the spikes in remote past or recent past. The rationale behind this is that depending on the neurotransmitter the signal of a neuron takes few minutes to influence other neurons connected to it. We note that discrete stochastic models similar to Equation (4.3) has been proposed in Rigat et al. (2006), Truccolo et al. (2005) and Chornoboy et al. (1988) but they did not have any notion of delay.

The likelihood will be a product of conditional Bernoulli random variables. The joint likelihood of the model can be written as the product of the conditional likelihood

for each neuron. The first few spikes for all the neurons in the network are generated from a Bernoulli distribution with the baseline firing rate as the success probability. We define t', similar to Chapter 2.4, as time when all the neurons in the network fire at least once for the first time. $t' = \min_t \left\{ \sum_{s=1}^{t} X_j(s) > 0 \text{ for all } j \right\}$

Thus, the joint likelihood of the model becomes,

$$L = \prod_{t=t'}^{T-1} \prod_{i=1}^{J} \left[p_i(t+1|\mathcal{H}_t)^{X_i(t+1)} (1 - p_i(t+1|\mathcal{H}_t))^{1-X_i(t+1)} \right]$$

where $p_i(t+1|\mathcal{H}_t) = P(X_i(t+1)|\mathcal{H}_t)$ is the conditional probability of spiking for neuron i at time $t+1$ given the history of spiking till time t. We get the expression of $p_i(t+1|\mathcal{H}_t)$ from Equation (4.3). The first spike for all the neurons in the network are generated from a Bernoulli distribution with the baseline firing rates as the success probability. The log-likelihood becomes:

$$\begin{aligned} \ell_T &= \log(L) = \sum_{t=t'}^{T-1} \Big[\sum_{i=1}^{J} X_i(t+1) \log(p_i(t+1|\mathcal{H}_t)) \\ &+ (1 - X_i(t+1)) \log(1 - p_i(t+1|\mathcal{H}_t)) \Big] \end{aligned} \tag{4.4}$$

The model is based on the assumption that the spike train for the ith neuron at time $t+1$, $X_i(t+1)$, is dependent on the spiking history only till time t that is on $\mathcal{H}_t$. In other words, $X_i(t+1)|\mathcal{H}_t$ is conditionally independent of $X_l(t+1)|\mathcal{H}_t$ for all $l \neq i$. We perform the estimation of the parameters of the model under the point process generalized linear model paradigm. The details of the estimation is given in the next section.

4.2.5 Likelihood and Estimation with Unknown Structure

We follow similar method of iterative reweighted least squares (IRWLS) algorithm (Agresti (1990)) as in Chapter III for the estimation of the connectivity strengths

and baseline parameters of this model. This method is the standard choice for the maximum likelihood estimation of GLMs because of its computational simplicity, efficiency, and robustness (Green (1984)). In practice we implemented Iteratively Re-Weighted Least Squares (IRWLS) method to get the estimates of W and b. This model can also incorporate different spiking delays with different connection strengths between neurons.

The log likelihood of the model becomes

$$\ell \;=\; \log(L) = \sum_{i=1}^{J} \Big[\sum_{t=t'}^{T} X_i(t) \log(p_i(t|\mathcal{H}_t)) + (1 - X_i(t)) \log(1 - p_i(t|\mathcal{H}_t)) \Big]$$

We can refer to Section 3.3 of Chapter III to see the detailed derivation of the estimation of parameters. For each run of our algorithm we need to estimate the likelihood of our model (4.4) for a given delay and the neighboring delays. In order to estimate the likelihood we estimate our connectivity strength and baseline parameters and plug in the estimated value of our parameters in (4.4).

The parameter σ is estimated by numerically maximizing the profile likelihood for a given value of W, b and d. In Chapter II we have considered the inter-event time constraints which represent the time delays in neuronal connections are constants. In general there would be variation in these delays. The parameter σ controls the variation of the delay d between each pair of neurons. We initialize σ at 1 for the purposes of the simulation study. We iteratively estimate σ and the connectivity strength matrix, W and choose that value of σ where the profile likelihood is maximized.

4.3 Results

4.3.1 Simulated Neuronal Networks

In this section, we illustrate the analysis of a simulated neuronal network by employing our implementation of model (4.3) and Algorithm 1 to estimate delay. We show two example networks with the configurations shown in Figure 4.3. Ten thousand data points or in other words 10 seconds worth of data were generated for each neuron. The baseline firing rates were set to suitable values in order to ensure that the simulated firing rates are comparable to those observed for in vitro multi-electrode recordings. The baseline coefficient b of each neuron of each network is taken to be -3 so our baseline firing rates become 5 Hz approximately. We embedded excitatory connections in the network with the connectivity strengths between the neurons, w, ranging from (0,2) in the network. We apply our estimation method to analyze the simulated data.

We used a total number of neurons, N as 10, 20 and 30. We present the result for the algorithm which is run for 2000 iterations. Table 4.2 shows the values of the measures for different sizes of the two networks in Figure 4.3. Table 4.2 shows the values of the measures for different sizes of the network 2 in Figure 4.3 similarly.

We use three measures to compare the performance of the algorithm and our estimation procedure. The measures are:

1) $w_{diff} = \sqrt{\frac{1}{J^2}\sum_{ij}(\hat{w}_{ij} - w_{ij})^2}$

2) $d_{diff} = \sqrt{\frac{1}{J^2}\sum_{ij}(\hat{d}_{ij} - d_{ij})^2}$

3) $\delta = \frac{1}{J^2}\sum_{ij}\mathbf{1}[sign(\hat{w}_{ij}) - sign(w_{ij})]$ where $sign(x) = 1$ if $x \neq 0$ and 0 if $x = 0$

The measures w_{diff} and d_{diff} are used to measure the deviation of the estimated parameters from the true parameter and the discrepancy measure δ is a function of both false positives and false negatives. We run the algorithm and at each iteration it

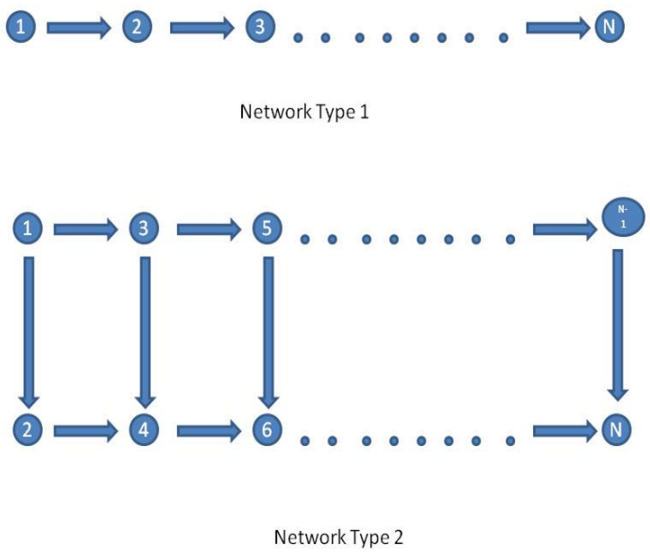

Figure 4.3: Simulated Neuronal Networks

selects the delay and connectivity strength having the maximum likelihood among it's neighbors. At the end of the procedure the estimated delay and connectivity strength matrix corresponds to the value having maximum estimated likelihood among all the visited delays. We compare the performance of the algorithm for different networks using these three measures.

Table 4.2: Showing the difference in connectivity strength and delay

	Size of Network	w_{diff}	d_{diff}	δ
Network 1	10 neurons	0.0034	0	0
	20 neurons	0.0543	0	0
	30 neurons	0.1119	0.1624	0.13
Network 2	10 neurons	0.0076	0	0
	20 neurons	0.0781	0	0
	30 neurons	0.1956	0.3671	0.14

We observe that the algorithm detects the correct network with the true delays for network size 10 and 20 from Table 4.2. But the performance of the algorithm gets worse with the increase in the number of neurons in the network if we fix the

number of iterations. Thus, for a fixed number of iterations Table 4.2 shows that the performance of the algorithm gets worse with the increase in number of neurons in the network. Also, the algorithm becomes computationally intensive as the number of neurons and the graph complexity increases as shown in the network having 30 neurons. As a result we cannot run the algorithm for larger of iterations. In such a setting the algorithm fails to search a reasonable span of possible values of the delay and thus loses its efficiency. Hence for complicated networks having a large number of neurons, we partition the graph into sub-clusters consisting of smaller number of neurons in order to increase the efficiency of the algorithm. For our endeavor we use the graph partitioning methods described in the next section.

4.3.2 Graph Partition

Graph Partition via clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, biology, social sciences and psychology. In virtually every scientific field dealing with empirical data, people attempt to get a first impression on their data by trying to identify groups of similar behavior in data. The intuition of clustering is to separate nodes in different groups according to their similarities. For data given in form of a similarity graph, this problem can be restated as a partition of the graph such that the edges between different groups have a very low weight and the edges within a group have high weight. Thus, a partition of the graph should be such that the points in different clusters are dissimilar from each other and the points within the same cluster are similar to each other. In this section, we will see how a special form of spectral clustering called the normalized cut can be derived as an approximation to such graph partitioning problems.

We follow the methods of Shi and Malik (2000) which propose a normalized cut criterion for segmenting a graph. The normalized cut criterion measures both the total

dissimilarity between the different groups as well as the total similarity within the groups. Shi and Malik (2000) show that an efficient computational technique based on a generalized eigen value problem can be used to optimize this criterion. We have applied this approach to our scenario thus partitioning the neuronal graphical model and found the results to be very encouraging. We also tried implementing other forms of spectral clustering such as Ng et al. (2002) and found that the normalized cut method works best in our scenario.

A graph $G = (V, E)$ can be partitioned into two disjoint sets, A, B, $A \bigcup B = V$, $A \bigcap B = \varnothing$ by simply removing edges connecting the two parts. The approach is related to the graph theoretic formulation of grouping. The set of points in an arbitrary feature space are represented as a weighted directed graph, where the nodes of the graph are the points in the feature space, and an edge is formed between every pair of nodes. The weight on each edge, $\omega_{i,j}$, is a function of the similarity between nodes i and j in Shi and Malik (2000). We use the 2-node counts matrix instead of the weights for each edge. In grouping, we seek to partition the set of vertices into disjoint sets $V_1, V_2, \cdots, V_m$, where the similarity among the vertices in a set V_i is high and similarly across different sets V_i, V_j is low. The degree of dissimilarity between these two pieces can be computed as total weight of the edges that have been removed. In graph theoretic language, it is called the cut and is defined as:

$$Cut(A, B) = \sum_{u \in A, v \in B} \omega_{u,v} \tag{4.5}$$

The optimal bipartitioning of a graph is the one that minimizes this cut value. Although there are an exponential number of such partitions, finding the minimum cut of a graph is a well-studied problem and there exist efficient algorithms for solving it (Shi and Malik (2000)). We apply the normalized cut partitioning algorithm to

neuron network data. Then, we implement the Algorithm 1 for each of the subgraph separately. Finally, we consider that $\hat{w}_{ij}$, combining the results of all the subgroups, as the final estimate of the connectivity strength between neurons i and j.

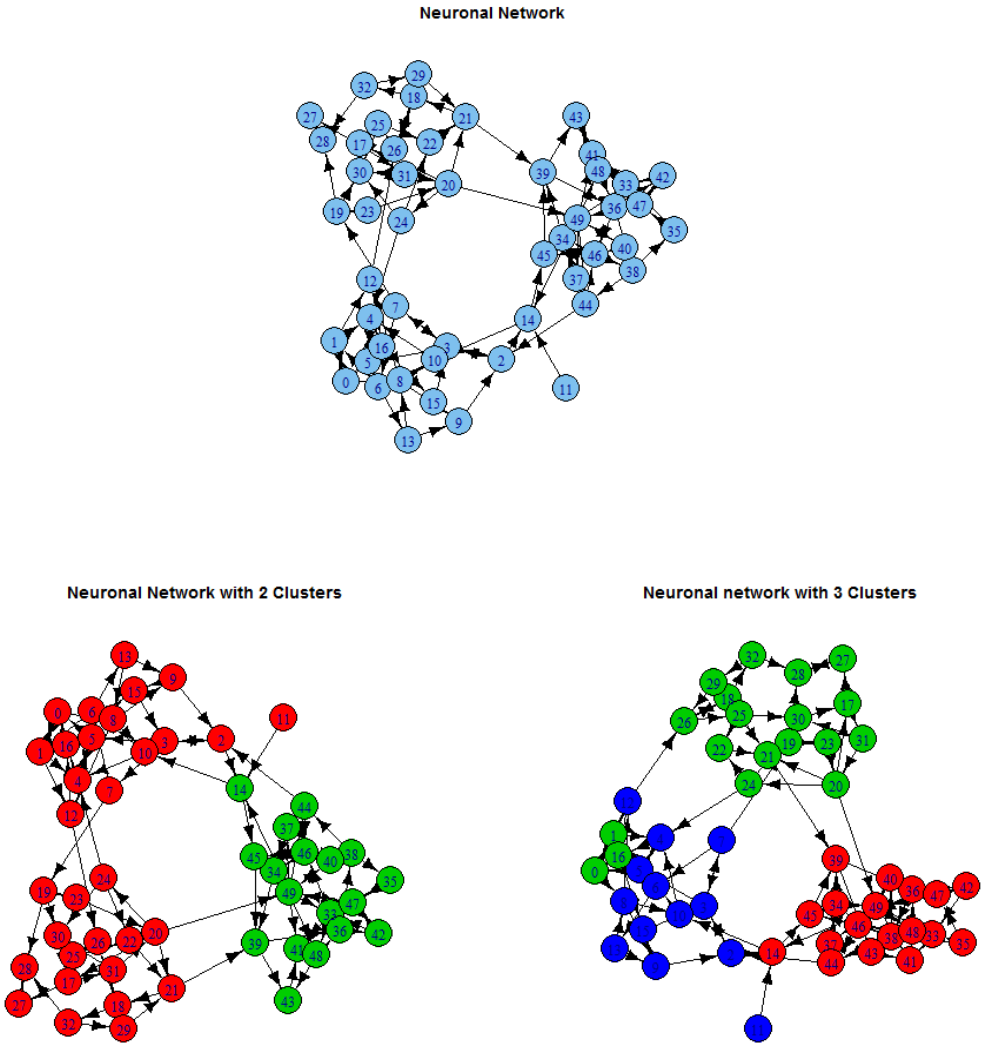

Figure 4.4: Partitioning of Neuronal Networks consisting of 50 neurons into two and three clusters respectively

We implement the normalized cut algorithm on an example network consisting of 50 neurons shown in Figure 4.4. We calculate the matrix of 2-node counts for the example network. Using the delay corresponding to the maximum 2-node counts

we compute a matrix of 2-node counts and use it as the weight matrix, $\omega_{i,j}$, for the partitioning. Figure 4.4 shows the network of neurons after and before partitioning using the normalized cut method. We partition the network into two and three clusters respectively and apply our algorithm to each of the clusters separately. We observe that the normalized cut method is able to find the inherent clusters in the data based on the total dissimilarity between the different groups as well as the total similarity within the groups. There are three clusters present in the original data with the edges between different clusters having a very low weight and the edges within a cluster having high weight. We apply the normalized cut method and specify the number of clusters as two and three respectively. The clustering method successfully isolates the most distinct cluster among the three clusters when we specify two clusters while implementing the clustering. We present the results of the combined connectivity strength from different clusters in Table 4.3. We are able to analyze larger number of neurons after the partition though we lose some of the connections due to the partitioning.

Table 4.3: Showing the differences in combined connectivity strength and delay for different networks presented in Figure 4.4

Type of Network	w_{diff}	k_{diff}	δ
50 Neuron Network (2 clusters)	0.0104	0.0877	0.08
50 Neuron Network (3 clusters)	0.0621	0.1272	0.13

We also apply the normalized cut method on the network with 30 neurons shown in Figure 4.3 of Section 4.3.1. We divide 30 neuron network data into two clusters consisting of 18 and 12 neurons such that the graph is the one that minimizes the cut value shown in Equation (4.5). Then we implement the Algorithm 1 for each of the clusters separately. Finally, we consider the connectivity matrix $\hat{W}$, combining all the clusters, as the final estimate of the connectivity strength. We present the results of the combined connectivity strength from the two subgroups in Table 4.4. We observe

that the performance of our algorithm improves after the partition though we lose some of the connections due to the partitioning. We conclude that the algorithm performs better if there is some inherent clustering in the network. However, we can apply our method to networks with chain structures without any clustering as well.

Table 4.4: Showing the differences in combined connectivity strength and delay for different networks presented in Figure 4.3

Type of Network	w_{diff}	k_{diff}	δ
30 Neuron Network 1	0.0618	0.1425	0.06
30 Neuron Network 2	0.0752	0.2564	0.09

4.3.3 Application to cultured cortical neurons

We used our methods to analyze spike train data recorded from cultures of cortical neurons. Wagenaar et al. (2006) made these recordings publicly available, and focused on characterizing "population bursts" in the spiking activity of the cultures. Such bursts, defined as brief periods of time during which the firing rate of several cells or electrodes greatly exceeds the baseline rate, are a common feature observed in cultures of many different types of neurons. Here, our focus is not to characterize the bursts but rather to detect precisely timed spiking patterns involving multiple neurons. From these patterns we can estimate the strength of functional connectivity between different neurons in the culture as well as the delay time of spiking. Our analysis methods assume that the firing rates of individual neurons are relatively stationary in the analysis window. Thus, for our analysis we chose a segment of the data which did not contain bursts. We analyzed 60 seconds of data from culture 2-1-35. We partition the network data of 30 neurons into 3 subnetworks containing 11, 9 and 10 neurons respectively as shown in Figure 4.5. To visualize the functional connectivity present in the culture, we draw network graphs based on our Algorithm 1 and estimation procedure described in Section 4.2.5. If there is a significant edge

with any delay for an $i - j$ pair, we draw a directed edge $i \to j$ connecting them. Table 4.5 and Figure 4.5 shows the 58 edges that were significant for culture 2-1-35.

Table 4.5: 2-node episodes of culture 2-1-35 that were significant using Algorithm 1

#	$i[k] - j$	$\hat{w}$		#	$i[k] - j$	$\hat{w}$
1	48[9]-32	1.28		30	57[6]-65	3.53
2	87[8]-32	2.15		31	65[5]-65	3.29
3	66[1]-32	2.96		32	75[6]-65	4.68
4	55[4]-48	1.42		33	84[3]-13	5.58
5	87[8]-48	2.96		34	78[8]-75	4.39
6	48[5]-14	1.57		35	78[8]-35	3.29
7	87[7]-14	2.43		36	25[2]-35	4.21
8	14[1]-56	1.01		37	27[8]-78	4.39
9	55[1]-56	3.65		38	57[8]-78	4.14
10	58[4]-56	1.21		39	78[8]-84	4.28
11	87[10]-55	2.17		40	13[3]-25	4.76
12	56[4]-87	3.75		41	22[1]-46	1.29
13	21[5]-87	2.73		42	82[1]-82	2.16
14	66[1]-87	3.05		43	34[10]-82	0.84
15	24[1]-24	2.32		44	42[3]-76	2.22
16	66[5]-24	1.08		45	34[9]-76	0.92
17	32[1]-21	3.51		46	68[4]-22	1.63
18	48[1]-21	1.93		47	86[1]-22	1.46
19	14[8]-66	1.31		48	42[3]-42	2.10
20	55[5]-66	1.80		49	34[2]-42	1.16
21	24[1]-66	1.28		50	77[4]-68	1.15
22	48[5]-58	1.80		51	86[1]-68	1.94
23	87[7]-58	2.80		52	72[1]-77	0.92
24	21[8]-58	1.80		53	77[1]-72	3.68
25	75[3]-85	2.03		54	86[4]-72	1.16
26	13[7]-27	5.27		55	68[1]-86	1.61
27	27[4]-57	5.78		56	34[2]-86	1.50
28	65[9]-57	4.87		57	22[1]-34	1.22
29	35[4]-57	4.63		58	72[1]-34	1.16

Although we considered episodes with delays of up to 50 ms, all the significant episodes had delays of less than 10 ms. Fast delays such as these are consistent with the timescale of the action for AMPA, a common excitatory neurotransmitter in the cortex. Overall, our results are consistent with the patterns reported in a previous study of precisely timed patterns in these cultures Rolston et al. (2007). Specifically,

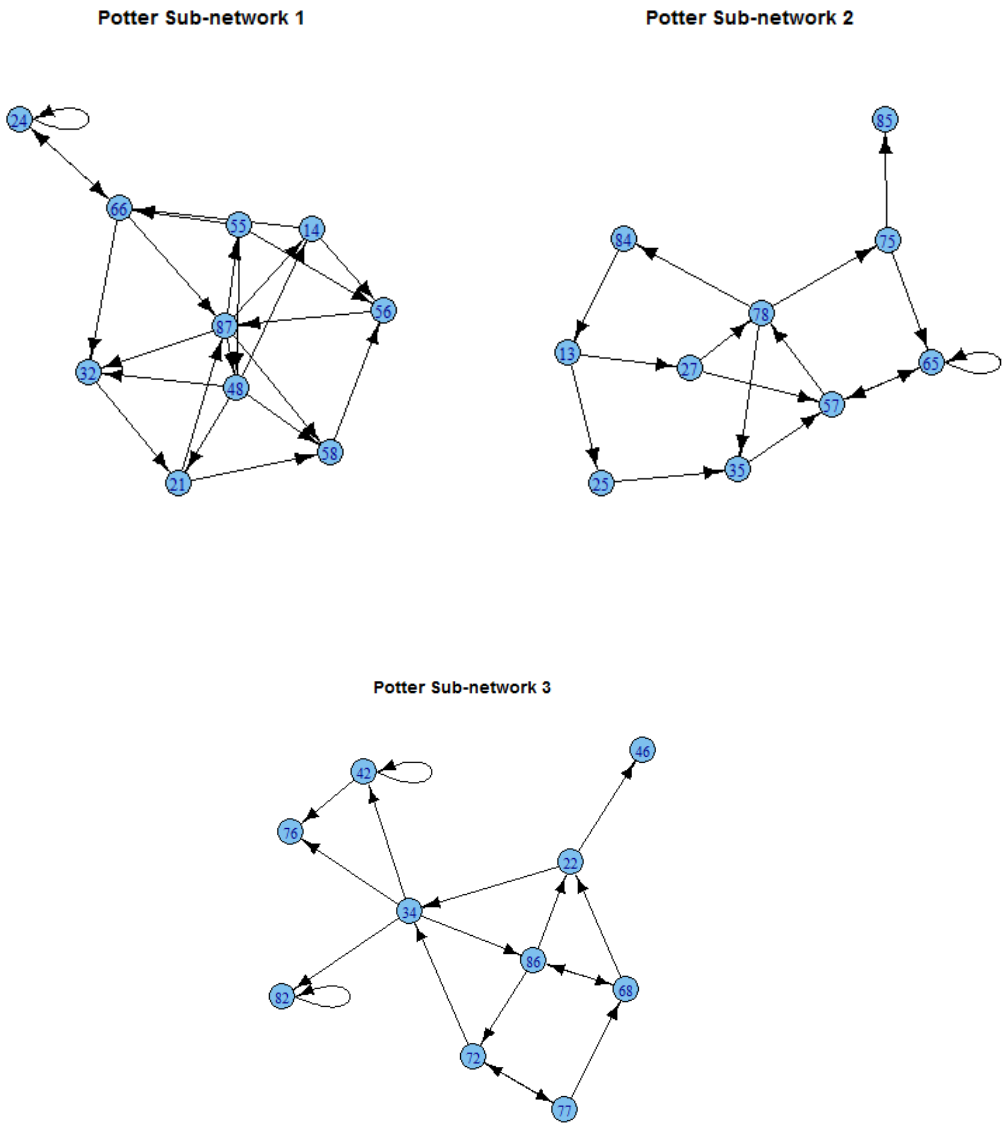

Figure 4.5: Functional connectivity in culture 2-1-35.

of the thirteen 2-node patterns mentioned in Rolston et al. (2007), six of them were also found to be significant in our analysis (68-78, 78-77, 77-48, 41-42, 42-32 and 37-38).

4.4 Discussion

In this chapter we introduced a new discrete time graphical model for the analysis of multiple spike train recordings. We defined the spiking probabilities over a finite grid of time intervals similar to the approaches taken in Brillinger (1988) and Brillinger and Villa (1997). We developed an iterative method to estimate delays and connectivity strength matrices. The method is used to reconstruct the graphical structure of the functional connectivity among a group of neurons. This delay is usually caused by the axonal delay of spike propagation and the delay at synapses due to chemical diffusion. In the graphical model, we have assumed that the mean time delays in neuronal connections are constants with some variations characterized by σ in the model.

We have shown that a technique based on generalized linear models can be used to perform consistent model selection in binary graphical models. Our analysis applies to the settings in which both the number of neurons and number of connectivities between them are allowed to increase with the number of neurons in the network. Simulation results show the accuracy of these predictions. We successfully used the normalized cut method (Shi and Malik (2000)) for graph partitioning and apply our method to bigger networks involving larger number of neurons, mostly not considered in existing literature. We demonstrate the effectiveness of our method both on simulated neuronal spike data as well as on data collected from in vitro neuronal cultures of cortical neurons.

In our current algorithm we are thresholding the insignificant connectivity strengths to zero. Thus, an interesting direction for future work is to apply l_1-regularized linear

regression to our model and thus thresholding the parameters. It would be interesting to obtain estimates in this setting and compare the performance with the model without thresholding. Ravikumar et al. (2010) have worked on estimating the graph associated with a binary Ising Markov random field and described a method based on l_1-regularized logistic regression, in which the neighborhood of any given node is estimated by performing logistic regression subject to an l_1-constraint. We are dealing with directed graphs while Ravikumar et al. (2010) have considered undirected graphs. Furthermore, incorporating regularization methods in our modeling framework will be another interesting area of future research.

APPENDICES

APPENDIX A

Derivations of Covariance of τ Based on N

In this section, we derive the covariance of the test statistic τ based on the total number of occurrences N. We define N_{AB} as the count of the joint occurrences of neurons A and B. All the estimators in this section are based on the total number of episodes N_{AB}. N is binomial$(L-k, P_{AB})$, where P_{AB} is the probability of the event $A[k]B$ defined in Section 2.2.1. Thus we get $Var(\hat{P}_{AB})$ as:

$$\hat{P}_{AB} = \frac{N_{AB}}{L-k}$$
$$Var(\hat{P}_{AB}) = \frac{P_E(1-P_E)}{(L-k)}$$

To compute $Var(\hat{P}_A\hat{P}_B)$, recall that $\hat{P}_A\hat{P}_B = \frac{N_A N_B}{(L-k)^2}$. So, $Var(\hat{P}_A\hat{P}_B) = \frac{1}{(L-k)^4}Var(N_A N_B)$. Note, however, that we are computing the variance under possible dependence, so N_A and N_B are not independent. We can write:

$$\begin{aligned} N_A N_B &= \sum_{t=1}^{L-k} I_A(t) \sum_{r=1}^{L-k} I_B(r+k) \\ &= \underbrace{\sum_{t=1}^{L-k} I_A(t) I_B(t+k)}_{N_{AB}} + \underbrace{\sum_{s=1}^{L-k} \sum_{r \neq s} I_A(s) I_B(r+k)}_{N'_{AB}}. \\ &= \end{aligned}$$

Therefore,

$$\begin{aligned} E[N_A N_B] &= E[N_{AB}] + E[N'_{AB}] \\ &= (L-k)P_E + (L-k)(L-k-1)P_A P_B \end{aligned}$$

Thus, finally we get,

$$Var(N_A N_B) = Var(N_{AB}) + Var(N'_{AB}) + 2Cov(N_{AB}, N'_{AB})$$

where $Var(N_{AB}) = (L-k)P_E(1-P_E)$ and the calculations of $Var(N'_{AB})$ and $Cov(N_{AB}, N'_{AB})$ are done using the expressions of N_A and N_B.

Lastly, we compute the third term $Cov(\hat{P}_{AB}, \hat{P}_A\hat{P}_B)$ as follows:

$$Cov(\hat{P}_{AB}, \hat{P}_A\hat{P}_B) = \frac{Cov(N_{AB}, N_A N_B)}{(L-k)^3}$$

$$\begin{aligned} Cov(N_{AB}, N_A N_B) &= E(N_{AB}N_A N_B) - E(N_{AB})E(N_A N_B) \\ E(N_{AB}N_A N_B) &= E[\sum_{t=1}^{L-k} I_A(t)I_B(t+k) \sum_{s=1}^{L-k} I_A(s)I_B(s+k) \\ &+ \sum_{t=1}^{L-k} I_A(t)I_B(t+k) \sum_{s=1}^{L-k} \sum_{r \neq s} I_A(s)I_B(r+k)] \\ &= E[N_{AB}^2] + E[N_{AB}N'_{AB}] \end{aligned}$$

$$E[N_{AB}N'_{AB}] = E[\sum_{t=1}^{L-k} I_A(tI_B(t+k) \sum_{s=1}^{L-k} \sum_{r\neq s} I_A(s)I_B(r+k)]$$

$$= \begin{cases} E[\sum_{t=1}^{L-k} \sum_{r\neq t} I_A(t)I_B(t+k)I_B(r+k)] & \text{if } t = s \neq r; \\ E[\sum_{t=1}^{L-k} \sum_{s\neq t} I_A(t)I_B(t+k)I_A(s)] & \text{if } t = r \neq s \\ E[\sum_{t=1}^{L-k} \sum_{s\neq t} \sum_{r\neq(t,s)} I_A(t^*)I_B(t+k)I_A(s)I_B(r+k)] & \text{if } s \neq r \neq t. \end{cases}$$

$$= \begin{cases} E[N_{AB}] \times (L-k-1)P_B & \text{if } t = s \neq r; \\ E[N_{AB}] \times (L-k-1)P_A & \text{if } t = r \neq s \\ E[N_{AB}] \times (L-k-1)(L-k-2)P_AP_B & \text{if } s \neq r \neq t. \end{cases}$$

APPENDIX B

Derivations of Covariance of False Edge Test ξ Based on N

Similarly as before, for the False edge elimination test we compute the variance and covariance terms one by one. We compute $Var(\hat{P}_{A\bar{B}C})$ as follows:

$$\hat{P}_{A\bar{B}C} = \frac{N_{A\bar{B}C}}{L-2k}$$
$$Var(\hat{P}_{A\bar{B}C}) = \frac{Var(N_{A\bar{B}C})}{(L-2k)^2}$$
$$Var(N_{A\bar{B}C}) = \frac{P_{A\bar{B}C}(1-P_{A\bar{B}C})}{(L-2k)}$$

Now, we compute $Var(\hat{P}_A\hat{P}_{\bar{B}}\hat{P}_C)$ as follows:

$$\hat{P}_A\hat{P}_{\bar{B}}\hat{P}_C = \frac{1}{(L-2k)^6}Var(N_A N_{\bar{B}} N_C)$$

Now, under H_0, N_A $N_{\bar{B}}$ and N_C are independent. Therefore we get:

$$\begin{aligned} Var(N_A N_{\bar{B}} N_C) &= Var(N_A)Var(N_{\bar{B}})Var(N_C) + Var(N_A)Var(N_{\bar{B}})E(N_C)^2 \\ &+ Var(N_A)Var(N_C)E(N_{\bar{B}})^2 + Var(N_{\bar{B}})Var(N_C)E(N_A)^2 \\ &+ E[N_A]^2 E[N_{\bar{B}}]^2 Var(N_C) + E[N_A]^2 E[N_C]^2 Var(N_{\bar{B}}) \\ &+ E[N_{\bar{B}}]^2 E[N_C]^2 Var(N_C) \end{aligned}$$

Lastly, we compute the third term $Cov(\hat{P}_{A\bar{B}C}, \hat{P}_A \hat{P}_{\bar{B}} \hat{P}_C)$ as follows:

$$Cov(\hat{P}_{A\bar{B}C}, \hat{P}_A \hat{P}_{\bar{B}} \hat{P}_C) = \frac{Cov(N_{A\bar{B}C}, N_A N_{\bar{B}} N_C)}{(L-2k)^4}$$

$$\begin{aligned} Cov(N_{A\bar{B}C}, N_A N_{\bar{B}} N_C) &= E(N_{A\bar{B}C} N_A N_{\bar{B}} N_C) - E(N_{A\bar{B}C})E(N_A N_{\bar{B}} N_C) \\ E(N_{A\bar{B}C} N_A N_{\bar{B}} N_C) &= E[N^2_{A\bar{B}C}] + E[N_{A\bar{B}C} N'_{A\bar{B}C}] \end{aligned}$$

$$\begin{aligned} E[N_{A\bar{B}C} N'_{A\bar{B}C}] &= E[\sum_{t=1}^{L-k} I_A(t) I_{\bar{B}}(t+k) I_C(t+2k) \sum_{t=1}^{L-k} \sum_{s \neq t} \sum_{r \neq s \neq t} I_A(t) I_{\bar{B}}(s+k) I_C(r+2k)] \\ &= \begin{cases} E[N_{A\bar{B}C}] \times (L-2k-1) P_A & \text{if } t = r = s \neq t; \\ E[N_{A\bar{B}C}] \times (L-2k-1) P_{\bar{B}} & \text{if } t = r = t \neq s \\ E[N_{A\bar{B}C}] \times (L-2k-1) P_C & \text{if } t = t = s \neq r \\ E[N_{A\bar{B}C}] \times (L-2k-1)(L-2k-2) P_A P_{\bar{B}} & \text{if } t = r \neq s \neq t \\ E[N_{A\bar{B}C}] \times (L-2k-1)(L-2k-2) P_A P_C & \text{if } t = s \neq r \neq t \\ E[N_{A\bar{B}C}] \times (L-2k-1)(L-2k-2) P_{\bar{B}} P_C & \text{if } t = t \neq s \neq r \\ E[N_{A\bar{B}C}] \times (L-2k-1)(L-2k-2) \times & \\ (L-2k-3) P_A P_{\bar{B}} P_C & \text{if } t \neq t \neq s \neq r \end{cases} \end{aligned}$$

APPENDIX C

Derivations of Covariance of τ Based on M

We show the calculation of the variance and covariance terms in the Equation (2.10) here. We compute $Var(\hat{P}_{AB})$ as follows:

$$\hat{P}_{AB} = \Big(\frac{L-k}{M_{AB}} - k\Big)^{-1}$$
$$Var(\hat{P}_{AB}) = [f'(E(M_{AB}))]^2 Var(M_{AB})$$
$$Var(M_{AB}) = \frac{(L-k)P_E(1-P_E)}{(1+kP_E)^3}$$
$$f'(M_{AB}) = \frac{L-k}{M_{AB}^2}\Big(\frac{L-k}{M_{AB}} - k\Big)^{-2}$$

We compute the third term $Cov(\hat{P}_{AB}, \hat{P}_A\hat{P}_B)$ as follows:

$$Cov(\hat{P}_{AB}, \hat{P}_A\hat{P}_B) = \frac{Cov(M_{AB}, N_AN_B)}{(L-k)E[M_{AB}]^2(\frac{L-k}{E[M_{AB}]} - k)^2}$$

$$
\begin{aligned}
Cov(M_{AB}, N_A N_B) &= E(M_{AB} N_A N_B) - E(M_{AB}) E(N_A N_B) \\
E(M_{AB} N_A N_B) &= E[\sum_{t^*=1}^{L-k} I_A(t^*) I_B(t^*+k) \sum_{s=1}^{L-k} I_A(s) I_B(s+k) \\
&+ \sum_{t^*=1}^{L-k} I_A(t^*) I_B(t^*+k) \sum_{s=1}^{L-k} \sum_{r \neq s} I_A(s) I_B(r+k)] \\
&= E[M_{AB} N_{AB}] + E[M_{AB} N'_{AB}]
\end{aligned}
$$

Here t^* refers to the non-overlapped t's $\rightarrow$ there are M_{AB} of them.

$$
\begin{aligned}
E[M_{AB} N_{AB}] &= E[M_{AB}(M_{AB} + \sum_{j=1}^{M_{AB}} R_j)] \\
&= E[M_{AB}^2] + E[M_{AB} \sum_{j=1}^{M_{AB}} R_j] \\
&= E[M_{AB}^2 (1 + k P_E)]
\end{aligned}
$$

$E[M_{AB} N'_{AB}]$

$$
\begin{aligned}
&= E[\sum_{t^*=1}^{L-k} I_A(t^*) I_B(t^*+k) \sum_{s=1}^{L-k} \sum_{r \neq s} I_A(s) I_B(r+k)] \\
&= \begin{cases}
E[\sum_{t^*=1}^{L-k} \sum_{r \neq t^*} I_A(t^*) I_B(t^*+k) I_B(r+k)] & \text{if } t^* = s \neq r; \\
E[\sum_{t^*=1}^{L-k} \sum_{s \neq t^*} I_A(t^*) I_B(t^*+k) I_A(s)] & \text{if } t^* = r \neq s \\
E[\sum_{t^*=1}^{L-k} \sum_{s \neq t^*} \sum_{r \neq (t^*, s)} I_A(t^*) I_B(t^*+k) I_A(s) I_B(r+k)] & \text{if } s \neq r \neq t^*.
\end{cases} \\
&= \begin{cases}
E[M_{AB}] \times (L-k-1) P_B & \text{if } t^* = s \neq r; \\
E[M_{AB}] \times (L-k-1) P_A & \text{if } t^* = r \neq s \\
E[M_{AB}] \times (L-k-1)(L-k-2) P_A P_B & \text{if } s \neq r \neq t^*.
\end{cases}
\end{aligned}
$$

APPENDIX D

Derivations of Covariance of False Edge Test ξ Based on M

Similarly as before, we compute the variance and covariance terms in the Equation (2.10). We compute $Var(\hat{P}_{A\bar{B}C})$ as follows:

$$\begin{aligned}
\hat{P}_{A\bar{B}C} &= \Big(\frac{L-2k}{M_{A\bar{B}C}} - 2k\Big)^{-1} = f(M_{A\bar{B}C}) \\
Var(\hat{P}_{A\bar{B}C}) &= [f'(E(M_{A\bar{B}C}))]^2 Var(M_{A\bar{B}C}) \\
Var(M_{A\bar{B}C}) &= \frac{(L-2k)P_{A\bar{B}C}(1-P_{A\bar{B}C})}{(1+2kP_{A\bar{B}C})^3} \\
f'(M_{A\bar{B}C}) &= \frac{L-2k}{M_{A\bar{B}C}^2}\Big(\frac{L-2k}{M_{A\bar{B}C}} - 2k\Big)^{-2}
\end{aligned}$$

Now, we compute $Var(\hat{P}_A\hat{P}_{\bar{B}}\hat{P}_C)$ as follows:

$$\begin{aligned}
\hat{P}_A\hat{P}_{\bar{B}}\hat{P}_C &= \frac{N_A N_{\bar{B}} N_C}{(L-2k)^3} = g(N_A N_{\bar{B}} N_C) \\
Var(\hat{P}_A\hat{P}_{\bar{B}}\hat{P}_C) &= Var(g(N_A N_{\bar{B}} N_C)) \\
&= [g'(N_A N_{\bar{B}} N_C)]^2 Var(N_A N_{\bar{B}} N_C) \\
&= \frac{1}{(L-2k)^6} Var(N_A N_{\bar{B}} N_C)
\end{aligned}$$

Now, under H_0, N_A $N_{\bar{B}}$ and N_C are independent. Therefore we get:

$$\begin{aligned} Var(N_A N_{\bar{B}} N_C) &= Var(N_A)Var(N_{\bar{B}})Var(N_C) + Var(N_A)Var(N_{\bar{B}})E(N_C)^2 \\ &+ Var(N_A)Var(N_C)E(N_{\bar{B}})^2 + Var(N_{\bar{B}})Var(N_C)E(N_A)^2 \\ &+ E[N_A]^2 E[N_{\bar{B}}]^2 Var(N_C) + E[N_A]^2 E[N_C]^2 Var(N_{\bar{B}}) \\ &+ E[N_{\bar{B}}]^2 E[N_C]^2 Var(N_C) \end{aligned}$$

Lastly, we compute the 3rd term $Cov(\hat{P}_{A\bar{B}C}, \hat{P}_A \hat{P}_{\bar{B}} \hat{P}_C)$ as follows:

$$\begin{aligned} Cov(\hat{P}_{A\bar{B}C}, \hat{P}_A \hat{P}_{\bar{B}} \hat{P}_C) &= Cov(f(M_{A\bar{B}C}), g(N_A N_{\bar{B}} N_C)) \\ &= Cov(M_{A\bar{B}C}, N_A N_{\bar{B}} N_C) f'(E[M_{A\bar{B}C}]) g'(E[N_A N_{\bar{B}} N_C]) \\ &= \frac{Cov(M_{A\bar{B}C}, N_A N_{\bar{B}} N_C)}{(L-2k)^2 E[M_{A\bar{B}C}]^2 (\frac{L-2k}{E[M_{A\bar{B}C}]} - 2k)^2} \end{aligned}$$

$$\begin{aligned} Cov(M_{A\bar{B}C}, N_A N_{\bar{B}} N_C) &= E(M_{A\bar{B}C} N_A N_{\bar{B}} N_C) - E(M_{A\bar{B}C}) E(N_A N_{\bar{B}} N_C) \\ E(M_{A\bar{B}C} N_A N_{\bar{B}} N_C) &= E[M_{A\bar{B}C} N_{A\bar{B}C}] + E[M_{A\bar{B}C} N'_{A\bar{B}C}] \end{aligned}$$

$$\begin{aligned} E[M_{A\bar{B}C} N_{A\bar{B}C}] &= E[M_{A\bar{B}C}(M_{A\bar{B}C} + \sum_{j=1}^{M_{A\bar{B}C}} R_j)] \\ &= E[M^2_{A\bar{B}C}] + E[M_{A\bar{B}C} \sum_{j=1}^{M_{A\bar{B}C}} R_j] \\ &= E[M^2_{A\bar{B}C}(1 + kP_E)] \end{aligned}$$

$$E[M_{A\bar{B}C}N'_{A\bar{B}C}]$$

$$= E[\sum_{t^*=1}^{L-k} I_A(t^*)I_{\bar{B}}(t^*+k)I_C(t^*+2k)\sum_{t=1}^{L-k}\sum_{s\neq t}\sum_{r\neq s\neq t} I_A(t)I_{\bar{B}}(s+k)I_C(r+2k)]$$

$$= \begin{cases} E[M_{A\bar{B}C}]\times(L-2k-1)P_A & \text{if } t^*=r=s\neq t; \\ E[M_{A\bar{B}C}]\times(L-2k-1)P_{\bar{B}} & \text{if } t^*=r=t\neq s \\ E[M_{A\bar{B}C}]\times(L-2k-1)P_C & \text{if } t^*=t=s\neq r \\ E[M_{A\bar{B}C}]\times(L-2k-1)(L-2k-2)P_AP_{\bar{B}} & \text{if } t^*=r\neq s\neq t \\ E[M_{A\bar{B}C}]\times(L-2k-1)(L-2k-2)P_AP_C & \text{if } t^*=s\neq r\neq t \\ E[M_{A\bar{B}C}]\times(L-2k-1)(L-2k-2)P_{\bar{B}}P_C & \text{if } t^*=t\neq s\neq r \\ E[M_{A\bar{B}C}]\times(L-2k-1)(L-2k-2)\times & \\ (L-2k-3)P_AP_{\bar{B}}P_C & \text{if } t^*\neq t\neq s\neq r \end{cases}$$

Here t^* refers to the non-overlapped t's $\rightarrow$ there are $M_{A\bar{B}C}$ of them.

APPENDIX E

Proof of Lemma III.1

Proof of Lemma III.1. The conditional probability for our model is:

$$p_k(t+1|\mathcal{H}_t) = \mathbb{P}\{X_k(t+1)=1|\mathcal{H}_t\} = F\Big(\sum_{i=1}^{J} w_{ik} g(X_i(\tau_k(t):t]) + b_k\Big)$$

Observe that,

$$\begin{aligned}
&\frac{\partial p_k(t+1|\mathcal{H}_t)}{\partial w_{ij}} \\
&\qquad = \begin{cases} \frac{\partial}{\partial w_{ij}} F\Big(\sum_{i=1}^{J} w_{ij} g(X_i(\tau_j(t):t]) + b_j\Big) & \text{if} \quad j=k \\ 0 & \text{otherwise.} \end{cases} \\
&\qquad = \begin{cases} F'\Big(\sum_{i=1}^{J} w_{ij} g(X_i(\tau_j(t):t]) + b_j\Big) g(X_i(\tau_j(t):t]) & \text{if} \quad j=k \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{E.1}$$

Now we assume logit link then the function $F(.)$ becomes inverse logit.

$$\begin{aligned} F(\eta_j) &= \frac{\exp(\eta_j)}{1+\exp(\eta_j)} = p_j(t+1|\mathcal{H}_t) \\ \text{where} \quad \eta_j &= \sum_{i=1}^{J} w_{ij} g(X_i(\tau_j(t):t]) + b_j. \\ \Rightarrow F'(\eta_j) &= \frac{\exp(\eta_j)}{(1+\exp(\eta_j))^2} = p_j(t+1|\mathcal{H}_t)(1-p_j(t+1|\mathcal{H}_t)) \end{aligned} \tag{E.2}$$

Therefore combining Equations (E.1) and (E.2) we get,

$$\frac{\partial p_k(t+1|\mathcal{H}_t)}{\partial w_{ij}} = \begin{cases} p_k(t+1|\mathcal{H}_t)(1-p_k(t+1|\mathcal{H}_t)) g(X_i(\tau_k(t):t]) & \text{if} \quad j=k \\ 0 & \text{otherwise.} \end{cases} \tag{E.3}$$

By Equation (E.3) and taking a partial derivative with respect to w_{ij} in Equation (3.2) we obtain

$$\begin{aligned} \frac{\partial \ell_T}{\partial w_{ij}} &= \sum_{t=1}^{T-1} \Big[\frac{X_k(t+1)}{p_k(t+1|\mathcal{H}_t)} \frac{\partial p_k(t+1|\mathcal{H}_t)}{\partial w_{ij}} - \frac{1-X_k(t+1)}{1-p_k(t+1|\mathcal{H}_t)} \frac{\partial p_k(t+1|\mathcal{H}_t)}{\partial w_{ij}} \Big] \\ &= \sum_{t=1}^{T-1} (X_j(t+1) - p_j(t+1|\mathcal{H}_t)) g(X_i(\tau_j(t):t]) \end{aligned} \tag{E.4}$$

Note that only the term corresponding to $k=j$ remains from the inner summation because of Equation (E.3). We can similarly derive the expression for $\partial \ell / \partial b_j$ and the derivation is in fact simpler. Since the parameter b_j in Equation (3.1) appears with a coefficient equal to 1. □

APPENDIX F

Proof of Propositions F.1, F.2 and F.4

Proposition F.1. *If model (3.1) is identifiable then condition (i) of Theorem III.4 holds.*

Proof. Consider now the log–likelihood

$$\ell_T(X;\theta) := \sum_{t=1}^{T-1}\sum_{j=1}^{J} X_j(t+1)\log(p_j(t+1|\mathcal{H}_t)) + (1-X_j(t+1))\log(1-p_j(t+1|\mathcal{H}_t)).$$

Then the log–likelihood can be written as follows:

$$\ell_T(X;\theta) := \sum_{t=1}^{T-1}\sum_{j=1}^{J} h_j(D(t+1))\log(p_j(D(t);\theta)) + (1-h_j(D(t+1)))\log(1-p_j(D(t);\theta)), \tag{F.1}$$

where θ is not necessarily equal to θ_0. Thus, the SLLN (3.7) for the underlying Markov chain in (F.1) implies that

$$\begin{aligned}\frac{1}{T}\ell_T(X;\theta) \quad &\xrightarrow{a.s.} \quad \sum_{j=1}^{J}\mathbb{E}_\pi\Big(p_j(D^*;\theta_*)\log(p_j(D^*;\theta)) \\ &+ \quad (1-p_j(D^*;\theta_*))\log(1-p_j(D^*;\theta))\Big), \qquad (F.2)\\ &=: \quad \sum_{j=1}^{J}\mathbb{E}_\pi H^-(p_j(D^*;\theta_*),p_j(D^*;\theta)) \equiv \ell_\infty(\theta_*;\theta),\end{aligned}$$

where $H^-(p_0, p) := (p_0 \log(p) + (1-p_0)\log(1-p))$ denotes the negative *cross–entropy* of the two Bernoulli distributions with parameters p_0 and p, respectively. The information inequality implies that

$$H^-(p_0, p) \leq H^-(p_0, p_0),$$

and the inequality is strict unless $p = p_0$. This argument shows that the log–likelihood converges, as $T \to \infty$ to the deterministic function $\theta \mapsto \ell_\infty(\theta_*; \theta)$, which has a maximum at $\theta = \theta_*$. This maximum is unique since the model is identifiable. Also see Lemma III.3 for a criterion of identifiability.

By lemma III.1 we have, $|\frac{1}{T}\frac{\partial \ell_T}{\partial w_{ij}}| \leq 2K$ since g is bounded by K and $|\frac{1}{T}\frac{\partial \ell_T}{\partial b_j}| \leq 2$. By a Taylor series expansion, we have

$$\frac{1}{T}(\ell_T(\theta) - \ell_T(\theta')) = \frac{1}{T}\Big\langle \nabla \ell_T(\tilde{\theta}), (\theta' - \theta) \Big\rangle$$

where $\tilde{\theta} = \theta + \lambda(\theta' - \theta)$ with $\lambda \in [0, 1]$.

Since $\left(\frac{1}{T}\frac{\partial \ell_T}{\partial \theta_k}\right)$ are almost surely bounded by a constant that does not depend on θ, we have

$$\Big|\frac{1}{T}\Big(\ell_T(\theta) - \ell_T(\theta')\Big)\Big| \leq C||\theta - \theta'||, \quad \theta, \theta' \in \Theta$$

For all $\epsilon > 0$ let us define a ϵ- mesh: $\alpha(n) = \{\alpha_{i(\theta)}(\epsilon)\}_{i=1,\ldots,M(\epsilon)}$ such that a compact set, $F \subset \bigcup_{i=1}^{M(\epsilon)} B(\theta_i(\epsilon), \epsilon)$. Thus we get,

$$\begin{aligned}
\sup_{\theta \in F}\Big|\frac{1}{T}\sum_{t=1}^{T-1} r(\theta, D(t)) - \ell(\theta)\Big| \quad &\leq \quad \sup_{\theta \in F}\Big\{\frac{1}{T}\sum_{t=1}^{T-1}|r(\theta, D(t)) - r(\alpha_{i(\theta)}, D(t))| \\
&+ \quad \Big|\frac{1}{T}\sum_{t=1}^{T-1} r(\alpha_{i(\theta)}, D(t)) - \ell(\theta)\Big|\Big\}
\end{aligned}$$

where $r(\theta, D(t)) := \sum_{j=1}^{J} h_j(D(t+1)) \log(p_j(D(t);\theta)) +$
$(1 - h_j(D(t+1))) \log(1 - p_j(D(t);\theta))$.

$$\begin{aligned}
\sup_{\theta\in F}\Big|\frac{1}{T}\sum_{t=1}^{T-1} r(\theta, D(t)) - \ell(\theta)\Big| &\leq C\epsilon + \sup_{\theta\in F}\Big|\frac{1}{T}\sum_{t=1}^{T-1} r(\alpha_{i(\theta)}, D(t)) - \ell(\theta)\Big| \pm \ell(\alpha_{i(\theta)}) \\
&\leq C\epsilon + \sup_{\theta\in F}\Big|\frac{1}{T}\sum_{t=1}^{T-1} r(\alpha_{i(\theta)}, D(t)) - \ell(\alpha_{i(\theta)})\Big| \\
&+ \sup_{\theta\in F}\Big|\ell(\theta) - \ell(\alpha_{i(\theta)})\Big| \\
&\leq C\epsilon + \max_{j=1,\ldots,M(\epsilon)}\Big|\frac{1}{T}\sum_{t=1}^{T-1} r(\alpha_{i(\theta)}, D(t)) - \ell(\alpha_{i(\theta)})\Big| \\
&+ \sup_{\theta\in F}\Big|\ell(\theta) - \ell(\alpha_{i(\theta)})\Big| \\
&\leq C\epsilon + \sup_{\theta\in F}\Big|\ell(\theta) - \ell(\alpha_{i(\theta)})\Big|
\end{aligned}$$

Thus, using the fact that ℓ is uniformly continuous we prove that $T^{-1}\ell_T(\theta) \xrightarrow{\text{a.s.}} \ell(\theta)$, on all compact subsets of the parameter space Θ and thus completes the proof of Property (i). □

Proposition F.2. *Condition (ii) of Theorem III.4 holds under the model (3.1). Moreover, if the model is identifiable then $\mathfrak{I}(\theta_*)$ is invertible for every $\theta_* \in \Theta$.*

Proof. To prove Property (ii) we show that $\nabla_\theta \ell_T(\theta)$ is a Martingale and apply the Martingale CLT proved in Brown (1971). By the Wold's Device (Theorem 29.4 in Billingsley (1995)) the multivariate limit in (3.10) holds provided that for all $a \in \mathbb{R}^d$ ($d = J^2 + J + 1$) we have:

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{d} a_i \partial l/\partial\theta_i \to \mathcal{N}(0, a^T \mathfrak{I}(\theta_*) a), \quad \text{as } T \to \infty. \tag{F.3}$$

For any fixed $a \in \mathbb{R}^d$, we will show next that (F.3) holds by using the martingale CLT. For notational convenience we treat the parameters b_i and w_{ij} separately. Namely,

we let $\beta_i = a_i,\ i = 1, ..., J$ and $\alpha_{ij} = a_{J+i+(j-1)J},\ i, j = 1, ..., J$. Then the L.H.S. of (F.3) equals,

$$\frac{1}{\sqrt{T}} N_T = \frac{1}{\sqrt{T}} \sum_{j=1}^{J} \beta_j \partial l / \partial b_j + \frac{1}{\sqrt{T}} \sum_{i,j=1}^{J} \alpha_{ij} \partial l / \partial w_{ij}$$

By Lemma III.1 we have,

$$\begin{aligned} N_T &= \sum_{t=1}^{T-1} \sum_{j=1}^{J} [X_j(t+1) - \mathbb{P}_j(t+1|\mathcal{H}_t)] \Big(\beta_j + \sum_{i,j=1}^{J} \alpha_{ij} g_{ij}(D(t)) \Big) \qquad \text{(F.4)} \\ &= \sum_{t=1}^{T-1} \zeta_t(\alpha, \beta) \end{aligned}$$

Observe that $\mathbb{E}(\zeta_t(\alpha, \beta)|\mathcal{H}_t) = 0$ since $\mathbb{P}_j(t+1|\mathcal{H}_t) = \mathbb{E}(X_j(t+1)|\mathcal{H}_t)$. Therefore, $\zeta_t(\alpha, \beta)$ is a martingale difference w.r.t. the filtration $\mathcal{H}_t, t = 1, 2, ...$, i.e. $\{N_T, T = 1, ...\}$ is a martingale w.r.t. the same filtration. This allows us to apply the martingale CLT to $\frac{1}{\sqrt{T}} N_T$. Following(3.7), let σ_t^2 be the conditional variance of $\zeta_t(\alpha, \beta)$. By Equation (F.5), we have $X_j(t+1)|\mathcal{H}_t$ is conditionally Bernoulli($p_j(D(t))$ and we get

$$\sigma_t^2 = \mathbb{E}(\zeta_t^2(\alpha, \beta)|\mathcal{H}_t) = \sum_{j=1}^{J} \Big(\beta_j + \sum_{i=1}^{J} \alpha_{ij} g_{ij}(D(t)) \Big)^2 p_j(D(t))(1 - p_j(D(t))) \qquad \text{(F.5)}$$

Now, let us define,

$$\begin{aligned} V_T^2 &= \sum_{t=1}^{T} \sigma_t^2 \qquad \text{(F.6)} \\ s_T^2 &= \mathbb{E}[V_T^2] = \mathbb{E}[N_T^2] = \sum_{t=1}^{T} \mathbb{E}[\sigma_t^2] \\ &= \sum_{t=1}^{T} \mathbb{E}\Big[\sum_{j=1}^{J} \Big(\beta_j + \sum_{i=1}^{J} \alpha_{ij} g_{ij}(D(t)) \Big)^2 p_j(D(t))(1 - p_j(D(t))) \Big] \end{aligned}$$

For this class of martingales, the Lindeberg condition is said to hold if

$$s_T^2 \sum_{t=1}^{T} \mathbb{E}[\zeta_t^2 I(|\zeta_t| \geq \epsilon s_T)] \to 0 \qquad \text{as } \ t \to \infty \tag{F.7}$$

Theorem F.3 (Martingale CLT)**.** *If* $V_T^2 s_T^{-2} \ \to_p \ 1 \quad$ *as* $n \to \infty$ *and if the Lindeberg condition hold, then*

$$\lim_{n\to\infty} \mathbb{P}[N_T/s_T \leq x] = \Phi(x) = (2\pi)^{(-1/2)} \int_{-\infty}^{x} \exp\{-\frac{1}{2}y^2\} dy$$

for all x*.*

Using the strong law of large numbers (SLLN) of Markov chains (3.7) we can verify the conditions in Theorem F.3 and also the Lindeberg condition. Indeed, by (3.7) and the Lebesgue Dominated Convergence Theorem, since $|g_i^2| \leq K^2$, it follows that

$$\frac{1}{T} V_T^2 = \frac{1}{T} \sum_{t=1}^{T} h_{\sigma,\alpha,\beta}(D(t)) \xrightarrow{a.s.} \mathbb{E}_\pi[h_{\sigma,\alpha,\beta}(D^*)] \text{ as } T \to \infty \tag{F.8}$$

where $h_{\sigma,\alpha,\beta}(D^*) = \sum_{j=1}^{J} (\beta_j + \sum_{i=1}^{J} \alpha_{ij} g_{ij}(D^*))^2 p_j(D^*)(1 - p_j(D^*))$ and D^* has the stationary distribution of the Markov Chain.

The SLLN applies to the function $h_{\sigma,\alpha,\beta}$ since the g_{ij}'s and p_j's are bounded and the α_{ij}'s, β_j's are some fixed constants. The fact that $|h_{\sigma,\alpha,\beta}| \leq$ Constant and the Lebesgue DCT also imply that

$$\frac{1}{T} \mathbb{E}[V_T^2] = \frac{1}{T} s_T^2 \longrightarrow \mathbb{E}[h_{\sigma,\alpha,\beta}(D^*)] \tag{F.9}$$

In view of (F.5), $|\zeta_t|$ is uniformly bounded by a constant. Since by (F.9), $s_T^2 \sim T, \ t \to \infty$, the Lindeberg condition trivially holds. Thus, the martingale C.L.T.

implies that,

$$\frac{1}{\sqrt{T}}N_T \sim \frac{N_T}{s_T}\sqrt{\mathbb{E}(h_{(\sigma,\alpha,\beta)}(D^*))} \longrightarrow \mathcal{N}(0, E[h_{\sigma,\alpha,\beta}(D^*)]) \text{ as } T \to \infty \tag{F.10}$$

It can be shown that,

$$\begin{aligned} E[h_{\sigma,\alpha,\beta}(D^*)] &= E\Big[\sum_{j=1}^{J}\Big(\beta_j + \sum_{i=1}^{J}\alpha_{ij}g_{ij}(D^*)\Big)^2 p_j(D^*)(1-p_j(D^*))\Big] \\ &= v^T\mathfrak{I}(\theta_*)v \end{aligned} \tag{F.11}$$

where $\vec{v} = (\beta_1, \cdots, \beta_J, \alpha_{11}, \alpha_{12}, \cdots, \alpha_{1J}, \cdots, \alpha_{JJ})$ and $\mathfrak{I}(\theta_*)$ is the Fisher information matrix □

Proposition F.4. *Condition (iii) of Theorem III.4 holds under the model (3.1).*

Proof. We shall give detailed proof of Property (iii) only on the case when θ_i and θ_j corresponding to some connectivity strength coefficients. With some misuse of notation we will denote these coefficients as $w_{i'j'}$ and w_{ij}. We only need to consider the case $j = j'$ by Lemma III.1. By applying the SLLN (recall (3.7)) to the underlying Markov chain in (G.2). By Equation (G.2) we have,

$$\Big|\frac{1}{T}\frac{\partial^2\ell_T}{\partial w_{i'j}\partial w_{ij}}(\theta) - \frac{1}{T}\frac{\partial^2\ell_T}{\partial w_{i'j}\partial w_{ij}}(\theta')\Big| \leq \frac{1}{T}\sum_{t=1}^{T-1}\Big|f(\theta, D(t)) - f(\theta', D(t))\Big| \tag{F.12}$$

where $f(\theta, D(t)) = p_j(D(t);\theta)(1-p_j(D(t);\theta))g_{ij}(D(t))g_{i'j}(D(t))$ and for simplicity we omit the dependence on the indices i, i' and j which are fixed. By the intermediate value theorem we get,

$$\Big|f(\theta, D(t)) - f(\theta', D(t))\Big| = \Big|\nabla f(\tilde{\theta})^T(\theta' - \theta)\Big| \tag{F.13}$$

where $\tilde{\theta} = \theta + \lambda(\theta' - \theta)$, $\lambda \in [0,1]$ using direct differentiation and by considering all

cases one can show that $\|\nabla f(\tilde{\theta})^T\|$ is bounded by a constant independent of D, θ', θ and λ. For the exact expression see Lemma G.1 in the Appendix. Then, (F.13) is bounded above as follows:

$$\Big|f(\theta, D(t)) - f(\theta', D(t))\Big| \leq C'||\theta - \theta'||, \quad \theta, \theta' \in \Theta \tag{F.14}$$

where C' is a constant independent of θ and $D(t)$.

Now, for the supremum on the L.H.S. of (3.11) by (F.14) and the Δ - inequality we obtain,

$$\begin{aligned}
\sup_{|\theta-\theta_*|\leq\delta} \quad & \Big|\frac{1}{T}\sum_{t=1}^{T-1}\Big\{f(\theta, D(t)) \pm f(\theta', D(t))\Big\} + \mathfrak{I}_{ij}(\theta_*)\Big| \\
\leq \quad & \sup_{|\theta-\theta_*|\leq\delta} C'||\theta - \theta_*|| + \Big|\frac{1}{T}\sum_{t=1}^{T-1} f(\theta_*, D(t)) - \mathfrak{I}_{ij}(\theta_*)\Big|, \quad \theta, \theta_* \in \Theta \\
\leq \quad & C\delta + \Big|\frac{1}{T}\sum_{t=1}^{T-1} f(\theta_*, D(t)) - \mathfrak{I}_{ij}(\theta_*)\Big|, \quad \theta, \theta' \in \Theta
\end{aligned} \tag{F.15}$$

The first term on the R.H.S. of (F.15) is less than $C'\delta$. The second term therein vanishes a.s. as $T \to \infty$ by the S.L.L.N (recall (3.7)). Indeed by (F.11) one can see that the element of the information matrix $\mathfrak{I}(\theta_*)$ corresponding to the coefficients w_{ij} and $w_{i'j}$ equals $\mathbb{E}_\pi\Big[p_j(D^*;\theta)(1 - p_j(D^*;\theta))g_{ij}(D^*)g_{i'j}(D^*)\Big]$. By the SLLN applied to Equation G.2 we have that the second term on the R.H.S. of (F.15) vanished as $T \to \infty$, which completes the proof of (iii). □

APPENDIX G

Proof of Lemma G.1

Lemma G.1. *The log-likelihood function ℓ is infinitely continuously differentiable with respect to all of its parameters $w_{ij} \in \mathbb{R}$ and $b_j \in \mathbb{R}$, $i, j = 1, 2, \cdots, J$. Moreover,*

$$\begin{aligned}
\frac{\partial^2 \ell_T}{\partial w_{i'j'} \partial w_{ij}} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1 - p_j(D(t);\theta)) g_{ij}(D(t)) g_{i'j}(D(t)) \Big) \\
&\times \quad \delta(j - j') \\
\frac{\partial^3 \ell_T}{\partial w_{i''j''} \partial w_{i'j'} \partial w_{ij}} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1 - p_j(D(t);\theta)) \\
&\times \quad (1 - 2p_j(D(t);\theta)) g_{ij}(D(t)) \Big) g_{i'j}(D(t)) g_{i''j}(D(t)) \times \\
&\times \quad \delta(j - j') \delta(j - j'')
\end{aligned} \tag{G.1}$$

$$
\begin{aligned}
\frac{\partial^2 \ell_T}{\partial w_{i'j'} \partial b_j} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta)) g_{i'j}(D(t)) \Big) \\
&\times \quad \delta(j-j') \\
\frac{\partial^2 \ell_T}{\partial b_{j'} \partial b_j} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta)) \Big) \times \delta(j-j') \\
\frac{\partial^3 \ell_T}{\partial w_{i''j''} \partial b_{j'} \partial b_j} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta))(1-2p_j(D(t);\theta)) \Big) \\
&\times \quad g(h_{i''j}(D(t))) \times \delta(j-j') \times \delta(j-j'') \qquad \text{(G.2)} \\
\frac{\partial^3 \ell_T}{\partial w_{i''j''} \partial w_{i'j'} \partial b_j} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta))(1-2p_j(D(t);\theta)) \Big) \\
&\times \quad g_{i''j}(D(t)) g_{ij}(D(t)) \times \delta(j-j')\delta(j-j'') \\
\frac{\partial^3 \ell_T}{\partial b_{j''} \partial b_{j'} \partial b_j} &:= -\Big(\sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta))(1-2p_j(D(t);\theta)) \Big) \\
&\times \quad \delta(j-j')\delta(j-j'')
\end{aligned}
$$

where $g(h_{ij}(D(t))) = \sum_{s=\tau_j(t)}^{t} \lambda^{t-s} X_i(s)$ *and* δ *stands for the kronecker symbol.* □

Proof. As in the proof of lemma III.1 we have that the L.H.S. in (G.2) and (G.2) is zero unless $j = j' = j''$. We therefore set $j = j' = j''$ and by differentiating (3.4) we get,

$$
\begin{aligned}
\frac{\partial^2 \ell_T}{\partial w_{i'j} \partial w_{ij}} &= \sum_{t=1}^{T-1} \frac{\partial p_j(t+1|\mathcal{H}_t)}{\partial w_{i'j}} g(h_{ij}(D(t))) \\
&= \sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta)) g_{ij}(D(t)) g_{i'j}(D(t))
\end{aligned}
$$

Similarly,

$$\begin{aligned}\frac{\partial^3 \ell_T}{\partial w_{i''j}\partial w_{i'j}\partial w_{ij}} &= \sum_{t=1}^{T-1} \frac{\partial p_j(D(t);\theta)(1-p_j(D(t);\theta))}{\partial w_{i''j}} g_{ij}(D(t))g_{i'j}(D(t)) \\ &:= \sum_{t=1}^{T-1} p_j(D(t);\theta)(1-p_j(D(t);\theta))(1-2p_j(D(t);\theta)) \\ &\times \; g_{i'j}(D(t))g_{i''j}(D(t))\end{aligned}$$

Note that only the term corresponding to $k = j$ remains from the inner summation because of Equation (E.3). We can similarly derive the expression for $\partial\ell/\partial b_j$, $\partial^2\ell_T/\partial b_{j'}\partial b_j$, $\partial^3\ell_T/\partial b_{j''}\partial b_{j'}\partial b_j$ and $\partial^3\ell_T/\partial w_{i''j''}\partial b_{j'}\partial b_j$ and the derivation is in fact simpler. Since the parameter b_j in Equation (3.1) appears with a coefficient equal to 1. □

BIBLIOGRAPHY

BIBLIOGRAPHY

M. Abeles and I. Gat. Detecting precise firing sequences in experimental data. *Journal of Neuroscience Methods*, 107:141–154, 2001.

M. Abeles and G. L. Gerstein. Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. *Journal of Neurophysiology*, 60, 1988.

A. Agresti. Categorical Data Analysis. *Probability and Mathematical Statistics, Wiley, New York*, 1990.

P. Billingsley. Probability and Measure. *John Wiley and Sons, New York*, 1995.

D. Brillinger. Nerve cell spike train data analysis: a progression of technique. *Journal of the American Statistical Association*, 87:260–271, 1992.

D. R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. *Biological Cybernetics*, 59:189–200, 1988.

D. R. Brillinger and E. Villa. Assessing connections in networks of biological neurons. *The Practice of Data Analysis: Essays in Honor of John W. Tukey*, pages 77–92, 1997.

B. M. Brown. Martingale Central Limit Theorems. *The Annals of Mathematical Statistics*, 42, 1971.

E. M. Brown, R. E. Kass, and P. M. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7, 2004.

O. Cappé, E. Moulines, and T. Rydén. Inference in Hidden Markov Models. *Springer Series in Statistics*, 2005.

E. Chornoboy, L. Schramm, and A. Karr. Maximum likelihood identification of neuronal point process systems. *Biological Cybernetics*, 59:265–275, 1988.

N. Cressie. Statistics for Spatial Data. *J. Wiley and Sons, Inc*, 1993.

C. O. Diekman, P. S. Sastry, and K. P. Unnikrishnan. Statistical significance of sequential firing patterns in multi-neuronal spike trains. *Journal of Neuroscience Methods*, 182:279–284, 2009.

C. O. Diekman, K. Dasgupta, V. N. Nair, and K. P. Unnikrishnan. Detecting Neuronal Connectivity from Serial Patterns in Spike Train Data. *In preparation*, 2012.

M. S. Fee, P. P. Mitra, and D. Kleinfeld. Automatic sorting of multiple unit neuronal signals in the presence of anisotrophic and non-Gaussian variability. *Journal of Neuroscience Methods*, 69:175–188, 1996.

G. L. Gerstein. Searching for significance in spatio-temporal firing patterns. *ACTA Neurobiologiae Experimentalis*, 64, 2004.

P. J. Green. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46, 1984.

S. Grun, M. Diesmann, and A. Aertsen. Unitary events in multiple single-neuron spiking activity: I. Detection and significance. *Neural Computation*, 14:43–80, 2001.

K. D. Harris, D. A. Henze, J. Csicsvari, H. Hirase, and G. Buzsaki. Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *Journal of Neurophysiology*, 84:401–414, 2000.

M. I. Jordan. Graphical Models. *Statistical Science*, 19, 2004.

R. E. Kass and V. Ventura. A spike train probability model. *Neural Computation*, 13:1713–1720, 2001.

R. E. Kass, V. Ventura, and E. M. Brown. Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology*, 94:8–25, 2005.

S. Koyama and R. E. Kass. Spike Train Probability Models for Stimulus-Driven Leaky Integrate-and-Fire Neurons. *Neural Computation*, 20:1776–1795, 2008.

L. Lauritzen. Graphical Models. *Clarendon Press, Oxford*, 1996.

S. Laxman, P. S. Sastry, and K. P. Unnikrishnan. Discovering frequent episodes and learning hidden markov models: A formal connection. *IEEE Transactions on Knowledge and Engineering*, 17:1505–1517, 2005.

M. S. Lewicki. A review of methods for spike sorting: the detection and classification of neural action potentials. *Network: Computation In Neural Systems*, 9:53–78, 1988.

H. Mannila, H. Toivonen, and A. Verkamo. Discovery of frequent episodes in event sequences. *In Data Mining and Knowledge Discovery*, 1:259–289, 1997.

L. Martignon, G. Deco, K. Laskey, M. Diamond, W. Freiwald, and E. Vaadia. A method for simulating non-normal distributions. *Psychometrika*, 12:2621–2653, 2000.

S. P. Meyn and R. L. Tweedie. Markov Chains and Stochastic Stability. *Springer-Verlag*, 2005.

A. Y. Ng, M. Jordan, and Y. Weiss. On Spectral Clustering Analysis and An Algorithm. *Advances in Neural Information Processing Systems (NIPS)*, 14, 2002.

M. Okatan, M. Wilson, and E. M. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Computation*, 17:1927–1961, 2005.

C. R. Olson, S. N. Gettner, V. Ventura, R. Carta, and R. E. Kass. Neuronal activity in macaque supplementary eye field during planning of saccades in response to pattern and spatial cues. *Journal of Neurophysiology*, 84:1369–1384, 2000.

L. Paninski. Inferring synaptic inputs given a noisy voltage trace via sequential Monte Carlo methods. *Journal of Computational Neuroscience*, 2011.

D. Patnaik, P. S. Sastry, and K. P. Unnikrishnan. Inferring neuronal network connectivity from spike data: A temporal datamining approach. *Scientific Programming*, 16, 2008.

G. Pipa, D. W. Wheeler, W. Singer, and D. Nikolic. NeuroXidence: reliable and efficient analysis of an excess or deficiency of joint spike events. *Journal of Computational Neuroscience*, 25:64–88, 2008.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, 38, 2010.

F. Rigat, M. Gunst, and J. Pelt. Bayesian Modelling and Analysis of Spatio-Temporal Neuronal Networks. *Bayesian Analysis*, 1:733–764, 2006.

J. Rolston, D. A. Wagenaar, and S. M. Potter. Precisely timed spatiotemporal patterns of neural activity in dissociated cell cultures. *Neuroscience*, 148:294–303, 2007.

P. S. Sastry and K. Unnikrishnan. Conditional probability based significance tests for sequential patterns in multi-neuronal spike trains. *Neural Computation*, 22: 1025–1059, 2010.

M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *Journal of Neuroscience Methods*, 18:3870–3896, 1998.

Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.

I. V. Tetko and A. E. Villa. A pattern grouping algorithm for analysis of spatiotemporal patterns in neuronal spike trains: 1. Detection of repeated patterns. *Journal of Neuroscience Methods*, 105:1–14, 2001.

W. Truccolo, U. Eden, M. Fellows, J. Donoghue, and E. M. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. *Journal of Neurophysiology*, 93:1074–1089, 2005.

R. Viswanathan, P. S. Sastry, and K. P. Unnikrishnan. Efficient Discovery of Large Synchronous Events in Neural Spike Streams. *arXiv:1006.1543*, 2010.

D. A. Wagenaar, J. Pine, and S. M. Potter. An extremely rich repertoire of bursting patterns during the development of cortical cultures. *BMC Neuroscience*, 7:294–303, 2006.

J. Whittaker. Graphical Models in Applied Multivariate Statistics. *Wiley Series in Probability and Statistics*, 2009.