A Statistical Approach to Functional Connectivity Involving Multichannel Neural Spike Trains

Ruiwen Zhang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research (Statistics).

Chapel Hill
2011

Approved by:
Advisor: Young K. Truong
Advisor: Haipeng Shen
Committee Member: Chuanshu Ji
Committee Member: Jan Hannig
Committee Member: Sidney A. Simon
Abstract

(Under the direction of Young K. Truong and Haipeng Shen.)

The advent of the multi-electrode has made it feasible to record spike trains simultaneously from several neurons. However, the statistical techniques for analyzing large-scale simultaneously recorded spike train data have not developed as satisfactorily as the experimental techniques for obtaining these data. This dissertation contributes to the literature of modeling simultaneous spike train data and inferring the functional connectivity in two aspects.

In the first part, we apply a point process likelihood method under the generalized linear model framework (Harris, 2003) for analyzing ensemble spiking activity from noncholinergic basal forebrain neurons (Lin and Nicolelis, 2008). The model can assess the correlation between a target neuron and its peers. The correlation is referred to as weight for each peer and is estimated through maximizing the penalized likelihood function. A discrete time representation is used to construct the point process likelihood, and the discrete 0-1 occurrence data are smoothed using Gaussian kernels. Ultimately, the entire peer firing information and the correlations can be used to predict the probability of target firing.

In the second part, we propose a regression spline model, which directly makes use of the neural firing times instead of using the smoothed version of spike train. The primary contribution of the model is that it can both capture the spontaneous dynamics and also infer functional connectivity for an arbitrary number of interactive
neurons in a given region or across different regions. In addition, it does not need
discretization, relaxes the parametric assumption, and offers high flexibility for esti-
mation via spline functions. The regression spline model selects the optimal spline
knots adaptively using the spike train data. Our model incorporates adaptive model
selection and is estimated through maximum likelihood. Asymptotic properties of
the proposed estimator are investigated as well.
Acknowledgments

I am largely indebted to Prof. Young Truong and Prof. Haipeng Shen, who have been such great advisors and mentors to me. Their ingenious ideas and deep insight have constantly inspired me and guided me through the past five years. And their knowledge of statistics has been a great source of enlightenment to me. I am especially gratefully for their great effort on my research topics and process.

My sincere gratitude also goes to other members of my committee. Prof. Sidney Simon from the Neurobiology department at Duke University has provided enormous support of my research. He interpreted the intricacy of neurobiological issues with patience especially as my background is in Statistics. His profound knowledge and valuable advice enlightened me throughout my research. Prof. Jan Hannig and Prof. Chuanshu Ji have read my dissertation carefully and provided numerous useful comments and feedbacks. I also learned a lot from the interesting courses they taught.

I cannot show enough appreciation to my husband. He always listens to me when I am frustrated, helping me out with his love at each and every moment. I also owe my deepest gratitude to my parents, who always love me, believe in me, and support me no matter what. I hope that I will make them proud by my accomplishment.
## Contents

Abstract ........................................................................................................ iii

1 Introduction .................................................................................................. 1

2 Background and Basic Concepts ................................................................. 6

  2.1 Point Process and Conditional Intensity Function ................................. 6

  2.2 The Likelihood Function of a Point Process Model ................................. 12

  2.3 Time-rescaling Theorem ....................................................................... 13

  2.4 Neural Synchrony ................................................................................. 15

3 Literature Review ......................................................................................... 17

  3.1 Parametric Models .............................................................................. 18

    3.1.1 A Special Interspike Interval (ISI) Probability Model .................. 18

    3.1.2 Brillinger’s GLIM Approach for Interacting Neurons ............... 23

    3.1.3 Relating Neural Spiking Activities to Covariate Effects
         via the Generalized Linear Model (GLM) Framework ............... 24

  3.2 Nonparametric Models ....................................................................... 26

    3.2.1 Kernel Smoothing ..................................................................... 26

    3.2.2 Adaptive Kernel Smoothing ...................................................... 29

    3.2.3 Kernel Bandwidth Optimization ............................................... 31

    3.2.4 Smoothing Splines .................................................................. 32

4 The Methods ............................................................................................... 35

  4.1 Continuous State-Space Model .............................................................. 35
4.2 Regression Spline Model ........................................... 37
   4.2.1 Model .......................................................... 37
   4.2.2 Maximum Likelihood Estimation ......................... 40
   4.2.3 Adaptive Model Selection ................................. 42

5 Numerical Studies ...................................................... 44
   5.1 Continuous State-Space Model ................................. 44
      5.1.1 Simulation ................................................. 44
      5.1.2 Real Data Analysis ..................................... 48
   5.2 Regression Spline Model ....................................... 54
      5.2.1 Simulation Study I: A General Algorithm of Neural
           Spike Train Generation .................................. 55
      5.2.2 Simulation Study II: Fixed-Knot Splines ............... 57
      5.2.3 Simulation Study III: Adaptive Knot Selection ....... 63
      5.2.4 Simulation Study IV: Predictability Comparison with
           Brillinger’s GLIM Approach .............................. 65
      5.2.5 Real Application I: Feature Extraction Using the Base-
           line Intensity ................................................ 68
      5.2.6 Real Application II: Linear Function for Peer Effect .. 70
      5.2.7 Real Application III: Quadratic Functions for Multiple
           Firings of Each Peer ....................................... 75

6 Consistency and Asymptotics for Spline Regression Model ...... 83
   6.1 Introduction ..................................................... 83
      6.1.1 Preliminaries ............................................. 83
      6.1.2 Polynomial Splines ...................................... 85
   6.2 Spline Approximation .......................................... 85
   6.3 Maximum Likelihood Estimation ............................... 92
   6.4 Asymptotic Distributions of the Estimates ................... 107

Bibliography ............................................................. 123
Chapter 1

Introduction

It is known that neurons, even when they are apart in the brain, often exhibit correlated firing patterns (Varela et al., 2001). For instance, coordinated interaction among cortical neurons is known to play an indispensable role in mediating many complex brain functions with highly intricate network structures (Yoshimura and Callaway, 2005). A procedure to examine the underlying connectivity between neurons can be stated in the following way (Chapter 5, Oweiss, 2010). For a neuron $i$ in a population of $N$ observed neurons, we need to identify a subset $\pi_i$ of neurons that affect the firing of neuron $i$ in some statistical sense.

In the study of neural plasticity and network structure, it is desirable to infer the underlying functional connectivity between the recorded neurons. In the analysis of neural spike trains, functional connectivity is defined in terms of the statistical dependence observed between the spike trains (Friston, 1994) from distributed and often spatially remote neuronal units. This can result from the presence of a synaptic link between neurons, or it can be observed when two unlinked neurons respond to a common driving input.

Characterization of the functional network requires simultaneous monitoring of neural constituents while subjects carry out certain functions. Technology developments in multi-electrode recording enable us to easily obtain the activities of ensembles of spiking neurons simultaneously. Easier access to data has underlined the need
for developing analysis methods that can process these data quickly and efficiently (Hatsopoulos et al., 1998; Nicolelis et al. 2003; Harris et al., 2003; Truccolo et al., 2005; Brown et al., 2004). In this dissertation, we apply and develop statistical approaches to analyze simultaneously recorded neural spike trains and infer functional connectivity between neurons that act in concert in a given brain region or across different regions.

In this dissertation, we contribute to the literature of modeling neural spike trains and inferring the functional connectivity in two aspects:

- We apply a point process likelihood method under the generalized linear model framework to analyze the ensemble activities from noncholinergic basal forebrain neurons (Lin and Nicolelis, 2008), which have never been studied under this scope before. The model constructs the rate function by smoothing the discrete ‘on/off’ data into continuous state space. It is referred to as the continuous state-space model in this thesis.

- We endeavor to model the occurrence data ‘directly’ without transformation and so consider the time to event instead. We propose a regression spline model for estimating the conditional firing intensity, which captures both the spontaneous dynamics and the time-varying peer effects, and so eventually yields the functional connectivity in the network.

Early methods of inferring connectivity usually focus on analyzing pairs of neurons using time series techniques (e.g., cross-correlograms or joint peri-stimulus time histograms) or frequency domain methods (e.g., cross-coherence) (Brown et. al, 2004), but pairwise methods generally provide only an incomplete picture of the connections among several neurons. More recent methods, such as gravitational clustering and spike pattern classification methods (Stevenson, 2009), have made steps towards estimating functional connectivity that is greater than pairwise, but they are still
mostly suitable for pairs or triplets. When large-scale recording of multiple single units became more common, using these methods to infer a complex dynamic network structure became nontrivial.

Now we describe in detail our first contribution. Likelihood methods under the generalized linear model (GLM) framework are increasingly popular for analyzing neural ensembles (Hatsopoulos et al., 1998; Harris et al., 2003; Stein et al., 2004; Paninski, 2004; Truccolo et al., 2005; Santhanam et al., 2006; Luczak et al., 2007; Pillow et al., 2008). We applied a point process likelihood model under the GLM framework (Harris, 2003) for some real applications. The spike train data come from noncholinergic basal forebrain neurons (Lin and Nicolelis, 2008), which have never been analyzed under the point process likelihood framework. Harris’ method approaches the problem similarly in the spirit of Brown et. al (1998), but it models the intensity function as the sum of weighted Gaussian kernels and adds a penalty term in the log likelihood. We will discuss more details in Section 4.1.

Our numerical studies suggest the following observations: First, the method estimates the correlations (which are referred to as weights) between a target neuron and its peers. Second, once we estimate the weights for each individual peer in a network, we can ultimately predict the firing of the target neuron.

This continuous state-space model does have some shortcomings, which motivate us to develop an approach based on the point process in the second part of this dissertation. As we mentioned earlier, the continuous state-space approach convolves the spike train and transforms it to a continuous process. Also, a discrete time representation is used to construct the corresponding likelihood, which limits the applicability of the method to some extent. Below we describe an alternative approach to modeling the spike trains, which directly makes use of the neuron firing times.

Brillinger (1988) formulates an approach for analyzing interacting nerve cells based
on interspike intervals (ISIs), which extends Brillinger and Segundo (1979) to the case of several spike train inputs. The approach articulates the spontaneous firing rate and interacting effects from arbitrary numbers of neurons. Brillinger’s approach also uses the discrete time likelihood, and the implementation begins with cutting the time into consecutive, non-overlapping small intervals. We will discuss more details about Brillinger’s method and provide a complete comparison with our proposed model later in Section 5.2.4.

We propose a regression spline model to extract information from an ensemble of neurons. The model can be used to describe the variation of a neuron’s spontaneous firing rate from the history of its own, and it can also assess time-varying correlations between interactive neurons in network. Our regression spline model inherits the flexibility for analyzing arbitrary number of interactive neurons. In addition, it does not need discretization, relaxes the parametric assumption, and offers high flexibility for estimation via spline functions. In the model, cubic splines are used to approximate the spontaneous firing intensity of a target neuron, while polynomial functions model the influences of peer activities. The regression spline model selects the optimal spline knots adaptively using the spike train data. Our model incorporates adaptive model selection and is estimated through maximum likelihood.

The model specifications will be described in detail in Section 4.2, and the maximum likelihood estimation will be discussed thoroughly in Section 4.2.2. We employ an adaptive model selection for allocating the knots in a manner similar to HEFT and HARE as described in Kooperberg, Stone and Truong (1995); see details in Section 4.2.3. Our method is numerically evaluated using simulation studies in Section 5.2.2 and 5.2.3. In the real applications, we analyzed a real data set from noncholinergic basal forebrain neurons (Lin and Nicolelis, 2008). We will use special cases in Section 5.2.6 and 5.2.7 to present how the model can be used to describe both
the spontaneous features of the individual neurons and also the functional connectivity. Our implementation of the model is written in C and is now readily available with an R interface. We also investigate the asymptotic properties of our method in Chapter 6. Our results show the $L_\infty$ rates of convergence and the asymptotic normality of the estimates of the conditional intensity function.

The rest of the dissertation is organized as follows. In Chapter 2, we introduce some background concepts and theorems about point process and the likelihood of point process models. Chapter 3 is the literature review of some existing parametric or nonparametric methods. Then we mainly study two models in Chapter 4, a point process-GLM likelihood method and our new proposed model, a regression spline model. Both the models have taken the ensemble firing influences into account. Chapter 5 presents the simulation studies and numerical applications for the two models.
Chapter 2
Background and Basic Concepts

2.1 Point Process and Conditional Intensity Function

A Point Process may be specified in terms of Spike Times, Spike Counts, and Interspike Intervals.

Let \((0, T]\) denote the observation interval, and suppose that a neuron fires at times \(\tau_i, i = 1, 2, ..., n\), where \(0 < \tau_1 < \tau_2 < ... < \tau_{n-1} < \tau_n \leq T\) is a set of \(n\) Spike Times. For \(t \in (0, T]\), a spike train is defined as

\[
S(t) = \sum_i \delta(t - \tau_i),
\]

where \(\tau_i\) is the time of the \(i^{th}\) spike and \(\delta(\cdot)\) is known as Dirac delta function which follows

\[
\delta(x) = \begin{cases} 
1 & \text{if } x = 0 \\
0 & \text{otherwise.}
\end{cases}
\]

When the \(\tau_i\) are random, one has a stochastic point process \(\{\tau_i\}\).

For spikes \(\{\tau_i\}\) randomly scattered along a line, the Counting Process \(N(t)\)
gives the number of points observed in the interval \((0, t]\)

\[
N(t) = \# \{ \tau_i \text{ with } 0 < \tau_i \leq t \}
\]  

(2.1)

where \# stands for the number of events (which have occurred up to time \(t\)). The counting process satisfies

\(i\) \(N(t) \geq 0\);

\(ii\) \(N(t)\) is an integer-valued function;

\(iii\) if \(s < t\), then \(N(s) \leq N(t)\);

\(iv\) for \(s < t\), \(N(t) - N(s)\) is the number of events in \((s, t)\).

A counting process is said to have independent increments if the number of events that occur in disjoint time intervals is independent. A counting process is said to have stationary increments if the number of events that occur in any time interval depends only on the length of the interval. Independent increment and stationary increment are strong conditions and rarely hold for neuronal data since spike trains depend on history, which we shall elaborate in the next section.
As we know, a Poisson process is a special case of point process. A counting process \( N(t) \) \( t \geq 0 \) is a homogeneous (stationary) Poisson process with rate \( \lambda \), \( \lambda > 0 \), if

\[(i) \quad N(0) = 0;\]

\[(ii) \quad \text{the process has independent increments;}\]

\[(iii) \quad P(N(t + \Delta) - N(t) = 1) = \lambda \Delta + o(\Delta);\]

\[(iv) \quad P(N(t + \Delta) - N(t) \geq 2) = o(\Delta).\]

where \( \Delta \) is a very small interval and \( o(\Delta) \) refers to all events of order smaller than \( \Delta \), such as two or more events that occur in an arbitrarily small interval.

The probability mass function is

\[
P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \tag{2.2}\]

for \( k = 0, 1, 2, \ldots \). The mean and the variance of the Poisson process are \( \lambda t \).

To construct the \textit{interspike interval (ISI) probability} for the Poisson process, we denote \( U \) as the \textbf{Interspike Interval} (ISI) between two conjoint events (also called the waiting time), and

\[
Pr(N(u) = 0) = Pr(U > u) = e^{-\lambda ut} \tag{2.3}
\]

or

\[
Pr(u \geq U) = 1 - e^{-\lambda u}. \tag{2.4}
\]

Differentiate Equation 2.4, we have the probability density of the interspike interval, which is the exponential density

\[
p(u) = \lambda e^{-\lambda u} \tag{2.5}
\]
We can show that given the assumption that the interspike interval (ISI) probability density of a counting process is exponential, the counting process is a Poisson process.

Once we understand the inter-event probability density, we can consider the spike train as segments of interspike intervals, \( \{u_i\} \), and the *waiting time density* function is the probability density of the time until the occurrence of the \( k^{th} \) event. Denote \( s_k = \sum_{i=1}^{k} u_i \) as the waiting time up to \( k^{th} \) spike. Then using the properties of a Poisson process, we obtain

\[
Pr(t < s_k < t + \Delta) = Pr(N(t) = (k - 1) \bigcap 1 \text{ event in } (t, t + \Delta]) + o(\Delta) \\
= Pr(N(t) = (k - 1)) \lambda \Delta + o(\Delta) \\
= \frac{(\lambda t)^{k-1}}{(k - 1)!} e^{-\lambda t} \lambda \Delta.
\]

(2.6)

Hence, the waiting time density function until the \( k^{th} \) event is

\[
p_k(t) = \lim_{\Delta \to 0} \frac{Pr(N(t) = k - 1) \lambda \Delta + o(\Delta)}{\Delta} = \frac{(\lambda t)^{k-1}}{(k - 1)!} e^{-\lambda t} \lambda,
\]

(2.7)

which also gives

\[
p_k(t) = \frac{\lambda^k t^{k-1}}{\Gamma(k)} e^{-\lambda t}.
\]

(2.8)

The waiting time density is a gamma distribution with parameters \( k \) and \( \lambda \).

The Poisson process is one of the simplest and most commonly used classes of neural spiking models. Poisson processes are characterized by lack of memory, meaning that the probability distribution of spikes at any point is independent of all previous activities. In some cases, especially when spikes are rare compared with the time scale of the intrinsic membrane dynamics or the effect of history has been averaged out by combining multiple spike trains, Poisson processes can accurately describe spiking activity.

However, Poisson processes are rarely realistic for various neural spike train data.
In particular, the biophysical properties of ion channels limit how fast a neuron can recover immediately following an action potential, leading to a refractory period during which the probability of firing another spike is zero immediately afterward and then significantly decreased further after the previous spike. This is perhaps the most basic illustration of history dependence in neural spike trains.

Since most neural systems have a history-dependent structure, it is necessary to define a firing rate function that depends on history. Any point process can be completely characterized by the **Conditional Intensity Function**, \( \lambda(t|H_t) \) (Daley and Vere-Jones, 2003), which is defined as follows,

\[
\lambda(t|H_t) = \lim_{\Delta \to 0} \frac{\Pr(N(t + \Delta) - N(t) = 1|H_t)}{\Delta},
\]

where \( H_t \) is the history of the spike train up to time \( t \), including all the information of the track and number of spikes in \( (0,t] \) and any covariances up to time \( t \) as well. Therefore, \( \lambda(t|H_t)\Delta \) is the probability of a spike in \( (t,t+\Delta) \) when there is history dependence in the spike train.

The conditional intensity function generalizes the definition of the Poisson rate [3,4]. If the point process is an inhomogeneous Poisson process, then \( \lambda(t|H_t) = \lambda(t) \). It follows that \( \lambda(t|H_t)\Delta \) is the probability of a spike in \( [t,t+\Delta) \) when there is a history-dependent spike train.

When we choose \( \Delta \) to be a small time interval, then Equation (2.9) can be rewritten as follows:

\[
\Pr(N(t + \Delta) - N(t) = 1|H_t) \approx \lambda(t|H_t)\Delta.
\]

Equation (2.10) states that the conditional intensity function multiplied by \( \Delta \) gives the probability of a spike event in a small time interval \( \Delta \).

There are two primary ways to characterize a point process. The first is in terms of
the interevent probability model or interspike interval probability model specifically, and the second is the conditional intensity function. Actually, defining one defines the other and vice-versa.

Under the framework of survival analysis, \( \lambda(t|H_t) \) can be defined in terms of the interspike interval (ISI) density at time \( t \), \( p(t|H_t) \), as

\[
\lambda(t|H_t) = \frac{p(t|H_t)}{1 - \int_0^t p(t|H_t)dt}, \quad t \geq 0. \tag{2.11}
\]

The conditional intensity function as completely characterizes the stochastic structure of the spike train. In any time interval \( (t, t + \Delta] \), \( \lambda(t|H_t)\Delta \) defines the probability of a spike given the history up to time \( t \). If the spike train is an inhomogeneous Poisson process, then \( \lambda(t|H_t) = \lambda(t) \) becomes the generalized definition of the Poisson rate.

We can write

\[
\lambda(t|H_t) = -\frac{d}{dt} \left[ \log[1 - \int_0^t p(s|H_s)ds] \right]. \tag{2.12}
\]

By integrating we have

\[
-\int_0^t \lambda(s|H_s)ds = \log \left[ 1 - \int_0^t p(s|H_s)ds \right]. \tag{2.13}
\]

Finally, exponentiating yields

\[
\exp \left\{ -\int_0^t \lambda(s|H_s)ds \right\} = 1 - \int_0^t p(s|H_s)ds. \tag{2.14}
\]

Therefore, by equation 2.11 and equation 2.14, we have

\[
p(t|H_t) = \lambda(t|H_t) \exp \left\{ -\int_0^t \lambda(s|H_s)ds \right\}. \tag{2.15}
\]

So, we have shown that given the conditional intensity function the interspike interval
density function is specified and vice versa. Hence, defining one completely defines the other.

2.2 The Likelihood Function of a Point Process Model

The likelihood function is formulated by deriving the joint distribution of the data and then, viewing this joint distribution as a function of the model parameters with the data fixed. Likelihood approach has many optimality properties that makes it a central tool in statistical theory and modeling.

With the interspike interval density function for any time $t$, the likelihood of a neural spike train is defined by the joint probability density function of all the data, or in another word, it is the joint probability density function of exactly $n$ events happening at times $\tau_i$, for $i = 1, 2, ... n$.

\[
\begin{align*}
  f(\tau_1, \tau_2...\tau_n \cap N(T) = n) &= \prod_{k=1}^{n} \lambda(\tau_k | H_{\tau_k}) \exp \left\{ - \int_0^T \lambda(s | H_s) ds \right\} \\
  &= \exp \left\{ \int_0^T \log \lambda(s | H_s) dN(s) - \int_0^T \lambda(s | H_s) ds \right\}
\end{align*}
\]

Please refer Proposition 1 in Appendix for proof. Proposition 1 shows that the joint probability density of a spike train process can be written in a canonical form in terms of the condition intensity function (Brown and et. al., 2002). That is, when formulated in terms of the conditional intensity function, all point process likelihoods have the form given in Equation 2.16.

If the density function in Equation 2.16 depends on an unknown $q$-dimensional parameter $\theta$ to be estimated, then the equation viewed as a function of $\theta$ is the
likelihood function defined as

\[ L(\theta) = \exp \left\{ \int_0^T \log \lambda(s|H_s, \theta) dN(s) - \int_0^T \lambda(s|H_s, \theta) ds \right\}. \tag{2.17} \]

One of the most compelling reasons to use maximum likelihood estimation in neural spike train data analyzes is that for a broad range of models, these estimates have other important optimality properties in addition to being asymptotically Gaussian. First, there is consistency which states that the sequence of maximum likelihood estimates converges in probability (or more strongly almost surely) to the true value as the sample size increases. Second, the convergence in probability of the estimates means that they are asymptotically unbiased. That is, the expected value of the estimate \( \hat{\theta} \) is \( \theta \) as the sample size increases. For some models and some parameters, unbiasedness is a finite sample property. The third property is invariance. That is, if \( \psi(\theta) \) is any transformation of \( \theta \) and \( \hat{\theta} \) is the maximum likelihood estimate of \( \theta \), then \( \psi(\hat{\theta}) \) is the maximum likelihood estimate of \( \psi(\theta) \). Finally, the maximum likelihood estimates are asymptotically efficient in that as the sample size increases, the variance of the maximum likelihood estimate achieves the Cramer-Rao lower bound. This lower bound defines the smallest variance that an unbiased or asymptotically unbiased estimate can achieve. Like unbiasedness, efficiency for some models and some parameters is achieved in a finite sample. Detailed discussions of these properties are given in the statistical inference book by Casella and Berger (1990).

### 2.3 Time-rescaling Theorem

A form of the time-rescaling theorem is well known in elementary probability theory. It states that any inhomogeneous Poisson process may be rescaled or transformed into a homogeneous Poisson process with a unit rate (Taylor, 1994). The
inverse transformation is a standard method for simulating an inhomogeneous Poisson process from a constant rate (homogeneous) Poisson process.

If the point process is an inhomogeneous Poisson process, then the conditional intensity function $\lambda(t|H_t)$ equals to $\lambda(t)$, which is simply the Poisson rate function not depending on the history of the process. Hence, the conditional intensity generalizes the definition of the Poisson rate. And the time-rescaling theorem states that any point process with an integrable conditional intensity function may be transformed into a Poisson process with unit rate (Papangelou, 1972).

We can rewrite the joint probability density function, Equation (2.16), to apply the time-rescaling theorem to spike train data series.

$$f(\tau_1, \tau_2, ..., \tau_n \cap N(T) = n) = f(\tau_1, \tau_2, ..., \tau_n \cap N(\tau_n) = n) \cdot \Pr(\tau_{n+1} > T|\tau_1, \tau_2, ..., \tau_n)$$

$$= \prod_{k=1}^{n} \lambda(\tau_k|H_{\tau_k}) \exp\left\{-\int_{\tau_{k-1}}^{\tau_k} \lambda(s|H_s) ds\right\} \cdot \exp\left\{-\int_{\tau_n}^{T} \lambda(s|H_s) ds\right\},$$

(2.18)

where $\tau_0 = 0$. The first term $f(\tau_1, \tau_2, ..., \tau_n \cap N\tau_n = n)$ is the joint probability density of exactly $n$ events in $(0, \tau_n]$, whereas the second term is the probability that the $(n + 1)^{th}$ event occurs after time $T$.

**Theorem 1. Time-rescaling** Let $0 < \tau_1 < \tau_2 < ... < \tau_n < T$ be a realization from a point process with a conditional intensity function $\lambda(t|H_t)$ satisfying $0 < \lambda(t|H_t)$ for all $t \in (0, T]$. Define the transformation

$$\Lambda(\tau_k) = \int_{0}^{\tau_k} \lambda(s|H_s) ds,$$

(2.19)

for $k = 1, 2, ..., n$, and assume $\Lambda(t) < \infty$ with probability one for all $t \in (0, T]$. Then the $\{\Lambda(\tau_k)\}$ are a Poisson process with unit rate.
Please refer Proposition 2 in Appendix for proof.

We may use the time-rescaling theorem to construct goodness-of-fit tests for a spike train model (Brown, 2001). Once a model has been fit to spike train data, we can compute the rescaled times by \( \eta_k = \Lambda(\tau_k) - \Lambda(\tau_{k-1}) \). If the model is correct, according to the time-rescaling theorem, all the \( \eta_k \)'s are independent exponential random variables with mean 1. We can make the further transformation, \( z_k = 1 - \exp(-\eta_k) \), then we get all the \( z_k \)'s are independent uniform random variables on the interval \((0, 1)\). Because both the transformations are one-to-one, any statistical assessment that measures how well a uniform distribution fits the \( z_k \)'s directly evaluates the goodness-of-fit of the spike train model. Brown (2001) presented two kinds of tests for the fit, Komogorov-Smirnov (KS) tests and quantile-quantile (Q-Q) plots. Both of methods are popular and well-known. Used together, the two plots help approximate upper and lower limits on the discrepancy between a proposed model and a spike train data.

2.4 Neural Synchrony

As we pointed out in Introduction, large-scale recording detect neuronal activities in the same region or across different regions in brain. The results reveal the fact that synchronized neural oscillations are a fundamental mechanism for enabling coordinated activity in the normally functioning brain (Buzsaki and Draguhn, 2004; Fries, 2009). A large body of evidence from both invasive and noninvasive experiments demonstrated that synchronization of oscillatory response is involved in a variety of cognitive functions, such as perceptual grouping, attention-dependent stimulus selection, routing of signals across distributed cortical networks, sensory-motor integration, working memory, and perceptual awareness (Uhlhaas et al., 2009). There are also a large number of studies suggest close correlation between abnormalities in
neuronal synchronization and brain disorders, such as schizophrenia, epilepsy, autism, Alzheimer's disease, and Parkinson’s disease (Uhlhaas et al., 2006). All the neurobiological findings enhance the importance of investigating the mechanism and generation of neural synchrony in a statistical context.

In Harris (2003), the activity of simultaneously monitored pyramidal cells within the CA1 region were rearranged so that synchronously firing cells were displayed near each other. The patterns suggest a cell assembly organization, with different sets of cells repeatedly showing synchronous activity at different times. He proposed a ”peer prediction” method to verify that the visualized assembly organization was not simply due to chance coincidences of spikes in a statistical sense. Under the peer prediction framework, if neurons are organized into assemblies whose firing is only partially determined by external factors, it should be possible to better predict when a neuron will fire, given the spike times of simultaneously recorded assembly members (Harris, 2003). We will discuss more details about the model in the Methods chapter.
In Chapter 1, the conditional intensity function is defined in terms of the instantaneous firing probability of the point process. However, this probability distribution is typically unknown. Therefore, much work has been done essentially constructing a model for the conditional intensity function.

This chapter is organized as follows. In Section 1, we discuss three typical parametric models. The advantages of parametric models are (1). often efficiently computable, (2). explainable parameters related to some well known factors and (3). nice asymptotic properties when the model is actually supported by abundant knowledge of the neural firing. In the field of nonparametric function estimation with stochastic data, known as data smoothing, vast advancements have been made in the past decades (Pouzat, Chaffiol and Gu, 2008). So, in Section 2, we review several popular firing rate methods that estimate the conditional intensity function in nonparametric ways. Those nonparametric approaches have the advantage due to fewer assumptions and they are often easier to implement and remain accurate when the parametric form is incorrect.
3.1 Parametric Models

3.1.1 A Special Interspike Interval (ISI) Probability Model

In some cases, with sufficient guidance of the neural characteristics, we can assume the ISI density function follows a certain distribution so that the likelihood function can be explicitly derived. For further understanding, the spike train data from goldfish retinal ganglion neurons recorded in vitro are studied in this example (see Figure 3.1). For this particular type of data, a gamma probability density is applied to model the ISI since retinal ganglion cells are relatively simple in mechanism and their firing follows a simple stochastic integrate-and-fire model with excitatory Poisson inputs. Inverse Gaussian (IG) probability density is an alternative model in which the inputs are assumed as a random walk with drift.

Let us introduce some background about the Integrate-and-Fire model here to help statisticians better acquaint with the neurobiological concepts and ideas.

Instead of considering the spike train as an isolated process without influence of inputs, from now on we always think in a more realistic situation: the integrate-and-fire model is part of a larger network, and we assume neuron firing is triggered by input currents, which are generated by the activity of presynaptic neurons.

Integrate-and-fire is one of the earliest models of a neuron, which was first investigated by Lapicque (1907). This model occurs as a particular case of the Hodgkins-Huxley equations (Knight, 2007). When an input current is applied, the membrane voltage increases with time until it reaches a constant threshold $V_{\text{thre}}$, at which point a delta function, a spike, occurs and the voltage is reset to its resting potential, after which the model continues to run. All integrate-and-fire neurons can be stimulated either by external current or by synaptic input from presynaptic neurons.

In the framework of the integrate-and-fire model, each presynaptic spike generates
a postsynaptic current pulse, and the total input current to a integrate-and-fire neuron is the sum over all current pulses. Taking the fact that the current for each firing is relatively stable, the first model we define here considers a neuron as a non-leaky integrator with excitatory Poisson inputs only. So the membrane voltage time course involves only the numbers of inputs as follows (Tuckwell, 1988),

$$dV(t) = \delta_E dN(t), \quad (3.1)$$

where $N(t)$ is a Poisson process with constant rate parameter $\lambda$, and $\delta_E$ is the magnitude of each excitatory input. Suppose the resting membrane potential at time $t_0$ is $V(t_0) = 0$ and the neuron discharges an action potential when $V(t) \geq V_{\text{thre}}$. For $V(t) \geq V_{\text{thre}}$, we must have $\delta_E N(t) \geq V_{\text{thre}}$ or $N(t) \geq V_{\text{thre}} \delta_E^{-1}$. If we let $[x]$ denote the greatest integer that is less or equal to $x$, then we require $N(t) > 1 + [V_{\text{thre}} \delta_E^{-1}]$ to generate a spike.

Recall that the waiting time probability we derived follows a gamma distribution with parameters $\lambda$ and $k$. For the primitive neuron model, we need $k = 1 + [V_{\text{thre}} \delta_E^{-1}]$ to observe an action potential. Hence, the interspike interval probability density is the gamma probability density with parameters $\lambda$ and $1 + [V_{\text{thre}} \delta_E^{-1}]$. The special case is when $k = 1$, the interspike probability density is exponential, which is a naive model and does not represent the ISI well in real cases.

For a more generalized application, the inputs can be considered as a random walk with both excitatory and inhibitory current instead of just the excitatory processes (Gerstein, 1964 and Tuckwell, 1988). So we define the membrane voltage equation for a non-leaky integrator neuron with Gaussian random walk inputs as

$$dV_R(t) = \delta_E dN_E(t) - \delta_I dN_I(t). \quad (3.2)$$
First, suppose that for the simplest case, $\delta_E = \delta_I = \delta$ and $\lambda_E = \lambda_I$, we have $V_R(t) \to W(t)$ in distribution, where $W(t)$ is a Wiener process. The Wiener process, $W(t)$, $t \geq 0$, is defined by the following three properties:

(i) $W(0) = 0$;

(ii) if $(t_j, t_{j+1}]$ and $(t_k, t_{k+1}]$ are non-overlapping intervals, then $W(t_{j+1}) - W(t_j)$ and $W(t_{k+1}) - W(t_k)$ are independent;

(iii) $[W(t_{k+1}) - W(t_k)] \sim N(0, \sigma^2(t_{k+1} - t_k))$, where $\sigma^2 = 2\delta^2\lambda$.

Next, from this model we then transform the Wiener process by adding a drift as follows:

$$V(t) = V_0 + \sigma W(t) + \beta t.$$  \tag{3.3}

The mean and the variance are

$$E[V(t)] = V_0 + \beta t,$$
$$V[V(t)] = \sigma^2 t.$$  \tag{3.4}

The first time the membrane voltage crosses the threshold is called the first passage time. We define the first passage time as the condition $t_{\text{thre}} = \inf \{u | V(u) = V_{\text{thre}}\}$ and $V(0) = V_0 < V_{\text{thre}}$. So the first passage time probability density for the Wiener process with drift is given by the inverse Gaussian probability density as

$$p_{\text{thre}}(t) = \frac{V_{\text{thre}} - V_0}{(2\pi \sigma^2 t^3)^{1/2}} \exp \left\{ -\frac{(V_{\text{thre}} - V_0 - \beta t)^2}{2\sigma^2 t} \right\} \quad t > 0.$$  \tag{3.5}

The probability of reaching threshold in finite time is defined by

$$\Pr(t_{\text{thre}} < \infty) = \begin{cases} 1 & \beta \geq 0 \\ \exp \left( -\frac{2|\beta(V_{\text{thre}} - V_0)}{\sigma^2} \right) & \beta > 0 \end{cases}$$  \tag{3.6}
Figure 3.1: (a): Thirty seconds of spike times from a retinal ganglion neuron recorded in vitro under constant illumination. (b): Interspike interval (ISI) histogram for the neural spike train in (a).

The mean and the variance of the inverse Gaussian probability density are

\[
\mu = E[t_{thre}] = \frac{V_{thre} - V_0}{\beta} \frac{V_{thre}}{\delta_E \lambda_E - \delta_I \lambda_I},
\]

\[
V[t_{thre}] = \frac{(V_{thre} - V_0) \sigma^2}{\beta^3} = \frac{(\delta_E^2 \lambda_E + \delta_I^2 \lambda_I) V_{thre}}{(\delta_E \lambda_E - \delta_I \lambda_I)^3},
\]

where \( \beta > 0 \). If \( \left( \frac{\mu}{\lambda} \right)^{-1} \) is large, then the inverse Gaussian probability density is well approximated by a Gaussian probability density.

In this first example, we study a spike train data series from a goldfish retinal ganglion cell recorded in vitro (Iyengar and Liao, 1997). Recordings of retina ganglion cells were made with an extracellular microelectrode under constant illumination. The plot of the spike train reveals a collection of short and long interspike intervals (ISI). To analyze these data, we consider that the Poisson process and the ISI probability
could follow a gamma distribution, as the special cases. The gamma and inverse Gaussian probability densities are, respectively,

\[ p_1(u_i|\theta) = \frac{\lambda^k}{\Gamma(k)} u_i^{k-1} \exp(-\lambda u_i), \quad (3.8) \]

where \( \theta = (k, \lambda), k > 0, \lambda > 0, \) and

\[ p_2(u_i|\theta) = \left( \frac{\lambda}{2\pi u_i^3} \right)^{1/2} \exp \left\{ -\frac{\lambda(u_i - \mu)^2}{2\mu^2 u_i} \right\}, \quad (3.9) \]

where \( \theta = (\mu, \lambda), \mu > 0, \lambda > 0. \)

Fitting this model into the spike train data requires construction of the likelihood and estimation of \( \theta. \) By our results in the previous section, the log likelihood can be presented either in terms of the conditional intensity or the ISI probability model. Here we use the latter. Given the set of interspike intervals (ISI), \( u = (u_1, u_2, ..., u_n), \) then under the assumption that the ISIs are independent, the likelihood functions for the two models are, respectively,

\[ L_1(\theta|u) = \prod_{i=1}^{n} p(u_i|\theta) = \left[ \frac{\lambda^k}{\Gamma(k)} \right]^{n} \prod_{i=1}^{n} u_i^{k-1} \exp(-\lambda u_i), \quad (3.10) \]

and

\[ L_2(\theta|u) = \prod_{i=1}^{n} p_2(u_i|\theta) = \prod_{i=1}^{n} \left( \frac{\lambda}{2\pi u_i^3} \right)^{1/2} \exp \left\{ -\frac{\lambda(u_i - \mu)^2}{2\mu^2 u_i} \right\}. \quad (3.11) \]

The maximum likelihood estimate of \( \theta \) for the gamma model cannot be computed.
in closed form, but rather numerically as the solution to the equations by differentiating Equation (3.10) with respect to $k$ and $\lambda$.

$$\hat{\lambda} = \hat{k} / \bar{u},$$

$$n \log \Gamma(\hat{k}) = n \hat{k} \log \left( \frac{\hat{k}}{\bar{u}} \right) + (\hat{k} - 1) \sum_{i=1}^{n} \log u_j.$$ \hspace{1cm} (3.12)

where $\bar{u} = n^{-1} \sum_{i=1}^{n} u_i$.

Similarly, differentiating Equation (3.11) with respect to $\mu$ and $\lambda$ yields the maximum likelihood estimates for the IG model

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} u_i,$$

$$\hat{\lambda}^{-1} = n^{-1} \sum_{i=1}^{n} (u_i^{-1} - \hat{\mu}^{-1}).$$ \hspace{1cm} (3.13)

### 3.1.2 Brillinger’s GLIM Approach for Interacting Neurons

Recall the “integrated-and-fire” model we introduced in the previous section. The presynaptic or other input currents are additive until an action potential is triggered when crossing its particular threshold. Based on the biological facts, Brillinger (1988) proposed an approach to estimate the firing rate of a neuron by its interacting nerve cells. The approach is flexible for arbitrary numbers of the input cells, and it allows biological interpretation of the results.

In the approach, spike trains are replaced by a 0-1 time series, which essentially discretizes the continuous time domain into finite numbers of small intervals for computation purposes. The computations are realized by the GLIM program. For example, in a simple network that includes only two neurons A and B, the ISI density function of neuron A is composed of the history behavior of its own and the influence
from neuron B. The influence, of course, can be both excitatory and inhibitory.

\[ p_t = \Phi(\sum_{u=0}^{\gamma_t-1} b_u \Gamma_{t-u} + \theta_1 \gamma_t + \theta_2 \gamma_t^2 + \theta_3 \gamma_t^3 - \theta), \]  

(3.14)

where \( \Phi(\cdot) \) is for the normal cumulative function, \( \gamma_t \) is the time elapsed since the last firing of neuron A, and \( \Gamma_t \) is a 0-1 time series that represents the spike times of neuron B. Those unknown parameters are estimated by maximizing the likelihood function.

### 3.1.3 Relating Neural Spiking Activities to Covariate Effects via the Generalized Linear Model (GLM) Framework

In the previous case in section 3.1.1, we have a strong assumption of the underlying distribution of ISI density function, which could be less appropriate if we lack knowledge of the mechanism of various kinds of neurons. Another currently widely used class of parametric models specifies the form of conditional intensity function, rather than the ISI density, and the point process model expresses the conditional intensity function as a function of time, history and other variables.

We have already seen that spiking history often plays an important role in determining when the next spike will occur. Therefore the history as one group of covariates should naturally be considered for neural point process models. If the spike train being modeled comes from a larger ensemble of recorded neurons that interact with each other, it may be useful to consider the firing histories of the other neurons as well. Also, for many experiments, there are other signals, (external covariates) besides the history terms. These external covariates are often recorded simultaneously with the point process. For example, in any stimulus-response experiment, it is expected that some function of the stimulus affects the firing probability.

Truccolo et al. (2005) proposed a statistical framework based on the point process
likelihood function to relate three typical types of covariates to a neural spiking probability. First, spiking activity is associated with extrinsic covariates such as sensory stimuli and behavior. For instance, retinal neurons respond to light intensity and contrast. Second, the current spiking activity of a neuron is also related to its past activity, reflecting biophysical properties such as refractoriness. Third, spiking activity in a given neuron is affected by concurrent ensemble spiking activity (called peer activity). So Truccolo et al. modeled the logarithm of the conditional intensity function as a linear combination of functions of the covariates, and they used the discrete time likelihood function for point process to carry out the analysis. The general form of the conditional log-intensity function was expressed as follows:

$$\log \lambda(t_k | H_k) = \sum_{i=1}^{q} \beta_i g_i(\varphi_i(t_k - \omega)),$$

where $g_i$ is a general function of a covariate $\varphi_i(t_k)$ at different time lags $\omega$, and $\beta_i$’s are the corresponding parameters with $q$ dimension.

Following the proof in Proposition 1 (see Appendix), the joint probability density function, Equation (2.16), has a discrete time representation, in which the probability can be expressed in terms of discretized spike trains and their conditional intensity function. Assume the partition of the observation time $(0, T]$, $(t_{k-1}, t_k]_{k=1}^K$, is sufficiently small so that there is at most one spike in any $(t_{k-1}, t_k]$ for arbitrary $k$; that is, $\Delta N_k = N(t_k) - N(t_{k-1})$ can be either 0 or 1. Then the probability of the point process with $n$ events in $(0, T]$ is approximately as follows:

$$\Pr \left( \tau_1, \tau_2, \ldots, \tau_n \cap N(T) = n \right) \approx \exp \left\{ \sum_{k=1}^{K} \log[\lambda(t_k | H_k)\Delta] \Delta N_k - \sum_{k=1}^{K} \lambda(t_k | H_k)\Delta \right\}.$$

(3.16)

This framework covers a very large class of models because Equation 3.15 allows for general functions of covariates.
3.2 Nonparametric Models

3.2.1 Kernel Smoothing

The firing rate is a fundamental concept for the description of a spiking neuron (a point process in general). The underlying firing rate $\lambda(t)$ is a non-negative deterministic function of time, such that the integral $\int_t^{t+\Delta} \lambda(u)du$ represents the expected number of spikes encountered in an observation of the neuron during the observation interval $(t, t+\Delta]$. The rate function underlying the spiking of a real neuron, however, cannot be observed directly, it must be reconstructed from the recorded spike trains. Nawrot et al. (1999) described a method to estimate the neuronal firing rate from single-trial spike trains by convolution with a fixed kernel function.

Consider a single spike train, comprised of a finite number of discrete spike events at times $\tau_1, ..., \tau_n$. We define the estimation of the time-varying rate function as

$$\lambda(t) \doteq \sum_{i=1}^{n} K(t - \tau_i) \quad (3.17)$$

where $K(t)$ is called kernel function. Thus, the desired underlying rate function $\rho(t)$ is estimated from a single-trial spike train by taking the sum over kernel functions $K(t - \tau_i)$, centered at spike occurrence times $\tau_i$.

$K(t)$ is required to be non-negative to avoid negative rates. Moreover, the kernel should be normalized such that each spike contributes with unit area to the rate function; this also guarantees that the integral of $\lambda(t)$ is equal to the total number of spikes $n$ recorded during the interval $(0, T]$. Finally, the first moment of $K(t)$ is required to be zero to preserve the center of mass of the spike train.

There are two important aspects to specify a kernel function: the shape of the kernel function and its width. The kernel shape determines the visual appearance of
Figure 3.2: Concept of single-trial rate estimation by means of the kernel approach. (A) The true underlying rate function. (B) One single-trial spike train. (C) Kernel functions centered at spike occurrence times. (D) The empirical rate function. In this particular example, a triangular kernel with a standard width of 40 ms was used.
Figure 3.3: Kernel functions

the estimated rate function. The kernel width is defined as

\[ \sigma = \sqrt{\int_{-\infty}^{\infty} t^2 K(t) dt} \]  \hspace{1cm} (3.18)

which can be viewed as a smoothing parameter. Table 1 listed four kernel functions of different shapes, parameterized by their standard width \( \sigma \).

<table>
<thead>
<tr>
<th>Kernel</th>
<th>( K(t, \sigma) )</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxcar</td>
<td>( \frac{1}{2\sqrt{3\sigma}} )</td>
<td>([-\sqrt{3\sigma}, \sqrt{3\sigma}] )</td>
</tr>
<tr>
<td>Triangle</td>
<td>( \frac{1}{6\sigma^2(\sqrt{6\sigma}-</td>
<td>t</td>
</tr>
<tr>
<td>Epanechnikov</td>
<td>( \frac{3}{4\sqrt{5\sigma}} \left(1 - \frac{t^2}{5\sigma^2}\right) )</td>
<td>([-\sqrt{5\sigma}, \sqrt{5\sigma}] )</td>
</tr>
<tr>
<td>Gauss</td>
<td>( \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{t^2}{2\sigma^2}\right) )</td>
<td>([-\infty, +\infty] )</td>
</tr>
</tbody>
</table>

*Outside the support, the kernels are defined to be zero. All kernel functions are normalized to unit area and have standard width \( \sigma \).
density estimator, is most often used.

\[ \lambda(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{i=1}^{n} \exp\left( -\frac{(t - \tau_i)^2}{2\sigma^2} \right) \]  

(3.20)

When \( \lambda(t) \) varies slowly, Gaussian kernels do a good job of estimating the rate and filtering out the noise. Nevertheless, when the firing rate varies quickly, Gaussian filters are not able to capture the variation without introducing artificial high-frequency fluctuations. In other words, to filter high-frequency noise, the Gaussian kernel density estimator must remove the high-frequency firing rate.

In general, the most obvious advantage of kernel smoothing is its simplicity. Kernel smoothing methods are extremely fast and simple to implement. However, as the results depend critically on the choice of the smoothing parameter \( \sigma \), the lack of a global choice of the bandwidth is typically considered a major shortcoming of kernel smoothing methods.

### 3.2.2 Adaptive Kernel Smoothing

There are two concerns about the standard fixed bandwidth kernel estimation. One is that it requires the investigator to choose the parameter \( \sigma \) arbitrarily, which produce significantly different estimates of firing rates. The other problem is that the bandwidth of the standard kernel estimator is constant throughout the time interval of the neuronal response. Richmond et al.(1990) tried to solve those two problems by varying the width of the estimation kernel throughout the trial, letting the data themselves determine how much to vary the width of the kernel. This process is called adaptive-kernel estimation, because the width of the kernel adapts to the local density of the data points (Richmond, 1990).

The procedure is basically as follows,
(i) Firstly, form a fixed bandwidth kernel estimate (called a pilot estimate) from the data, the bandwidth is $\sigma_p$;

(ii) Then, this pilot estimate of the density is used as a measure of the activity over small time periods throughout the entire response interval. The definition of ”small” here depends on the choice of the fixed kernel bandwidth, $\sigma_p$;

(iii) These local kernels are then used to produce a smoothed firing rate that changes more rapidly in regions of high firing, and less in regions of less firing.

The pilot estimate is used to define a set of local bandwidth factors, $\lambda_i$,

$$
\lambda_i = \sqrt{\frac{f(i)}{\mu}} \tag{3.21}
$$

where $f(i)$ is the pilot estimate at the $i^{th}$ point, and $\mu$ is the geometric mean of the pilot estimates,

$$
\mu = \exp \left[ \frac{1}{n} \sum_{i=0}^{n-1} \ln f(i) \right] \tag{3.22}
$$

Finally, the adaptive kernel estimate is computed by convolving each point with a kernel density function having a width that is the product of the fixed bandwidth $\sigma_p$ and the factor for each point

$$
m(k) = \frac{1}{n} \sum_{0}^{n-1} K\left( \frac{t - \tau_i}{\sigma \lambda_i} \right) \tag{3.23}
$$

Adaptive kernel smoothing benefits from the simplicity of kernel smoothing methods, and the additional complexity of the local kernel widths increases the computational effort only very slightly. Further, this approach lifts the strict stationarity requirement of many models. A possible shortcoming is that, even though it adapts the kernel width, the adaptive kernel smoothing still requires an specific choice of kernel width for the pilot estimate.
3.2.3 Kernel Bandwidth Optimization

Kernel smoother and a time histogram are classical tools for estimating an instantaneous firing rate. The optimization method was initially proposed for the joint peristimulus time histogram (PSTH) of spike counts over multiple neurons (Shimazaki, 2007b). The method can select the bin width of the time histogram automatically based on the principal of minimizing the mean integrated squared error (MISE), defined as follows, without knowing the underlying rate.

\[
MISE = \int_0^T E(\lambda(t) - \hat{\lambda}(t))^2 dt,
\]

where \(\lambda(t)\) is the underlying rate and \(\hat{\lambda}(t)\) is the estimation, and \(E\) refers to the expectation with respect to the spike generation process under a given time-dependent rate \(\lambda(t)\).

Similarly, we consider a kernel rate estimator as \(\hat{\lambda}(t)\) and select the width of a kernel under the MISE criterion. Suppose independently and identically obtained \(m\) spike trains which contain \(M\) spikes as a whole, then a superposition of the spike trains can be regarded as being drawn from an inhomogeneous Poisson point process, due to the general limit theorem of the sum of independent point process (Shimazaki, 2009). Let’s define \(\bar{Y}_t = \frac{1}{m} \sum_{i=1}^M \delta(t - \tau_i')\), where \(\tau_i'\) is the \(i\)th spike of the superimposed spike sequence and \(\delta(\cdot)\) is still the Dirac delta function. The estimator \(\hat{\lambda}(t)\) can be constructed by a kernel function \(K_w(\cdot)\) as \(\hat{\lambda}(t) = \int \bar{Y}_s K_w(t - s) ds\), where \(w\) refers to the bandwidth.

Following Equation 3.24, the integrand can be decomposed into three terms: \(\lambda(t)^2 - 2\lambda(t)E(\hat{\lambda}(t)) + E(\hat{\lambda}(t)^2)\). Since the first component does not depend on the choice of the kernel, we subtract it from the MISE and define a cost function as a
function of the bandwidth $w$,

$$C_m(w) = MISE - \int_0^T \lambda(t)^2 dt = -2 \int_0^T \lambda(t) E(\lambda(t)) dt + \int_0^T E(\lambda(t)^2) dt. \quad (3.25)$$

From a general decomposition rule of a covariance of the two random variables, we obtain

$$\int_0^T \lambda(t) E(\lambda(t)) dt = \int_0^T E(\widetilde{Y}_t \lambda(t)) dt - \int_0^T E(\widetilde{Y}_t - E(\widetilde{Y}_t))(\lambda(t) - E(\lambda(t))) dt$$

$$= E \left( \int_0^T \widetilde{Y}_t \lambda(t) dt \right) - \frac{K_w(0)}{n} E \left( \int_0^T \widetilde{Y}_t dt \right). \quad (3.26)$$

To obtain the next equality, we used the assumption that the spike sequence is a Poisson process so that the spikes are independent to each other. Hence, the cost function is estimated as

$$\hat{C}_m(w) = \frac{2K_w(0)}{m} \int_0^T \widetilde{Y}_t dt - 2 \int_0^T \widetilde{Y}_t \lambda(t) dt + \int_0^T \lambda(t)^2 dt$$

$$= \frac{2K_w(0)}{m^2} M - \frac{2}{m^2} \sum_{i=1}^M \sum_{j=1}^M K_w(\tau'_i - \tau'_j) + \frac{1}{m^2} \sum_{i=1}^M \sum_{j=1}^M \psi_w(\tau'_i - \tau'_j), \quad (3.27)$$

where $\psi_w(t)$ is given by $\psi_w(t) = \int_0^T K_w(s) K_w(s + t) ds$.

The optimal bandwidth $w^*$ minimizes the score function $\hat{C}_m(w)$.

### 3.2.4 Smoothing Splines

The penalty smoothing estimate the rate function by maximized the penalized likelihood function (Gu, 2008). We first consider a nonparametric regression problem $Y_i = \eta(x_i) + \varepsilon_i$, $i = 1, 2, ..., n$, where $n$ is the total number of points of a spike train, $x_i \in [0, T]$, $\varepsilon_i \sim N(0, \sigma^2)$ and $\eta(\cdot)$ is the smoothing spline function. Then we may
estimate \( \eta(\cdot) \) via the penalized least square

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 + \kappa \int_0^1 (\eta''(x))^2 \, dx
\] (3.28)

where the first term, \( \sum_{i=1}^{n} (Y_i - \eta(x_i))^2 \), measures the goodness-of-fit of the smoothing function \( \eta \) to the data, the second term, \( J(\eta) = \int_0^1 (\eta''(x))^2 \, dx \), penalizes the roughness of \( \eta(x) \) and the smoothing parameter \( \kappa \) controls the trade-off between the two conflicting goals. The solution of (3.28) is known as natural cubic spline. As \( \kappa \rightarrow \infty \), the estimate ”shrinks” into the null space of the roughness penalty, \( \{ \eta : \int_0^1 \eta''(x) \, dx \} \).

When \( \kappa \rightarrow 0 \), it converges to the minimum curvature interpolation, i.e. \( Y_i = \eta(x_i) \) for \( i = 1, 2, \ldots, n \).

For the point process data, log-likelihood is used as the goodness-of-fit measure, instead of the least square. As we known, when the error terms of the regression are assumed to be independent and identically distributed Normal random variables, ordinary least square (OLS) turns out to be equivalent to maximum likelihood estimation (ML).

Consider the so-called exponential family distributions with densities of the form

\[
f(y|\eta, \phi) = \exp\left\{ (y\eta - b(\eta))/a(\phi) + c(y, \phi) \right\},
\] (3.29)

where \( a > 0 \), \( b \) and \( c \) are known functions, \( \eta \) is the parameter of interest and \( \phi \) is either known or considered as a nuisance parameter. Observing \( Y_i f(y|\eta(x_i), \phi) \), one may estimate \( \eta(x) \) via the general penalized likelihood,

\[
-\frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \eta(x_i) - b(\eta(x_i)) \right\} + \kappa J(\eta)
\] (3.30)

where \( J(\eta) = \int_0^1 (\eta''(x))^2 \, dx \) and \( c(y, \phi) \) is dropped as it does not depend on \( \eta \).
A special case of the penalized likelihood method is for Poisson regression. For the density function \( f(y|\eta) = \exp\{y\eta - e^\eta - \log y!\} \), one may minimize the penalized likelihood
\[
-\frac{1}{n} \sum_{i=1}^{n} \{Y_i \eta(x_i) - e^{\eta(x_i)}\} + \kappa J(\eta),
\]
where \( \eta(x_i) = \log \lambda(x_i) \) is the log intensity and the \( \lambda(x_i) > 0 \) but \( \eta(x_i) \) is free of constraint.
Chapter 4

The Methods

Synchronized firing of related neurons is an interesting fact which motivates us to develop statistical analysis using the valuable information of concurrent spiking activities (or called peer activities). Our research focuses on applying and developing statistical approaches to analyze simultaneously recorded neural spike trains and infer functional connectivity between interactive neurons in brain. In this chapter, we introduce two models for spike train analysis: the first one is a point process likelihood method under the generalized linear model framework, and the second one is a proposed regression spline model for modeling both the spontaneous firing and also the influence of interactive nerve cells. Simulation results and real data applications will be presented in Chapter 5.

4.1 Continuous State-Space Model

We have shown that the point process likelihood function can be analyzed under the GLM framework and that the conditional log-intensity function can be expressed as a general function consisting of extrinsic covariates, past activity of its own and concurrent spiking activities. See section 3.1.3 for detail. Harris’ (2003) introduced a “peer prediction” method for the situations when external factors cannot explain some unpredictable spike times. The author also proved that the firing rate can be predicted better by incorporating the spike times of a simultaneously recorded
assembly population than by using only external variables. There are three major steps in the method. Initially, all the peer spike trains are smoothed with a Gaussian filter in the time domain.

\[ s(t)^\alpha = \frac{1}{\sqrt{2\pi \sigma^2}} \sum_{j=1}^{n^\alpha} \exp\left(-\frac{(t - \tau^\alpha_j)^2}{2\sigma^2}\right), \]  

(4.1)

where \( \alpha \) is the index of peer cells, \( \{\tau^\alpha_j\} \) is the spike train of neuron \( \alpha \), and \( n^\alpha \) is the number of spikes of peer \( \alpha \). \( \sigma \) is the smoothing bandwidth of the Gaussian kernel, termed peer prediction timescale by the author.

Then, under the generalized linear model, the predicted intensity function at time \( t \) is given by

\[ \lambda(t) = g\left(\sum_{\alpha} s(t)^\alpha w^\alpha\right), \]  

(4.2)

where \( w^\alpha \) is the prediction weight of peer \( \alpha \). The sign of the weight represents the positive or negative correlation between a certain member neuron and the target neuron. Here \( g(\cdot) \) is the link function and has the following form

\[ g(x) = \begin{cases} 
\exp(x) & x < 0 \\
 x + 1 & \text{otherwise} 
\end{cases} \]  

(4.3)

The advantage of this link function is that it will not lead to excessively high intensity when many positively correlated peer cells fire simultaneously compared with a simple exponential link function.

The final step is to estimate the weight by maximizing the penalized log-likelihood on the training set.

\[ \log(L_f) = \sum_t [-\lambda(t) \Delta + \Delta N_t \log(\lambda(t) \Delta)] - \frac{1}{4} \sum_{\alpha} (w^\alpha)^2. \]  

(4.4)
The maximum is carried out by Newton’s method with an analytically calculated Hessian matrix.

A 10-fold cross-validation procedure is used to repeatedly divide the recorded data into a training set and a test set. For each training set, the mean firing rate is calculated, $f_0$, as the number of spikes during the training period divided by the length of the training period. The prediction quality on the test set, termed predictability, is defined as the difference, $\log(L_f) - \log(L_{f_0})$, over the base of $\log(2)$. Then, the predictability of the entire data set is defined by a cross-validation procedure, where the data are divided into 10 segments, each segment is used in turn as test set, and the log likelihood ratios for each segment are summed and divided by the total time.

As the spike train is smoothed by the Gaussian kernel, then the choice of the bandwidth, $\sigma$ in Equation (4.1), is critical. In Harris (2003), the optimal bandwidth, $\sigma$, is chosen by optimizing the predictability; therefore, it maximizes the log likelihood function, $\log(L_f)$, among all the values of $\sigma$.

4.2 Regression Spline Model

The spontaneous firing rate for one particular neuron depends on its own natural characteristics. Also, the firings of other neurons (also called peers) within the neural network have impacts on the target neuron. In this section, we propose a regression spline model for neural spike train data with interactive neural activities.

4.2.1 Model

Suppose there are $M$ covariates and their firing times before time $t$ are given by $x(t) = (x_1(t), ..., x_M(t))$. Let $T$ be an nonnegative random variable ranging over a compact interval $T$. The distribution of $T$ may depend on the peers $x(t)$. Let $f(t|x)$
and $F(t|x)$ denote the conditional density and conditional distribution function, respectively, of $T$ given $x \in \mathbb{R}^M$. Let $\lambda(\cdot|x(s))$ denote the conditional intensity function of $T$ given $x(s)$ so that

$$\lambda(t|x(t)) \, dt = P(T \in (t, t + dt)|x(t)).$$

Let $\alpha(\cdot|x(t))$ denote the log conditional intensity function. To simplify the discussion, we assume that the conditional intensity function at time $t$ depends on the value of the covariates only up to that time; that is, we assume that $\lambda(t|x(s), s > 0) = \lambda(t|x(s), 0 \leq s \leq t)$ and hence that $\alpha(t|x(s), s > 0) = \alpha(t|x(s), 0 \leq s \leq t)$.

In this paper, we model the log conditional intensity function via

$$\alpha(t|x(t)) = \mu(t) + \beta \cdot Q(t|x(t)), \quad t \geq 0,$$  \hspace{1cm} (4.5)

where $\mu(t)$ is the baseline function which represents the spontaneous firing rate, $Q(\cdot|\cdot) = (Q_1, \ldots, Q_M)$ is the vector of $M$-dimensional polynomial functions which describe the effects of the covariates, and $\beta = (\beta_1, \ldots, \beta_M)$ are the coefficients. This model is flexible for an arbitrary number of interacting neurons. Brillinger (1988) gave a step function to approximate the peer effects on the target neuron. In our method, the components of $Q(\cdot|\cdot)$ are all continuous functions whose explicit forms would be discussed in detail in Section 4.

Let us explain a bit how we can understand Equation 4.5 intuitively. We take into account two basic facts: First, a nerve cell has a spontaneous firing rate along the time that is considered as a baseline without any intrinsic or extrinsic stimulus. Second, under the proportionality assumption, any type of inputs from the peer activities would accumulatively affect the conditional log-intensity function.
Set $\mathbf{x} = \mathbf{x}(t)$. The log-likelihood based on $(T, \mathbf{x})$ is given by

$$l = \log \lambda(T|\mathbf{x}) - \int \lambda(u|\mathbf{x})du = \alpha(T|\mathbf{x}) - \int \text{ind}(T \geq u) \exp \alpha(u|\mathbf{x})du$$

$$= \{\mu(T) + \beta \cdot Q(T|\mathbf{x})\} - \int \text{ind}(T \geq u) \exp(\mu(u) + \beta \cdot Q(u|\mathbf{x}))du.$$

The expected log-likelihood is given by

$$E \left( \log \lambda(T|\mathbf{x}) - \int \text{ind}(T \geq u) \lambda(u|\mathbf{x})du \right) = \int [\log \lambda(t|\mathbf{x}) f(t|\mathbf{x}) - (1 - F(t|\mathbf{x})) \lambda(t|\mathbf{x})] dt.$$

Let $\tau_1, \ldots, \tau_n$ be independent random variables having distribution functions $F(\cdot|\mathbf{x}_i)$, and let $\mathbf{x}_i \in \mathbb{R}^M$ denote the vector of covariates for the $i^{th}$ individual, $1 \leq i \leq n$. Let $G$ denote a linear space of polynomial spline functions on $T$. Let function $h(t|\mathbf{x}_i) = g(t) + \beta \cdot Q(t|\mathbf{x}_i)$, where $g \in G$. The expected log-likelihood function $\Lambda(\cdot)$ is defined by

$$\Lambda(h) = \sum_i \int [h(t|\mathbf{x}_i) f(t|\mathbf{x}_i) - (1 - F(t|\mathbf{x}_i)) \exp h(t|\mathbf{x}_i)] dt.$$ 

Observe that $\Lambda(\cdot)$ is maximized at $\alpha = \log(f/1 - F) = \mu + \beta \cdot Q$. $\mu(t)$ may or may not be in $G$, but we can define the best approximation to $\mu$, $\mu^* \in G$, so that $\alpha^* = \mu^* + \beta \cdot Q$ maximizes $\Lambda(\cdot)$ over $G$.

The best approximation will be chosen from the linear space $G$. Say for $1 \leq p < \infty$ the baseline function in Model 4.5 is spanned by the $B$-spline functions: $B_1, \ldots, B_p$, and

$$\alpha(t|\mathbf{x}) = \sum_{i=1}^p \theta_i B_i(t) + \beta \cdot Q(\mathbf{x}), \quad t \geq 0,$$ 

where $\theta = (\theta_1, \ldots, \theta_p)$ are the coefficients for $p$ basis functions.

We do not assume that $\mu$ is exactly equal to a spline, but we still can make use of spline approximation. In order for this method to be accurate, we need the error of approximation to tend to zero as the sample size $n$ tends to infinity; for this, it
is necessary that the dimension $p$ of the approximation space $G$ tend to infinity. To control the error of estimation, we need this dimension to increase more slowly than $n^{1/2}$.

The model aggregates the influences from peers using polynomial functions, and each covariate extends the model in terms of adding a polynomial term. The aggregating methodology offers high flexibility for real situations. In this paper, we will see the applications for applying different order functions for single or multiple covariates, and also we will discuss how to explain the applications in realistic cases later. For illustration purposes, we now use a special case with the polynomial function of order 1 to describe our proposed methodology. Equation 4.7 is a simplified version of 4.6 with only one covariate. The general case with multiple covariates is just a straightforward extension.

\[
\alpha(t|x(t)) = \sum_{i=1}^{p} \theta_i B_i(t) + \beta_1 Q_1(t|x_1(t)), \quad t \geq 0. \tag{4.7}
\]

### 4.2.2 Maximum Likelihood Estimation

Given the vector of covariates $x(t)$ and the baseline function spanned by a set of basis functions, $B = (B_1, ..., B_p)$, we can estimate the coefficients in Equation 4.7 by maximum likelihood. The partial log-likelihood for a neuron can be written as

\[
\phi(t|x(t)) = \log \lambda(t|x(t)) - \int_0^t \lambda(s|x(s))ds = \alpha(t|x(t)) - \Lambda(t|x(t)). \tag{4.8}
\]
We take the partial derivatives of $\phi(\cdot)$ to examine its concavity and maximize the log likelihood. Set

\[
D_\theta^j = \frac{\partial \Lambda(t|x)}{\partial \theta_j} = \int_0^t B_j(s) \exp (\theta \cdot B(s) + \beta \cdot Q(s|x)) \, ds \quad \text{for } j = 1, \ldots, p,
\]
\[
D_\beta^k = \frac{\partial \Lambda(t|x)}{\partial \beta_k} = \int_0^t Q_k(s|x_k(s)) \exp (\theta \cdot B(s) + \beta \cdot Q(s|x)) \, ds \quad \text{for } k = 1, \ldots, M,
\]

(4.9)

and

\[
E_{\theta,j,l}^\theta = \frac{\partial^2 \Lambda(t|x(t))}{\partial \theta_j \partial \theta_l} = \int_0^t B_j(s)B_l(s) \exp (\theta \cdot B(s) + \beta \cdot Q(s|x)) \, ds
\]

for $j, l = 1, \ldots, p$,
\[
E_{\theta,k}^\theta = \frac{\partial^2 \Lambda(t|x(t))}{\partial \theta_j \partial \beta_k} = \int_0^t B_j(s)Q_k(s|x_k(s)) \exp (\theta \cdot B(s) + \beta \cdot Q(s|x)) \, ds
\]

for $j = 1, \ldots, p$ and $k = 1, \ldots, M$,
\[
E_{\beta,q}^\beta = \frac{\partial^2 \Lambda(t|x(t))}{\partial \beta_k \partial \beta_q} = \int_0^t Q_k(s|x_k(s))Q_q(s|x_k(s)) \exp (\theta \cdot B(s) + \beta \cdot Q(s|x)) \, ds
\]

for $k, q = 1, \ldots, M$.

(4.10)

Then

\[
\frac{\partial \phi}{\partial \theta_j} = B_j(t) - D_\theta^j, \quad \frac{\partial \phi}{\partial \beta_k} = Q_k(t|x_k(t)) - D_\beta^k,
\]
\[
\frac{\partial^2 \phi}{\partial \theta_j \partial \theta_l} = -E_{\theta,j,l}^\theta, \quad \frac{\partial^2 \phi}{\partial \theta_j \partial \beta_k} = -E_{\theta,k}^\theta, \quad \frac{\partial^2 \phi}{\partial \beta_k \partial \beta_q} = -E_{\beta,k,q}^\beta.
\]

(4.11)

It follows from Equation 4.11 that $\phi(t|x(t))$ is a concave function. Hence there exists a unique maximum likelihood estimate $\hat{h} = \hat{\theta} \cdot B \in G$ [see Kooperberg et al. (1995b)] and $\hat{\beta}$ so that $[\hat{\theta}, \hat{\beta}] = \max l(\theta, \beta)$.

The neural spike train data are collected as the time-series of electrical impulses
generated by individual neurons, and the typical form of a neural spike train is a
temporal point process that shows precisely the times of firing. The interval between
two consecutive firings as shown in Figure 5.13 are referred to as interspike interval
(ISI), and here we denote the interspike intervals for a spike train with \( N \) spikes as
\( \{ u_k \}_{k=1}^N \).

The log-likelihood function is a sum for all \( N \) consecutive ISIs \( \{ u_k \}_{k=1}^N \) given by

\[
l(\theta, \beta) = \sum_{k=1}^{N} \alpha(u_k|x(u_k)) - \sum_{k=1}^{N} \int_{0}^{u_k} \lambda(s|x(s))ds \quad (4.12)
\]

The maximum likelihood estimate \( \hat{\theta} \) and \( \hat{\beta} \) will be obtained by maximizing the
log-likelihood, \( l(\theta, \beta) \).

Under certain conditions, Kooperberg, Stone and Truong (1993) obtained the \( L_2 \)
rate of convergence in estimating the log intensity function.

### 4.2.3 Adaptive Model Selection

A useful feature in this model for neural spike train analysis is that the space \( G \) of
approximation is chosen adaptively. This can be really useful in capturing respective
features of firing rate for various neurons in a large network. The methodology is
similar in spirit to MARS (multivariate adaptive regression splines), and the choice
of the space \( G \) and its dimension \( p \) are resolved adaptively.

The selection of the dimension and the basis functions of \( G \) employs stepwise
addition and stepwise deletion of basis functions. Initially, we use minimal allowable
space to model \( \alpha(t|x(t)) \), so that the one dimensional model is fit. Then we proceed
with stepwise addition, successively replacing a \( (p - 1) \)-dimensional space \( G_0 \) by a \( p \)-
dimensional space \( G \) that contains \( G_0 \) as a subspace. At each step a candidate basis
function is added to the model. When multiple regression is possible for evaluating
candidates for a new basis function, we choose among the various candidates by a
heuristic search, which maximizes the Rao statistics.

The addition would be stopped when one of the following conditions is satisfied; see Kooperberg et al. (1995) for detail.

1. The number $P$ of basis functions equals $P_{\text{max}}$, where the default value for $P_{\text{max}}$ is $\min(4n^{1/5}, n/4, 30)$.

2. The search algorithm yields no possible new basis function.

3. $\hat{l}_P - \hat{l}_p < \frac{1}{2}(P - p) - 0.5$ for some $p$ with $3 \leq p \leq P - 3$, where $\hat{l}_p$ is the log-likelihood for the model with $p$ basis functions.

Upon stopping the stepwise addition stage, we proceed to stepwise deletion by successively replacing the $p$-dimensional allowable space $G$ by a $(p-1)$-dimensional allowable subspace $G_0$ until we arrive at the minimal allowable space. For each step, the basis function that would be removed in going from $G$ to $G_0$ has the smallest Wald statistic in space $G$. The Rao statistics during the stepwise addition and Wald statistics during the stepwise deletion give an approximation of the change in the log-likelihood due to adding or deleting a basis function that does not require finding the new maximum likelihood estimation of parameters.

During the combination of stepwise addition and deletion, we get a sequence of models. Let $p_{\nu}$ denote the number of parameters and $\hat{l}_{\nu}$ be the log-likelihood of the $\nu^{th}$ model. The method selects the model that minimize the Bayes information criterion $\text{BIC} = -2\hat{l}_{\nu} + p_{\nu} \log n$. 

43
This chapter presents the simulation results and real data applications according to the methods introduced in Chapter 4. Section 5.1 and 5.2 follow the models in Section 4.1 and Section 4.2 respectively.

5.1 Continuous State-Space Model

5.1.1 Simulation

First, we consider the simplest situation, when there is only one peer with the target spike train. To test the prediction methods when the true underlying firing rate is known as $\lambda_1(t)$, we simulated two inhomogeneous Poisson spike trains positively correlated with weight 3, the target spike train with rate $\lambda_1(t)$, and its peer with rate $\lambda_2(t)$, so that $\lambda_1(t) = 3 \times \lambda_2(t) + 1$ for all time $t$ from 0 to 50 seconds. Both the training and testing sets are generated in the same way.

The spike train generated from the time-rescaling algorithm is shown in Figure 5.1. The non-uniform distribution of the ticks on the raster plot indicates a model based on a inhomogenous Poisson process for this spike train, and the ISI clearly follows an exponential distribution. The underlying rate function is shown in Figure 5.2. The firing rate, which is intense in the middle of the raster plot and less frequent on both sides, matches the underlying intensity function.
Figure 5.1: (a): Inhomogeneous spike train generated from the time-rescaling algorithm with the raster plot shown at the bottom. (b): Interspike interval (ISI) histogram for the neural spike train in (a).
To select the optimal $\sigma$, the procedure was repeated 100 times given each $\sigma$ fixed. We got 100 estimated weights, and the kernel density function of those 100 $w$’s is plotted in Figure 5.3. For $\sigma$ from 0.1 to 2, $\sigma = 5.3$ has the kernel density with mean 3, which is the true weight of the peer. So, from the estimated weight, the optimal timescale is 5.3. This $\sigma = 5.3$ is also the optimal choice by the maximal predictability criteria.

For Figure 5.4, the x axis is $\sigma$ from 0.1 to 2, and the y axis is the counts of those optimal $\sigma$’s where the predictabilities achieve the maxima. The kernel density function is skewed, and the median is 5.35. So, from both the estimated weights and predictabilities, the optimal peer prediction timescale is 5.3, and it was used to predict the target spike train in the testing data set.

We got unbiased prediction by setting $\sigma = 5.3$, as can be seen in Figure 5.5. Except for the boundaries, the prediction is close to the true rate function for both the mean curve of 100 predicted intensity functions and the mean curve of the intensity functions of the 100 spike trains in the testing set. Unsurprisingly, the variation of the predictions is bigger than that of the intensity functions from testing data.
Figure 5.3: The kernel density follows approximately normal distribution. The density of $\sigma = 5.3$ has mean 3 which is the true weight of the peer.

Figure 5.4: Kernel density function of 100 optimal $\sigma$’s in terms of maximizing the predictability.
When we have multiple peers, we cannot select the optimal $\sigma$ from estimated weights, but the maximal predictability criteria still works well. Keeping the underlying rate function of the target spike train unchanged, two unequally weighted peers are generated with weight 1 and 2 respectively. Figure 5.6 shows the kernel density of the optimal $\sigma$’s in terms of maximizing the predictability of those two peers, and we can see the median is $\sigma = 4.5$.

Finally, we can still have unbiased prediction by setting $\sigma = 4.5$ in the two peer case. The unbiasedness is shown in Figure 5.7.

### 5.1.2 Real Data Analysis

The 16 spike trains for the analysis were recorded simultaneously from noncholinergic basal forebrain neurons (Lin and Nicolelis, 2008). Let us first explore neuron 1, our target spike train, in Figure 5.8.
Figure 5.6: Kernel density function of 100 optimal \( \sigma \)’s in terms of maximizing the predictability.

Figure 5.7: The blue solid curve in the middle is the underlying firing rate function, the red solid curve is the mean curve of 100 predicted intensity functions, and the black solid curve is the mean curve of the intensity functions of the 100 spike trains in the testing set. Red and black dashed lines are the 95% boundaries for the predictions and smoothing functions respectively.
Figure 5.8: (a): The non-uniform distribution of the ticks on the raster plot shown at the bottom of figure (a) indicates a model based on an inhomogenous Poisson process for this spike train. (b): Interspike interval (ISI) histogram for the neural spike train of neuron 1.
Another spike train from neuron 5 that is positively related to the target spike train was selected as the first peer to predict the firing rate of neuron 1. Figure 5.9 is the cross-correlogram (CCG) which peaks at lags close to 0. An intuitive interpretation of the CCG is that the target neuron is more likely to fire immediately before or after the peer. We can expect an positive weight estimated between this target neuron and its peer due to the synchronous firing.

Figure 5.9: Cross-correlogram of neuron 1 (the target neuron) and neuron 5 shows a peak for lags that are less than 0.1 second.

Predictability based on log-likelihood ratio is first introduced in predicting earthquake occurrence. One unit of predictability (information bit) would mean that uncertainty of earthquake occurrence is reduced on average by a factor of 2 by using a particular model. We quantify the predictability of a spike train in the similar way, so that we can reduce the uncertainty as much as possible. Comparisons were made between predictability from one peer (neuron 5) or two peers (neuron 4 and neuron 5). Also, predictability from two positively correlated peers (neuron 4 and neuron 5)
or one of them (neuron 15) has no significant correlation with the target neuron.

Figure 5.10: Both neuron 4 and neuron 5 are positively correlated to the target neuron. Predictabilities were estimated by one peer (neuron 4) or two peers (neuron 4 and neuron 5). Prediction by those two peers save more information bits overall.

Neuron 4 has an excitatory impact on the the target neuron as well. When we estimate the rate function based on the two peers (neuron 4 and neuron 5), we can see in Figure 5.10 that the green curve overall is above the blue one which is the predictability from one peer (neuron 5) only.

What if we involve another spike train that has no strong evidence to the synchronization with the target spike train? We can see in Figure 5.11(b) that involving neuron 15, which has no correlation with the target neuron, does not improve the predictability.

In the peer prediction framework, it is possible to estimate the timescale with which neurons are coordinated into synchronization by varying the temporal bandwidth, $\sigma$. Figure 5.12 shows the predictability as a function of timescale for peer neuron 5.
Figure 5.11: (a): Cross-correlogram of neuron 1 and neuron 15. There is no significant relation between the target neuron and predicting neuron. (b): Predictability from only one peer (neuron 5) and from two peers (neuron 5 and neuron 15).
5.2 Regression Spline Model

The model in Section 4.2 defines the conditional intensity function in terms of the baseline intensity and the covariates. Omitting the covariate terms, cubic splines approximate the spontaneous firing of the individual neurons, which can be helpful for describing the variation of a neuron’s firing rate across time and then understanding neural characteristics. We can use the information to determine the similarity of the characteristics of spike trains (see Section 5.2.5). For two spike trains that have similar curves of the spontaneous firing rate, the way we model the interactive term is then simple but sufficient to represent the peer effects. Section 5.2.6 shows an example where the covariate term in model 4.5 is a linear function. The advantage of this linear model is that we can infer the association between a target and a peer directly from the coefficient. This simple application may be less suitable when the neurons are not alike, which we can tell from the baseline curves, for example, when
two neurons have dramatically different rates of firing or the resting period of one neuron is significantly longer than some others. For these cases, we need to allow more flexibility of the covariate terms; say for a certain target, more spikes from each peer will be considered in the model. Section 5.2.7 gives a real application as we just described; that section also embeds the details of the model selection using AIC to select the appropriate number of firings for each peer.

Simulation studies are conducted to validate the method in Sections 5.2.2 and 5.2.3 and to compare the results with Brillinger’s approach for the spline estimation in Section 5.2.4. Section 5.2.1 starts with the generation of a neural spike train in which we introduce a general algorithm for a given conditional intensity function. We obtain correlated spike trains linked through the covariates terms in model 4.6. We then estimate the parameters in two stages. In 5.2.2, the knot numbers and locations of the spline are fixed, and the optimization is a simple MLE problem for the fixed knot case. If the knot locations are also to be optimized, we apply the adaptive model selection scheme in 5.2.3. We observe that both the spline fit with either fixed or adaptive knots can give almost unbiased estimates, and we observe that the kernel density of the estimations is approximately normal.

5.2.1 Simulation Study I: A General Algorithm of Neural Spike Train Generation

Generating the spike train according to the log-intensity function is a challenge because the log-intensity function could be arbitrary and because calculating the probability distribution is not always feasible. We introduce an algorithm for generating spike trains from any log-intensity function.

1. We divide the entire experiment time into finite numbers of small intervals
(called bins), $\Delta$. The bins need to be sufficiently small compared to the actual ISI. For example, if the actual ISI is expected to be around 1 second, then the bin size should not be larger than 1 millisecond.

2. For any given sets of spline knots $(t_1, \ldots, t_p)$, coefficients $(\theta_1, \ldots, \theta_p)$, peer spike times, and the coefficient $\beta_1$, we can calculate the log-intensity for each bin, one at a time. Beginning at time 0, the log-intensity for the first bin is $\alpha(\Delta|v_1) = \sum_{j=1}^{p} \theta_j B_j(\Delta) + \beta_1[K - (\Delta - v_1)]\mathbb{1}_{\Delta \geq v_1}$. The first spike of the peer does not impact the log-intensity function if $v_1 > \Delta$.

3. After taking the logarithm of the log-intensity function, the probability of one event occurring within the interval $\Delta$ equals the product of the intensity function and $\Delta$ (see Equation 2.10).

4. A random number is drawn following a Bernoulli distribution and the probability is what we calculated by step 3: 0 for no occurrence and 1 for one spike occurrence in the small interval.

5. If the outcome of the Bernoulli trial fails, we move onto the next bin and the time in step 2 is accumulated, from $\Delta$ to $2\Delta$, or from $2\Delta$ to $3\Delta$, and so on, until a time at which the Bernoulli trial succeeds.

6. If the outcome of the Bernoulli trial succeeds, the time is reset back to 0, and we move onto the next available spike of peer as the covariate and repeat step 2 to 5.

This procedure offers high flexibility for generating the neural spike train under the framework that the neuron has its spontaneous firing rate and also influenced by its peers. The spike train generation for any isolated neuron, though this may not be realistic, is a special case of the general procedure where the peer effects are omitted.
However, it is still not trivial to obtain the simulated sequences with correlation as desired. Let’s see an example. The design of the example requires that the interspike interval for the target neuron contains exactly one peer firing. In other word, the target neuron is less likely to fire until the peer fires, and such impact is strong enough to trigger one target firing almost sure. The spontaneous firing rate, therefore, is set to be close to 0, and the increment of the intensity due to peer activity will raise rapidly as the lag time increases. To control the consecutive firings and make the repetitions behave similarly and stably, the ISI of the peer can be set as a constant without loss of generality.

Similar design can be easily extend to more sophisticated cases, like multiple peer effects or negative peer effects.

5.2.2 Simulation Study II: Fixed-Knot Splines

We follow the principle ideas in an “integrate-and-fire” model. Stimulus from peers in the form of current inputs are accumulative until the next firing of target cell. After that triggering, the stimulus before would have no impact on any coming events, such as resetting the clock and counting from zero. In the meantime, the voltage of the membrane naturally decays along the time so that the impact of peer firing diminishes gradually, referred to as “leaky integrate-and-fire”. Based on those neurobiological supports, we proposed a model of interspike interval (ISI) data which follows Equation 4.7 while $Q$ is a decreasing function of time.

For two interacting neural spike trains, the target spike train has $N$ spikes or $N$ ISIs, $\{u_k\}_{k=1}^N$. Then each spike of the peer belongs to an individual ISI of the target spike train and only affects the target within that interval. So, for any $u_k$, we denote the distance from the beginning of the interval to a particular spike of peer within the interval as $v_k$, which is then the covariate of the $k^{th}$ observation (see Figure 5.13).
When more than one spike of a peer occur within an ISI of target spike train, we consider only the last one as the covariate.

![Diagram of peer and target spike trains]

Figure 5.13: Target and peer spike trains are recorded simultaneously. Between two consecutive spikes of the target, the distance is our observation and $u_k$ is for the interval ending at the $k^{th}$ spike. The peer may fire within the interval, and then we record the distance from the beginning of $u_k$ to the peer firing time as our covariate $v_k$.

The conditional log-intensity function for the $k^{th}$ observation is

$$
\alpha(t|v_k) = \sum_{i=1}^{p} \theta_i B_i(t) + \beta_1 [K - (t - v_k)]1_{t \geq v_k}, \quad t \geq 0,
$$

(5.1)

where $1_{t \geq v_k}$ is an indicator function and $K$ is a constant that is deterministic for each target spike train. In the model, the target spike train keeps its spontaneous firing rate before it is stimulated by the peer firing. Therefore the covariate term is zero until $t \geq v_k$.

Following the recipe above, we simulated 100 spike trains, each of which has 650 spike times given a deterministic peer spike train. In Figure 5.14 and Figure 5.15, the estimators of $\theta_1$ to $\theta_4$ correspond to knot locations $\{0.2, 0.4, 0.6, 0.8\}$ and the estimator of $\beta_1$ are shown respectively after 100 simulations.

By setting up the conditional log-intensity function as shown in Equation 5.1, we obtain nice results with unbiased mean and less variation. Nevertheless, the beauty of our method lies in the flexibility of handling multiple peers’ influence simultaneously or multiple spiking stimulations from a single peer. In other words, in the model we proposed, function $h$ could be not only a simple linear term of one covariate as shown...
Figure 5.14: Kernel density estimations of $\theta$’s from 100 simulations. In the setup, the spline knot locations are fixed as $\{0.2, 0.4, 0.6, 0.8\}$, and only one covariate is considered in the model. For each of the four plots, the estimates have an unbiased mean and less variation.
Figure 5.15: Kernel density estimations of $\beta_1$ from 100 simulations. In the setup, the spline knot locations are fixed as $\{0.2, 0.4, 0.6, 0.8\}$, and only one covariate is considered in the model. The estimate has an unbiased mean and less variation.

Figure 5.16: Q-Q plot for the $\beta_1$ estimates. The envelop is the 95% boundary, and the estimates follow approximately a normal distribution.
in Equation 5.1 but also a summation of functions for various covariates.

So, going one step further, we add on a second linear term of another spike of peer. The conditional log-intensity function is then

\[
\alpha(t|v_k) = \sum_{i=1}^{p} \theta_i B_i(t) + \beta_1 [K - (t - v_{k}^{(1)})] 1_{t \geq v_{k}^{(1)}} + \beta_2 [K - (t - v_{k}^{(2)})] 1_{t \geq v_{k}^{(2)}}, \quad t \geq 0. \quad (5.2)
\]

We certainly can expand the model by adding more covariate terms. Equation 5.2 yields the results in Figure 5.17 and Figure 5.18.

![Figure 5.17: Kernel density estimations of \(\theta\)’s from 100 simulations. In the setup, the spline knot locations are fixed as \(\{0.2, 0.4, 0.6, 0.8\}\), and two covariates are considered in the model. The estimates have an unbiased mean and less variation.](image)

61
Figure 5.18: Kernel density estimations of $\beta_1$ and $\beta_2$ from 100 simulations. In the setup, the spline knot locations are fixed as $\{0.2, 0.4, 0.6, 0.8\}$, and two covariates are considered in the model. The estimates have an unbiased mean and less variation.

Figure 5.19: Q-Q plot for the $\beta_1$ and $\beta_2$ estimates. The envelop is the 95% boundary, and the estimates follow approximately a normal distribution.
5.2.3 Simulation Study III: Adaptive Knot Selection

As was the case in section 5.2.2, we do not fix the knot locations for the following regression so that the baseline function would be fitted by an adaptive sequence of basis. The same data sets as in section 5.2.2 are used here and yield the results through model selection. The mean curve of the baseline function from 100 simulations is plotted in Figure 5.20 along with its 95% boundary and the true function.

Figure 5.20: The black solid line is the mean curve of the baseline intensity function from 100 simulations, the black dashed line is its 95% boundary, and the red solid line is the true function. The knot locations for the estimates are adaptively selected as described, and two covariates are considered in the model. The estimates have an unbiased mean and less variation.
Figure 5.21: Kernel density estimations of $\beta_1$ and $\beta_2$ from 100 simulations. The knot locations for the estimates are adaptively selected as described, and two covariates are considered in the model. For each of the two plots, the estimates have an unbiased mean and less variation.

Figure 5.22: Q-Q plot for the $\beta_1$ and $\beta_2$ estimates. The envelop is the 95% boundary, and the estimates follow approximately a normal distribution.
5.2.4 Simulation Study IV: Predictability Comparison with Brillinger’s GLIM Approach

We mentioned some drawbacks of Brillinger’s approach (Brillinger, 1988). In this section, we compare systematically the performance of Brillinger’s GLIM method and our proposed regression spline model.

First of all, the GLIM method has to discretize the experimental time into small sub-intervals with size $\Delta$, so it limits the program capacity. Take one of our real spike train data as an example for the problem. For $spk1$, the first 100 spikes occur in about 123 seconds. When $\Delta = 0.005s$, the time then spans to 24,639 sub-intervals, which is a huge expansion and causes computers to run out of memory easily.

Secondly, the choice of the size $\Delta$ is subject to the data, and so it is data specific. As we explained in section 3.1.3, the discrete likelihood function of a point process assumes the partition of the time to be sufficiently small so that there is at most one spike in any sub-interval. In other words, $\Delta$ must be no larger than the minimal ISI, which is the optimal choice in most cases. The smaller $\Delta$ value not only leads to computational crisis but also sensitively affects the likelihood function (we will provide more details in the simulation study).

Thirdly, for peer effects, Brillinger’s method traces back 13 bins and aggregates the counts of peer spikes within each bin. The bin size is not adjustable over the entire spike train. That is not a flexible way to handle the peer effects since the ISIs vary in length. The coverage of the 13 bins could be too much when ISI is smaller, or too few when ISI is larger.

Lastly, we conduct a simulation study to compare the results from the GLIM method and our proposed regression spline model in terms of the predictability. Here, only the spontaneous firing based on the history is considered to make a fair comparison between the two methods. Brillinger’s probability model then can be expressed
as

\[ p_t = \Phi(\mu(\gamma_t)) = \Phi(\theta_1 \gamma_t + \theta_2 \gamma_t^2 + \theta_3 \gamma_t^3 - \theta), \]

(5.3)

where \( \Phi(\cdot) \) is for the normal cumulative function and \( \gamma_t \) is the time elapsed since the last firing of the neuron. Through a probit link function, the coefficients are estimated by maximizing the likelihood function.

Each sub-interval with size \( \Delta \) is a Bernoulli trial with probability \( p_t \), and ultimately a spike train is generated as a Bernoulli process. For the simulation study, the parameters are set: \( \theta = -3 \), \( \theta_1 = 3 \), \( \theta_2 = 3 \), and \( \theta_3 = -3 \). Both Brillinger’s method and our model are applied to the generated spike train data in 100 simulations. To verify the unbiasedness of the estimation, the procedure was repeated 100 times and the estimated function \( \hat{p}_t \) was plotted along with the true function highlighted in black (Figure 5.23).

When \( \Delta \) is small enough, the conditional intensity function, which the regression spline model estimates, multiplies by \( \Delta \) gives the probability of a spike event in a small time interval \( \Delta \).

Brillinger’s method has less bias in this simulation study due to the setup; the 95% envelops indicate less variation of Brillinger’s method, especially when \( \gamma_t \) is less than 0.5 second.

According to the definition of predictability in section 4.1, the mean predictability of the Brillinger’s method is 836.9 while the mean predictability of the regression spline model is 781.33. However, the predictability of Brillinger’s method is sensitive to the sub-interval size \( \Delta \), which can be understood intuitively. The response variable will have more 0’s when the \( \Delta \) tends to be smaller. In other words, the process is more likely to fail, and so the change would be reflected by smaller likelihood. That is obviously another disadvantage of the Brillinger’s method.
Figure 5.23: The red solid lines are the mean and 95% boundary of the estimation by GLM with probit link function; The blue dashed lines are the mean and 95% boundary of the estimation by regression spline model. The black bold dashed line is the true probability function while the parameters are set, $\theta = -3$, $\theta_1 = 3$, $\theta_2 = 3$, and $\theta_3 = -3$. 
5.2.5  Real Application I: Feature Extraction Using the Baseline Intensity

The 16 neural spike trains were recorded simultaneously from noncholinergic basal forebrain neurons (Lin and Nicolelis, 2008). We analyze the data using our model in the following three levels. First, as we mentioned before, the baseline intensity function can easily capture the different firing features of various neurons in the real world. So before we start to investigate the correlations among the neurons, we explore their own characteristics individually by fitting the baseline function using adaptive knots in Section 5.2.5.

Then, we focus on how can we study the relations between the target spike train and its peers through two examples. Section 5.2.6 and Section 5.2.7 are the special cases of Model 4.6 where the polynomial function has order 1 or 2. In the linear example, the sign of the coefficients $\beta$ reflect the impacts of peers as being positive, negative, or of no significance. After standardizing by taking the log-likelihood ratio of the full model over the baseline, the absolute value of the ratio complements the use of the sign of the coefficients and also indicates the strength of the peer effects. To validate the model, we generate more spike trains based on the estimators from the real data and check the coverage probability of the resampling in Section 5.2.6.

The first example considers only the most recent one firing from a single peer. In contrast, the second example in Section 5.2.7 considers all the firings within the ISI for any peer. Then we can observe the influence of the peer activities as a dynamic process. The covariate terms are constructed by quadratic polynomial functions. Tracing back for the most recent peer firing, we then use AIC to select the appropriate number of firings from a sequence of choices starting from the most recent one. More details about the AIC selection can be found in Section 5.2.7.

As we mentioned earlier, omitting the covariate terms (which is a special usage
of our model), the baseline intensity function can help us describe the variation of a neuron’s firing rate, and we can use it to determine the similarity of spike trains. The left panel in Figure 5.24 exhibits that $spk_1$ and $spk_6$ have very similar characteristics.

![Baseline Intensity Function](image)

Figure 5.24: Baseline intensity functions. Left panel: The black solid line is for $spk_1$, and the dashed line is for $spk_6$. Right panel: The red solid line is for $spk_3$, and the black dashed line is for $spk_7$.

In contrast, the right panel in Figure 5.24 shows the opposite case. If we take $spk_3$ and $spk_7$ as an example, we can see the clear contrast. $spk_3$ fires almost five times more frequently than $spk_7$ on average. If we see that $spk_3$ fires, we can expect another firing to occur very soon, within 0.1 second. That implies that neuron 3 does not require a relatively long resting period and it can fire again shortly. In contrast, the gap times between two consecutive spikes of $spk_7$ range much more widely. The baseline intensity of $spk_7$ has a small peak that appears at 0.002s, but overall the curve is flat although not uniform of course.

We also noticed that neuron 3 is an extremely active neuron (see Figure 5.25) and has very frequent firings on average. Neuron 3 fires almost five times faster than neuron 1, and almost seven times faster than neuron 2. In Section 5.2.7, we will discuss this special neuron further.
Figure 5.25: Baseline intensity functions. The red solid line is for spk3, and the other black dashed lines are for the other 15 neural spike trains in the data set.

5.2.6 Real Application II: Linear Function for Peer Effect

Table 5.1 and Table 5.2 contain the estimation results out of the single peer model 5.1 and the multiple peers model 5.2, respectively.

When we consider only an individual spike train as peer for one particular target, then most of the 16 neurons have excitatory impacts on others. In Table 5.1, a cell in bold implies a significantly positive coefficient $\beta_1$. The rows index the neural spike trains, which are treated as targets in each round, while the columns index the peers in the single peer model, one peer at a time. For example, cell (1,2), is for the regression with spk1 as the target and spk2 as the peer. The numbers underlined imply a significantly negative estimator, and no bolding or underlining implies no significant effect.

neuron 7 exhibits some interesting behaviors when it acts as a peer. It excites
Table 5.1: The rows index the neural spike trains which are treated as targets in each round, while the columns index the peers in the single peer model, one peer at a time. A red and bold number means a significantly positive estimator, blue and underlining indicates a significantly negative estimator, and no bolding or underlining indicates no significant effect.

<table>
<thead>
<tr>
<th>spk1</th>
<th>spk2</th>
<th>spk3</th>
<th>spk4</th>
<th>spk5</th>
<th>spk6</th>
<th>spk7</th>
<th>spk8</th>
</tr>
</thead>
<tbody>
<tr>
<td>spk1</td>
<td>0.44</td>
<td>1.52</td>
<td>0.44</td>
<td>0.69</td>
<td>0.56</td>
<td>5.22e-04</td>
<td>0.47</td>
</tr>
<tr>
<td>spk2</td>
<td>0.77</td>
<td>1.76</td>
<td>0.46</td>
<td>0.83</td>
<td>0.84</td>
<td>-1.20e-04</td>
<td>0.62</td>
</tr>
<tr>
<td>spk4</td>
<td>0.69</td>
<td>0.56</td>
<td>1.46</td>
<td>0.71</td>
<td>0.68</td>
<td>-9.18e-04</td>
<td>0.60</td>
</tr>
<tr>
<td>spk5</td>
<td>0.46</td>
<td>0.37</td>
<td>1.61</td>
<td>0.41</td>
<td>0.47</td>
<td>-6.46e-04</td>
<td>0.34</td>
</tr>
<tr>
<td>spk6</td>
<td>0.29</td>
<td>0.27</td>
<td>1.46</td>
<td>0.24</td>
<td>0.38</td>
<td>5.44e-04</td>
<td>0.34</td>
</tr>
<tr>
<td>spk7</td>
<td>0.53</td>
<td>0.45</td>
<td>1.63</td>
<td>0.39</td>
<td>0.63</td>
<td>0.73</td>
<td>0.49</td>
</tr>
<tr>
<td>spk8</td>
<td>0.73</td>
<td>0.47</td>
<td>1.79</td>
<td>0.45</td>
<td>0.77</td>
<td>0.69</td>
<td>1.51e-03</td>
</tr>
<tr>
<td>spk9</td>
<td>0.42</td>
<td>0.28</td>
<td>1.44</td>
<td>0.29</td>
<td>0.50</td>
<td>0.62</td>
<td>8.38e-04</td>
</tr>
<tr>
<td>spk10</td>
<td>0.41</td>
<td>0.31</td>
<td>1.58</td>
<td>0.37</td>
<td>0.52</td>
<td>0.52</td>
<td>-5.53e-04</td>
</tr>
<tr>
<td>spk11</td>
<td>0.84</td>
<td>0.68</td>
<td>1.62</td>
<td>0.55</td>
<td>0.84</td>
<td>0.80</td>
<td>1.95e-04</td>
</tr>
<tr>
<td>spk12</td>
<td>0.41</td>
<td>0.29</td>
<td>1.24</td>
<td>0.34</td>
<td>0.43</td>
<td>0.43</td>
<td>3.35e-03</td>
</tr>
<tr>
<td>spk13</td>
<td>0.47</td>
<td>0.40</td>
<td>1.43</td>
<td>0.34</td>
<td>0.51</td>
<td>0.51</td>
<td>1.30e-03</td>
</tr>
<tr>
<td>spk14</td>
<td>1.10</td>
<td>0.67</td>
<td>1.83</td>
<td>0.61</td>
<td>1.07</td>
<td>0.86</td>
<td>6.43e-06</td>
</tr>
<tr>
<td>spk16</td>
<td>0.92</td>
<td>0.70</td>
<td>1.77</td>
<td>0.59</td>
<td>0.91</td>
<td>0.91</td>
<td>3.39e-03</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>spk9</th>
<th>spk10</th>
<th>spk11</th>
<th>spk12</th>
<th>spk13</th>
<th>spk14</th>
<th>spk15</th>
<th>spk16</th>
</tr>
</thead>
<tbody>
<tr>
<td>spk1</td>
<td>0.52</td>
<td>0.66</td>
<td>0.52</td>
<td>0.75</td>
<td>0.64</td>
<td>0.34</td>
<td>0.65</td>
</tr>
<tr>
<td>spk2</td>
<td>0.81</td>
<td>1.00</td>
<td>0.65</td>
<td>0.96</td>
<td>0.85</td>
<td>0.36</td>
<td>0.84</td>
</tr>
<tr>
<td>spk4</td>
<td>0.64</td>
<td>0.75</td>
<td>0.60</td>
<td>0.82</td>
<td>0.68</td>
<td>0.44</td>
<td>0.79</td>
</tr>
<tr>
<td>spk5</td>
<td>0.41</td>
<td>0.48</td>
<td>0.41</td>
<td>0.55</td>
<td>0.51</td>
<td>0.32</td>
<td>0.59</td>
</tr>
<tr>
<td>spk6</td>
<td>0.36</td>
<td>0.34</td>
<td>0.19</td>
<td>0.43</td>
<td>0.36</td>
<td>0.15</td>
<td>0.55</td>
</tr>
<tr>
<td>spk7</td>
<td>0.68</td>
<td>0.68</td>
<td>0.44</td>
<td>0.76</td>
<td>0.67</td>
<td>0.30</td>
<td>0.72</td>
</tr>
<tr>
<td>spk8</td>
<td>0.59</td>
<td>0.72</td>
<td>0.49</td>
<td>0.88</td>
<td>0.70</td>
<td>0.39</td>
<td>0.75</td>
</tr>
<tr>
<td>spk9</td>
<td>0.47</td>
<td>0.31</td>
<td>0.56</td>
<td>0.48</td>
<td>0.22</td>
<td>0.62</td>
<td>0.26</td>
</tr>
<tr>
<td>spk10</td>
<td>0.47</td>
<td>0.42</td>
<td>0.64</td>
<td>0.49</td>
<td>0.27</td>
<td>0.67</td>
<td>0.27</td>
</tr>
<tr>
<td>spk11</td>
<td>0.73</td>
<td>1.10</td>
<td>1.01</td>
<td>0.79</td>
<td>0.39</td>
<td>0.84</td>
<td>0.59</td>
</tr>
<tr>
<td>spk12</td>
<td>0.40</td>
<td>0.47</td>
<td>0.34</td>
<td>0.45</td>
<td>0.22</td>
<td>0.53</td>
<td>0.30</td>
</tr>
<tr>
<td>spk13</td>
<td>0.55</td>
<td>0.52</td>
<td>0.41</td>
<td>0.62</td>
<td>0.26</td>
<td>0.59</td>
<td>0.33</td>
</tr>
<tr>
<td>spk14</td>
<td>0.77</td>
<td>0.89</td>
<td>0.69</td>
<td>1.00</td>
<td>0.87</td>
<td>0.89</td>
<td>0.64</td>
</tr>
<tr>
<td>spk16</td>
<td>0.80</td>
<td>1.13</td>
<td>0.75</td>
<td>1.28</td>
<td>0.84</td>
<td>0.44</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Table 5.2: The rows index the neural spike trains which is treated as target each round, while the columns index the peers in the multiple peers model, 15 peers each round. A red and bold number means a significantly positive estimator, blue and underlining indicates a significantly negative estimator, and no bolding or underlining indicates no significant effect.

<table>
<thead>
<tr>
<th></th>
<th>spk1</th>
<th>spk2</th>
<th>spk3</th>
<th>spk4</th>
<th>spk5</th>
<th>spk6</th>
<th>spk7</th>
<th>spk8</th>
</tr>
</thead>
<tbody>
<tr>
<td>spk1</td>
<td>0.03</td>
<td>1.13</td>
<td>0.09</td>
<td>0.19</td>
<td>0.12</td>
<td>-1.49e-05</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>spk2</td>
<td>0.15</td>
<td>1.19</td>
<td>0.05</td>
<td>0.14</td>
<td>0.15</td>
<td>-1.31e-04</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>spk3</td>
<td>0.12</td>
<td>3.33e-03</td>
<td>1.39</td>
<td>0.05</td>
<td>0.06</td>
<td>-7.40e-04</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>spk4</td>
<td>0.05</td>
<td>-0.02</td>
<td>1.33</td>
<td>0.04</td>
<td>0.08</td>
<td>1.52e-04</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>spk5</td>
<td>0.04</td>
<td>0.05</td>
<td>1.35</td>
<td>0.04</td>
<td>0.01</td>
<td>0.08</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>spk6</td>
<td>0.17</td>
<td>0.01</td>
<td>1.34</td>
<td>0.05</td>
<td>0.13</td>
<td>0.10</td>
<td>8.66e-04</td>
<td></td>
</tr>
<tr>
<td>spk7</td>
<td>0.05</td>
<td>0.02</td>
<td>1.24</td>
<td>0.03</td>
<td>0.07</td>
<td>0.20</td>
<td>1.86e-04</td>
<td></td>
</tr>
<tr>
<td>spk8</td>
<td>0.07</td>
<td>0.05</td>
<td>1.30</td>
<td>0.09</td>
<td>0.12</td>
<td>0.13</td>
<td>-1.93e-04</td>
<td></td>
</tr>
<tr>
<td>spk9</td>
<td>0.11</td>
<td>0.07</td>
<td>1.17</td>
<td>0.05</td>
<td>0.09</td>
<td>0.07</td>
<td>7.65e-04</td>
<td></td>
</tr>
<tr>
<td>spk10</td>
<td>0.33</td>
<td>-0.05</td>
<td>1.37</td>
<td>0.05</td>
<td>0.24</td>
<td>-0.01</td>
<td>-3.03e-04</td>
<td></td>
</tr>
<tr>
<td>spk11</td>
<td>0.18</td>
<td>0.08</td>
<td>1.11</td>
<td>0.08</td>
<td>0.15</td>
<td>0.17</td>
<td>9.13e-04</td>
<td></td>
</tr>
<tr>
<td>spk12</td>
<td>0.08</td>
<td>0.13</td>
<td>0.09</td>
<td>0.26</td>
<td>0.18</td>
<td>0.10</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>spk13</td>
<td>0.16</td>
<td>0.27</td>
<td>0.10</td>
<td>0.25</td>
<td>0.23</td>
<td>-3.74e-03</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>spk14</td>
<td>0.10</td>
<td>0.15</td>
<td>0.07</td>
<td>0.25</td>
<td>0.10</td>
<td>0.02</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>spk15</td>
<td>0.07</td>
<td>0.08</td>
<td>0.03</td>
<td>0.13</td>
<td>0.12</td>
<td>0.04</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>spk16</td>
<td>0.14</td>
<td>0.08</td>
<td>-0.02</td>
<td>0.15</td>
<td>0.08</td>
<td>-0.01</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>spk17</td>
<td>0.11</td>
<td>0.10</td>
<td>0.09</td>
<td>0.14</td>
<td>0.12</td>
<td>-0.06</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>spk18</td>
<td>0.08</td>
<td>0.11</td>
<td>0.03</td>
<td>0.25</td>
<td>0.13</td>
<td>0.02</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>spk19</td>
<td>0.07</td>
<td>0.12</td>
<td>0.17</td>
<td>0.25</td>
<td>0.13</td>
<td>-0.01</td>
<td>-0.02</td>
<td></td>
</tr>
<tr>
<td>spk20</td>
<td>0.11</td>
<td>0.38</td>
<td>0.34</td>
<td>0.17</td>
<td>3.60e-03</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spk21</td>
<td>0.07</td>
<td>0.12</td>
<td>0.05</td>
<td>0.11</td>
<td>-0.01</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spk22</td>
<td>0.14</td>
<td>0.08</td>
<td>0.06</td>
<td>0.15</td>
<td>-0.01</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spk23</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.01</td>
<td>0.16</td>
<td>0.10</td>
<td>0.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spk24</td>
<td>0.13</td>
<td>0.24</td>
<td>0.14</td>
<td>0.48</td>
<td>0.14</td>
<td>-0.01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

neurons 1, 6, 8, 9, 12, 13 and 16; it inhibits neurons 4, 5 and 10; it barely affects neurons 2, 11 and 14.

What if we consider 15 peers at a time and look at the covariates out of the multiple peers model. Table 5.2 contains all the estimation results for each individual target spike train corresponding to its 15 peers. Again, the rows index the neural spike trains which are treated as targets in each round, while the columns index the peers in the multiple peers model, 15 peers each round. So, for example, the first row is for the regression with spk1 as the target and spk2 through spk16 as the peers.
Besides Neuron 7, there are more negative coefficients estimated under the multiple peers model. For example, cell (2,14) is positive under the single-peer model, but here turns negative. One explantation is that given all the positive stimuli from the other peers, neuron 14 prevents the peer stimulation from getting too big, which is a natural mechanism in biology.

Table 5.1 and Table 5.2 are informative in the sense that we can distinguish the positively or negatively correlated neurons. However, peers with the same sign for the coefficients cannot be ordered based on the absolute value of the coefficients. For completeness, we also perform the log-likelihood ratio test to see how the peers benefit the log likelihood and thus infer which peer may have stronger impact on the target.

For any target, the spontaneous firing is modeled according to its own history, and so treated as the baseline without any peer effects. The baseline intensity function in our model can be explicitly written as $\omega(t) = \sum_{i=1}^{p} \theta_i B_i(t)$. The likelihood of the baseline is $L_0$, and the likelihood of the complete model (Equation 5.1) is $L_1$. So the log-likelihood ratio is $-2 \ast \log\left(\frac{L_0}{L_1}\right)$. For single-peer model, the log ratios are calculated for all the target and peer combinations (see Figure 5.31); large values of the log ratio correspond to small p-values\(^1\). The cells are also colored to better visualize the peer effects; warm colors represent for large ratios. spk3 benefits the likelihood the most among all the 15 peers; in contrast, spk7 raises the likelihood the least and even sometimes reduces it.

**The Coverage Probability**

Now, let us verify the coverage probability by estimating the sampling distribution of our estimator. Consider spk1 and spk2 from our real data; spk2 is treated as the target while spk1 is the peer. The estimator is 0.772 from Table 5.1 which is the coefficient of the covariate. To estimate the precision, we repeated the step of

\[\chi^2 \text{ statistic (df = 1): 3.841 for p-value < 0.05}\]
Figure 5.26: The log-likelihood ratios for all the target and peer combinations. Warm colors represent large ratios. spk3 benefits the likelihood the most among all the 15 peers; in contrast, spk7 raises the likelihood the least.
re-sampling, generating the spike trains depending on the estimators out of the real data, and then analyzing following the procedure discussed in Section 5.2.2 and 5.2.3.

The conditional log-intensity function for the $k^{th}$ spike is

$$\alpha(t|v_k) = \sum_{i=1}^{p} \hat{\theta}_i B_i(t) + \hat{\beta}_1[K - (t - v_k)]1_{t \geq v_k}, \quad t \geq 0. \tag{5.4}$$

where $v_k$ is the most recent peer firing after the last firing of the target $\tau_{k-1}$, and $1_{t \geq v_k}$ is an indicator function, which is 0 if there is no peer firing between $\tau_{k-1}$ to $t$. $\{\hat{\theta}_i\}_{i=1}^{p}$, $\hat{\beta}_1$ and the knot locations are the estimators from the real data. For more details about generating the spike train and estimating the parameters with adaptive knots selection, see Section 5.2.2 and Section 5.2.3.

In the initial setup, we generated 100 spike trains according to the conditional intensity function 5.4. Each spike train contains 400 spikes (observations), and we computed the confidence interval for each repetition. Eventually, 86.9% of the estimated confident intervals cover the value 0.772. We then doubled the sample size a couple of times to see that the coverage converges as shown in Figure 5.2.6.

To check the confidence intervals for a given sample size, we also plotted the density of the standard errors as in Figure 5.2.8. The values spread in a relatively small range, and they roughly center at 0.131 which is the standard error we got from the real data.

5.2.7 Real Application III: Quadratic Functions for Multiple Firings of Each Peer

In this section, we extend the model 5.1 in two aspects: (1). We replace the linear function for the peer firing $v_k$, $k = 1, \ldots, n$, by a quadratic polynomial; (2). We trace back the multiple firings of one peer that happen between two consecutive firings of
Figure 5.27: The coverage proportion increases as we double the sample size, and the coverage eventually converges.

Figure 5.28: Density of the standard error. The red line is 0.131, which is the standard error we got from the real data.
the target. These extensions will make the model more flexible to the real situation, and we will show an example later.

By replacing the linear function \( \beta_1[K - (t - v_k)]1_{t \geq v_k} \) with a quadratic function \( \beta_10(t - v_k)1_{t \geq v_k} + \beta_11(t - v_k)^21_{t \geq v_k} \), we also relaxed the conditions of the preset value of \( K \), which sometimes could be unexplainable. Then for the single-peer case, the model is

\[
\alpha(t|v_k) = \sum_{i=1}^{p} \theta_i B_i(t) + \beta_10(t - v_k)1_{t \geq v_k} + \beta_11(t - v_k)^21_{t \geq v_k}.
\]

When multiple peer firings are considered, the peer effects are additive as

\[
\alpha(t|v_k) = \sum_{i=1}^{p} \theta_i B_i(t) + \beta_10(t - v_k^{(1)})1_{t \geq v_k^{(1)}} + \beta_11(t - v_k^{(1)})^21_{t \geq v_k^{(1)}}
+ \beta_20(t - v_k^{(2)})1_{t \geq v_k^{(2)}} + \beta_21(t - v_k^{(2)})^21_{t \geq v_k^{(2)}} \tag{5.5}
+ \ldots.
\]

Similar to Data Analysis I, we will discuss the single-peer firing case first. The analysis is done on the same data set as before. From Figure 5.30 we can see that the influence of \textit{spk3} on \textit{spk1} is large in magnitude but does not last long, while the other peers have weaker but longer effects. We also checked on all the other setups. When each one from \textit{spk2} to \textit{spk16} is treated as the target, \textit{spk3} always shows the strong but short-term effect which is quite different from the rest.

Due to the nature of neuron 3, we may lose some information when we take only the most recent firing of \textit{spk3} as the covariate. To illustrate an example by taking \textit{spk1} as the target and \textit{spk3} as the peer, we trace back five firings of \textit{spk3} which occur within the ISI of the target. The reason we use five firings of \textit{spk3} is out of the sequential selection according to AIC which will be explained in the coming section. If there are fewer than five firings, then some terms will be zero and so have no effect in Model 5.5. Figure 5.31 shows the estimated fit for the dynamic effect of \textit{spk3} on
Figure 5.29: Quadratic functions of peer effects. The red solid curve is for spk3, while the black curves are for the other peers.

spk1 when spk3 fires at 0.046375s, 0.110450s, 0.181725s, 0.221050s and 0.245200s (values from the real data). We can see that those five firings play different roles from the change of curve. The result provides realtime guidance for how spk3 affects the firing probability of spk1.

Model Selection Criteria for the Number of Peer Firings

For some spike trains such as spk3 in our data set which have much higher firing frequency, we may need to take multiple firings of the spike train into account rather than only the most recent one. So, we use AIC to select the appropriate number of firings from a sequence of choices for those special cases. The idea is that we trace back for a multiple number of firings starting from the most recent one within each ISI of the target, and then we obtain sequential models with $AIC = -2(\log - \text{likelihood}) + 2k$, where $k$ is the number of $\theta$’s and $\beta$’s in the model. Let us take spk3 as an example, and here spk1 is its target. From Figure 5.32, we can see that the AIC is rather flat.
Figure 5.30: Quadratic functions of peer effects. The red solid curve is for $spk_3$, and the black curves are for the other peers.
Figure 5.31: Fitted function for peer effect. The peer, spk3, fires at 0.046375s, 0.110450s, 0.181725s, 0.221050s and 0.245200s.

in the region from six and above, so the best model we choose includes five firings of spk3.

We also conducted a simulation study to validate whether this model selection can pick the appropriate number for us.

The way we generate the data indicates a high probability of the target firing after every two close firings of the peer. Without the peer effects, it barely occurs. The model is ultimately as Equation 5.5, with $\beta_{10} = 7.2$, $\beta_{11} = -6.2$, $\beta_{20} = 4.5$, $\beta_{21} = -2.3$. As we discussed earlier, the generation of the spike train is a Bernoulli process. With the setup, the spike occurrence follows our design in most cases, but there is still a small portion of data generated which may not be associated with two peer firings exactly. It could be more or less as will see from the results. Figure 5.33 shows the average curves from 25 repetitions, and both AIC and BIC are calculated for simulated data. Overall, the correct model is picked by both the criteria. We
Figure 5.32: Sequential selection for number of firings using AIC. The x-axis is the number of peer firings.

Figure 5.33: The average AIC (left panel) and BIC (right panel) curves from 25 repetitions. The x-axis is the number of peer firing.
also noticed that the difference between number 2 and 3 is not significant, and it is actually supported by the generated data. As we explained earlier, a small portion of the generated data may not exactly follow the design as our simulation.
Chapter 6

Consistency and Asymptotics for Spline Regression Model

6.1 Introduction

6.1.1 Preliminaries

Let $T$ be nonnegative random variables ranging over a compact interval $\mathcal{T}$. Without loss of generality, let $\mathcal{T} = [0, 1]$. $T$ has a distribution that depends on an $M$-dimensional vector $\mathbf{x} = (x_1, ..., x_M)$ of covariates ($M = 0$ when there are no covariates). Let $f(t|x)$ and $F(t|x)$ denote the conditional density and conditional distribution function, respectively, of $T$ given $\mathbf{x} \in \mathbb{R}^M$. Recall that in the regression spline model, $T$ is the inter-spike interval (ISI).

Let $\lambda(t|x)$ denote the intensity function, and take the logarithm $\alpha(t|x) = \log \lambda(t|x) = \mu(t) + \beta \cdot \mathbf{Q}(t|x)$, where $\mathbf{Q}(\cdot|\cdot) = [Q_1, ..., Q_P]$ is the vector of $P$ pre-determined polynomial functions depending on the covariates $\mathbf{x}$ and $P$ is fixed. The log-likelihood based on $(T, \mathbf{x})$ is given by

\[
\begin{align*}
    l &= \log \lambda(T|\mathbf{x}) - \int \lambda(u|\mathbf{x})\,du = \alpha(T|\mathbf{x}) - \int \text{ind}(T \geq u) \exp \alpha(u|\mathbf{x})\,du \\
    &= \{\mu(T) + \beta \cdot \mathbf{Q}(T|\mathbf{x})\} - \int \text{ind}(T \geq u) \exp(\mu(u) + \beta \cdot \mathbf{Q}(u|\mathbf{x}))\,du.
\end{align*}
\]
The expected log-likelihood is given by

\[ E \left( \log \lambda(T|x) - \int \text{ind}(T \geq u) \lambda(u|x) du \right) = \int \left[ \log \lambda(t|x) f(t|x) - (1 - F(t|x)) \lambda(t|x) \right] dt. \]

Let \( \tau_1, \ldots, \tau_n \) be independent random variables having distribution functions \( F(\cdot|x_i) \), and \( x_i \in \mathbb{R}^M \) denote the vector of covariates for the \( i^{th} \) individual, \( 1 \leq i \leq n \). Let \( G \) denote a linear space of polynomial spline functions on \( T \). Let function \( h(t|x_i) = g(t) + \beta \cdot Q(t|x_i) \), where \( g \in G \). The expected log-likelihood function \( \Lambda(\cdot) \) is defined by

\[ \Lambda(h) = \sum_i \int \left[ h(t|x_i) f(t|x_i) - (1 - F(t|x_i)) \exp h(t|x_i) \right] dt. \]

Observe that \( \Lambda(\cdot) \) is maximized at \( \alpha = \log(f/1 - F) = \mu + \beta \cdot Q \). \( \mu(t) \) may or may not be in \( G \), but we can define the best approximation to \( \mu \), \( \mu^* \in G \) so that \( \alpha^* = \mu^* + \beta \cdot Q \) that maximizes \( \Lambda(\cdot) \) over \( G \). \( G \) has dimension \( J \).

The first goal is to prove that \( \Lambda(\cdot) \) has a maximum in \( G \). Suppose that vectors \( x_1, \ldots, x_n \) of covariates take values in a compact interval \( \mathcal{X} \subset \mathbb{R}^M \). Let \( T \) denote a compact interval of the form \([0, \tau]\) for some positive number \( \tau \). Without loss of generality, we assume that \( T = [0, 1] \) and \( \mathcal{X} = [0, 1]^M \). We will prove the existence of the best approximation in \( G \).

**Condition 1.** The density function \( f \) is bounded away from zero and infinity on \( T \times \mathcal{X} \). Moreover, the distribution \( F \) is bounded away from 1 on \( T \times \mathcal{X} \).

This condition implies that \( |\alpha| \) is bounded away from infinity on \( T \times \mathcal{X} \).

**Condition 2.** \( |Q(\cdot, \cdot)| \) is the pre-determined polynomial functions, which are all bounded away from infinity on \( T \times \mathcal{X} \).

So, condition 1-2 imply that \( |\mu| \) is bounded away from infinity.
6.1.2 Polynomial Splines

The best approximation will be chosen from the linear space $G$ of polynomial splines. Specifically, let $K = K_n$ be a positive integer and let $I_k$, $1 \leq k \leq K$, denote the subintervals of $[0, 1]$ defined by $I_k = [(k-1)/K, k/K]$ for $1 \leq k \leq K$ and $I_K = [1-1/K, 1]$. Let $m$ and $q$ be fixed integers such that $m \geq 0$ and $m \geq q \geq -1$.

Let $S$ denote the space of functions $s$ on $[0,1]$ such that

(i) the restriction of $s$ to $I_k$ is a polynomial of degree $m$ (or less) for $1 \leq k \leq K$;

and, if $q \geq 0$, then

(ii) $s$ is $q$-times continuously differentiable on $[0,1]$.

A function satisfying (i) is called a piecewise polynomial, and it is called a spline if it satisfies both (i) and (ii). Let $B_j$, $1 \leq j \leq J$, denote the usual basis of $S$ consisting of $B$-splines (de Boor, 1978). Then $J = (m+1)K - (q-1)(K-1)$, so $K + m \leq J \leq (m+1)K$. Also, $B_j > 0$ on $[0,1]$, $B_j = 0$ on the complement of an interval of length $(m+1)/K$ for $1 \leq j \leq J$, and $\sum_j B_j = 1$ on $[0,1]$. Moreover, for $1 \leq j \leq J$, there are at most $2m+1$ values of $j' \in \{1, ..., J\}$ such that $B_j B_{j'}$ is not identically zero on $[0,1]$. Set $\theta = (\theta_1, ..., \theta_J) \in \mathbb{R}^J$ and let $|\theta| = (\sum_j \theta_j^2)^{1/2}$ denote the Euclidean norm of $\theta$. According to Theorem 4.2 of DeVore and Lorentz (1993), (6.1) holds.

There is a positive constant $M_0$ such that

$$M_0^{-1}J^{-1}|\theta|^2 \leq \int |\theta \cdot B|^2 \leq M_0 J^{-1}|\theta|^2, \quad \theta \in \Theta := \mathbb{R}^J. \quad (6.1)$$

6.2 Spline Approximation

Under Condition 1-2, there exists an essentially uniquely determined function $\mu^* \in G$ such that $\Lambda(\mu^*) = \max_{g \in G} \Lambda(g)$. Moreover, if $\mu \in G$, then $\mu^* = \mu$ almost
everywhere (Kooperberg et al., 1995). Let $\theta^*$ denote the vector of parameters that is associated with $\mu^*$ and $\phi^* = [\theta^*, \beta]$, $U = [B(t), Q(t|x)]$, so that $\mu^*(t|x) = \theta^* \cdot B(t)$, $\alpha^*(t|x) = \log \lambda^*(t|x) = \theta^* \cdot B(t) + \beta \cdot Q(u|x) = \phi^* \cdot U$. These are referred to as spline approximations.

The errors resulting from spline approximation will be quantified in terms of a smoothness condition that will now be described. Let $0 < \delta \leq 1$, a function $g$ on $T$ is said to satisfy a Hölder condition with exponent $\delta$ if there is a positive number $\gamma$ such that $|g(z) - g(z_0)| \leq \gamma |z - z_0|^{\delta}$ for $z, z_0 \in T$. Let $m$ be a nonnegative integer and set $p = m + \delta$. A function $g$ on $T$ is said to be $p$-smooth if it is $m$ times continuously differentiable on $T$ and $g^{(m)}$ satisfies a Hölder condition with exponent $\delta$. The following smoothness condition will be used to describe the errors resulting from spline approximation.

**Condition 3.** $\mu$ is a $p$-smooth function with $p > \frac{1}{2}$.

We do not assume that $\mu$ is exactly equal to a spline, but we still can make use of spline approximation. In order for this method to be accurate, we need the error of approximation to tend to zero as the sample size $n$ tends to infinity; for this, it is necessary that the dimension $J$ of the approximation space $G$ tend to infinity. To control the error of estimation we need this dimension to increase more slowly than $n^{1/2}$.

**Condition 4.** $J = J_n \to \infty$ and $J^2 = o(n^{1-\epsilon})$ for some $\epsilon > 0$.

For a real-valued function $h$ on $T \times \mathcal{X}$, set $\|h\|_2 = \left[ \int_{T \times \mathcal{X}} |h(t|x)|^2 \right]^{1/2}$ and $\|h\|_\infty = \sup_{T \times \mathcal{X}} |h(\cdot|\cdot)|$. Also, set $\rho = \inf_{g \in G} \|g - \mu\|_\infty$. Our first result gives the error bounds for the spline approximation.

**Theorem 2.** Under Conditions 1-4,

$$\|\mu^* - \mu\|_\infty = O(\rho),$$  (6.2)
Proof of Theorem 2

Let $M_1, M_2, ...$ denote constants greater than 1. According to Condition 1 and 2,

$$M_1^{-1} \leq \exp(\beta \cdot Q(t|x_i))(1 - F(t|x_i)) \leq M_1, \quad (t, x) \in \mathcal{T} \times \mathcal{X}. \quad (6.7)$$

Let $\mathcal{A}$ denote a collection of functions $\mu$ on $\mathcal{T}$ satisfying the Hölder condition

$$|\mu(z) - \mu(z_0)| \leq \gamma |z - z_0|^\delta, \quad z, z_0 \in \mathcal{T}, \quad (6.8)$$

and the boundedness condition

$$||\mu||_\infty \leq M_2, \quad \mu \in \mathcal{A}. \quad (6.9)$$

Note that if $\mu \in \mathcal{A}$ and $0 \leq u < 1$, then $u\mu \in \mathcal{A}$. Set $\rho = \rho(\mu) = \inf_{g \in G} ||\mu - g||_\infty$, $\mu \in \mathcal{A}$, and note that $\rho(\mu) \leq M_2$ for $\mu \in \mathcal{A}$. Writing $\mu^*$ as $Q\mu$ and following the argument in Stone (1989), we will obtain an inequality of the form, which will be verified later,

$$||\mu - Q\mu||_\infty \leq M \rho(\mu), \quad \mu \in \mathcal{A}, \quad (6.10)$$

where the positive constant $M$ depends on $\mathcal{A}$ and the degree $m$ of $G$, but not on the dimension $J$ of $G$. We conclude from 6.10 that (6.2) holds. Then (6.3) is trivial since

$$\alpha^* - \alpha = \mu^* + \beta \cdot Q - (\mu + \beta \cdot Q) = \mu^* - \mu.$$
Equation (6.4) follows from (6.3) and the following.

\[ \lambda^* - \lambda = \exp \phi^* - \exp \phi = (\phi^* - \phi) \int_0^1 \exp(\phi + u(\phi^* - \phi)) du. \]  
(6.11)

The proof of (6.5) follows from (6.4) and the following.

Let \( V = \int_0^t \lambda du \) and \( V^* = \int_0^t \lambda^* du \), so \( V - V^* = \int_0^t (\lambda - \lambda^*) du \).

\[ F - F^* = 1 - \exp(-V) - (1 - \exp(-V^*)) \]
\[ = -(V^* - V) \int_0^1 \exp(-V - u(V^* - V)) du. \]

Observe that \( f^* - f = \lambda^*(F - F^*) + (\lambda - \lambda^*)F \). We conclude from (6.4) and (6.5) that (6.6) holds.

Now we start to verify (6.10). Let \( \psi \) be a function on \( T \) such that

\[ M_{\Delta}^{-1} \leq \psi(t) \leq M_{\Delta}, \quad t \in T. \]  
(6.12)

Consider the \( J \times J \) matrix \( \mathcal{M} \) whose \((j, l)^{th}\) entry is \( \sum_i \int B_j(t) B_l(t) \psi(t) dt \) for \( j \) and \( l \) range over \( A \). It follows from (6.1) that \( \mathcal{M} \) is invertible. Let \( \gamma_{jl} \) denote the \((j, l)^{th}\) entry of \( \mathcal{M}^{-1} \). Then \( ||\mathcal{M}^{-1}||_\infty = \max_{1 \leq j, l \leq J} \sum_l |\gamma_{jl}| \). By a slight extension of a result in de Boor (1976), \( ||\mathcal{M}^{-1}||_\infty \leq M_4 n^{-n} J \). This has the following consequence.

**Lemma 1.** Set \( g = \sum_j \theta_j B_j \), then

\[ \max_j |\theta_j| \leq M_4 n^{-1} J \max_j \left| \sum_i \int g(t) B_j(t) \psi(t) dt \right|. \]  
(6.13)
Choose \( \mu \in A \) and \( g \in G \). Since

\[
\sum_i \int \{ [ug(t) + Q\mu(t) + \beta \cdot Q(t|x_i)] \exp(\mu(t) + \beta \cdot Q(t|x_i)) - \exp(ug(t) + Q\mu(t) + \beta \cdot Q(t|x_i)) \} \\
(1 - F(t|x_i)) dt
\]

is maximized at \( u = 0 \),

\[
\sum_i \int g(t) [\exp \mu(t) - \exp Q\mu(t)] \exp(\beta \cdot Q(t|x_i))(1 - F(t|x_i)) dt = 0.
\]

Consequently, for \( j \in A \),

\[
\sum_i \int B_j(t) [\exp \mu(t) - \exp Q\mu(t)] \exp(\beta \cdot Q(t|x_i))(1 - F(t|x_i)) dt = 0. \tag{6.14}
\]

Let \( \mu \in A \). Then there is an \( \bar{\mu} \in G \) such that \( ||\mu - \bar{\mu}||_\infty = \rho(\mu) \). Note that \( Q\bar{\mu} = \bar{\mu} \). Note also that \( ||\bar{\mu}||_\infty \leq 2M_2 \) and hence that \( \exp(-2M_2) \leq \exp \bar{\mu} \leq (2M_2) \) and

\[
||\exp \bar{\mu} - \exp \mu||_\infty \leq \exp(2M_2)\rho(\mu). \tag{6.15}
\]

By (6.7), (6.1), (6.14), and (6.15),

\[
| \sum_i \int B_j(t) [\exp \mu(t) - \exp Q\mu(t)] \exp(\beta \cdot Q(t|x_i))(1 - F(t|x_i)) dt |
\leq M_1M_5nI^{-1}\exp(2M_2)\rho(\mu), \quad j \in A. \tag{6.16}
\]

Write \( Q\mu - \bar{\mu} = \sum_j \theta_j B_j \) and set \( \epsilon = \max_j |\theta_j| \). Now \( ||Q\mu - \bar{\mu}||_\infty \leq \epsilon \) and hence

\[
||\mu - Q\mu||_\infty \leq \epsilon + \rho(\mu). \tag{6.17}
\]

By repeatedly applying (viii) on Page 155 of de Boor (1978), there is a positive
constant $M_6$, such that
\[
\epsilon \leq M_6 \| Q\mu - \bar{\mu} \|_{\infty}.
\]  
(6.18)

Since $\exp Q\mu = \exp \bar{\mu} \exp(\sum_j \theta_j B_j)$,
\[
\| \exp Q\mu - \exp \bar{\mu} - \exp \bar{\mu} \exp(\sum_j \theta_j B_j) \|_{\infty} \leq \exp(2M_2 + \epsilon) \frac{\epsilon^2}{2}.
\]

We now conclude from (6.1) and (6.16) that, for $j \in A$,
\[
\left| \sum_i \int B_j(t) \sum_l \theta_l B_l(t) \exp \bar{\mu}(t) \exp(\beta \cdot Q(t|x_i))(1 - F(t|x_i)) dt \right|
\leq M_1 M_5 n I^{-1} \exp(2M_2)(\rho(\mu) + \exp(\epsilon) \frac{\epsilon^2}{2}).
\]

According to (6.12), (6.16) and Lemma (1) applied to $\psi = \exp(\beta \cdot Q)(1 - F) \exp \bar{\mu}$, with $M_3 = M_1 \exp(2M_2)$,
\[
\epsilon \leq M_1 M_4 M_5 \exp(2M_2) \left( \rho(\mu) + \exp(\epsilon) \frac{\epsilon^2}{2} \right)
\]  
(6.19)

Suppose now that
\[
M_1 M_4 M_5 \exp(2M_2 + \epsilon) \epsilon \leq 1.
\]  
(6.20)

Then $\epsilon \leq 2M_1 M_4 M_5 \exp(2M_2) \rho(\mu)$ and hence, by (6.17),
\[
\| \mu - Q\mu \|_{\infty} \leq M_7 \rho(\mu),
\]  
(6.21)

where $M_7 = 2[M_1 M_4 M_5 \exp(2M_2) + 1]$. According to (6.18), a sufficient condition for (6.20) and hence for (6.21) is
\[
\| Q\mu - \bar{\mu} \|_{\infty} \leq M_8^{-1}.
\]  
(6.22)
Let $0 < \rho_0 < 2^{-1}M_7^{-1}M_8^{-1}$. There is a positive integer $I_0$, depending on $M_1$ and the degree of $G$, such that

$$\rho(\mu) \leq \rho_0, \quad I \geq I_0 \quad \text{and} \quad \mu \in \mathcal{A}, \quad (6.23)$$

see Theorem 12.8 of Schumaker (1981). Let $I \geq I_0$. Suppose that

$$||\mu - Q\mu||_\infty \leq 2^{-1}M_8^{-1}. \quad (6.24)$$

Since $||\mu - \bar{\mu}||_\infty = \rho(\mu) \leq 2^{-1}M_8^{-1}$, (6.22) holds.

We will verify that (6.24) necessarily holds for $I \geq I_0$. Suppose not. Now $||u\mu - Q(u\mu)||_\infty$ is continuous in $u$ for $0 \leq u < 1$ (since the expected log-likelihood is a strictly concave function of $\theta_1, ..., \theta_I$ and it is continuous in $u, \theta_1, ..., \theta_I$) and it approaches 0 as $u \to 0$. Thus there is a value of $u \in (0, 1)$ such that $||u\mu - Q(u\mu)||_\infty = 2^{-1}M_8^{-1}$. By the previous argument, (6.21) and (6.23) hold with $\mu$ replaced by $u\mu$; hence

$$||u\mu - Q(u\mu)||_\infty \leq M_7\rho(u\mu) \leq M_7\rho_0 < 2^{-1}M_8^{-1},$$

which yields a contradiction.

We have now shown that

$$||\mu - Q\mu||_\infty \leq M_8\rho(\mu), \quad I \geq I_0 \quad \text{and} \quad \mu \in \mathcal{A}. \quad (6.25)$$

To complete the proof of (6.10), we need to show that

$$||\bar{\mu} - Q\mu||_\infty \leq M_9\rho(\mu), \quad I < I_0 \quad \text{and} \quad \mu \in \mathcal{A}. \quad (6.26)$$

But this result, for each $I$, follows in a straightforward manner by a compactness
6.3 Maximum Likelihood Estimation

The log-likelihood function corresponds to the data $(\tau_1, x_1), ..., (\tau_n, x_n)$

\[
    l(h) = \sum_i h(\tau_i|x_i; \phi) - \sum_i \int_0^{\tau_i} \exp h(u|x_i; \phi) du,
\]

(6.25)

where $\phi = [\theta, \beta]$, $\phi \in \Phi$ and $h(\cdot; \phi) = g(\cdot; \theta) + \beta \cdot Q(\cdot; \beta)$, $g \in G$.

Under Condition 1-4, the log-likelihood function $l$ is strictly concave and hence there exists a unique maximum likelihood estimation $\hat{h} = \hat{\theta} \cdot B + \hat{\beta} \cdot Q$ so that $\hat{\phi} = \max_{\phi \in \Phi} l(\phi)$. The maximum likelihood estimates of the log intensity, intensity functions are given, respectively, by $\hat{\alpha}(t|x) = \hat{\theta} \cdot B(t) + \hat{\beta} \cdot Q(t|x)$ and $\hat{\lambda}(t|x) = \exp \hat{\alpha}(t|x)$. The next result bounds the $L_2$ and $L_\infty$ norms of the error of the estimates.

Denote the score function $S$ at $\phi$ as

\[
    S(\phi) = \frac{\partial}{\partial \phi} l(h),
\]

(6.26)

which is the $I$-dimensional vector, $I = J + P$, having entries

\[
    \frac{\partial l(h)}{\partial \theta_j} = \sum_i B_j(\tau_i) - \sum_i \int_0^{\tau_i} B_j(u) \exp h(u|x_i, \phi) du, \quad j = 1, ..., J
\]

\[
    \frac{\partial l(h)}{\partial \beta_p} = \sum_i Q_p(\tau_i|x_i) - \sum_i \int_0^{\tau_i} Q_p(u|x_i) \exp h(u|x_i, \phi) du \quad p = 1, ..., P.
\]

(6.27)

Recall the notation $\phi = [\theta, \beta]$ and $U = [B, Q]$.

\[
    U_k = \begin{cases} 
        B_k(t) & \text{when } k = 1, ..., J \\
        Q_k(t|x) & \text{otherwise}
    \end{cases}
\]
So the entries can be rewritten as

\[
\frac{\partial l(h)}{\partial \phi_k} = \sum_i U_k - \sum_i \int_0^{r_i} U_k \exp h(u|x_i, \phi) du.
\]

Denote the Hessian matrix of \( l(h) \) as

\[
\frac{\partial^2}{\partial \phi \partial \phi^T} l(h),
\]

which is the \( I \times I \) matrix having entries

\[
\frac{\partial^2 l(h)}{\partial \theta_j \partial \theta_k} = -\sum_i \int_0^{r_i} B_j(u) B_k(u) \exp h(u|x_i, \phi) du \quad j = 1, \ldots, J; k = 1, \ldots, J,
\]

\[
\frac{\partial^2 l(h)}{\partial \theta_j \partial \beta_p} = -\sum_i \int_0^{r_i} B_j(u) Q_p(u|x_i) \exp h(u|x_i, \phi) du \quad j = 1, \ldots, J; p = 1, \ldots, P,
\]

\[
\frac{\partial^2 l(h)}{\partial \beta_p \partial \beta_q} = -\sum_i \int_0^{r_i} Q_p(u|x_i) Q_q(u|x_i) \exp h(u|x_i, \phi) du, \quad p = 1, \ldots, P; q = 1, \ldots, P.
\]

(6.29)

Similarly, we can also write as

\[
\frac{\partial^2 l(h)}{\partial \phi \partial \phi^T} = -\sum_i \int_0^{r_i} U_k U_q \exp h(u|x_i, \phi) du.
\]

(6.30)

This can further be written as

\[
D(\hat{\phi} - \phi^*) = -S^*, \quad \text{where } D \text{ is the } I \times I \text{ matrix given by }
\]

\[
D = \int_0^1 \frac{\partial^2}{\partial \phi \partial \phi^T} l(\phi^* + u(\hat{\phi} - \phi^*)) du.
\]

(6.31)
Theorem 3. Under Conditions 1-4,

\[ |\hat{\phi} - \phi^*| = O_p(I/\sqrt{n}), \]  
\( (6.32) \)

\[ ||\hat{\alpha} - \alpha^*||_2 = O_p(\sqrt{I/n}), \]  
\( (6.33) \)

\[ ||\hat{\lambda} - \lambda^*||_2 = O_p(\sqrt{I/n}), \]  
\( (6.34) \)

\[ \max_{1 \leq j \leq J} |\hat{\phi}_j - \phi^*_j|^2 = O_p(n^{-1}I \log I), \]  
\( (6.35) \)

\[ \max_{t \in T, x \in X} |\hat{\alpha}(t|x) - \alpha^*(t|x)| = O_p(\sqrt{n^{-1}I \log I}), \]  
\( (6.36) \)

\[ \max_{t \in T, x \in X} |\hat{\lambda}(t|x) - \lambda^*(t|x)| = O_p(\sqrt{n^{-1}I \log I}), \]  
\( (6.37) \)

Proof of Theorem 3

(Proof of (6.32) and (6.33)) It follows from the maximum likelihood equation that

\[ (\hat{\phi} - \phi^*)^T D(\hat{\phi} - \phi^*) = -(\hat{\phi} - \phi^*)^T S^*. \]  
\( (6.38) \)

By the definition of \( \phi^* \), we have

\[ E(\frac{\partial l(\alpha^*)}{\partial \phi_j}) = 0. \]

Hence

\[ E(\frac{\partial l(\alpha^*)}{\partial \phi_j})^2 = \text{var}(\sum_{i} U_j - \sum_{i} \int_{0}^{\tau_i} U_j \exp h). \]

Since

\[ \sum_{j} \text{var}(U_j \sum_{i} \int_{0}^{\tau_i} U_j \exp h) = O(1), \]
we conclude that $E[S(\phi^*)]^2$, and by Markov inequality, $\exists \epsilon$ s.t.

$$\text{Prob}(\frac{|S^*|^2}{n} > M) \leq \frac{E|S^*|^2}{nM} < \epsilon, \quad \text{for any } M,$$

so

$$|S^*|^2 = O_p(n) \quad (6.39)$$

Also according to (3.6) of Kooperberg et al. (1995b), there is a positive constant $M_1$ such that

$$(\hat{\phi} - \phi^*)^T D(\hat{\phi} - \phi^*) \leq -M_1 n I^{-1} |\hat{\phi} - \phi^*|^2 \quad (6.40)$$

except on an event whose probability tends to zero with $n$. Since

$$|((\hat{\phi} - \phi^*)^T S^*)| \leq |\hat{\phi} - \phi^*||S^*|,$$

it follows from (6.38)- (6.40) that $|\hat{\phi} - \phi^*|^2 = O_p(I^2/n)$. According to Cauchy-Schwarz inequality, $||\hat{\alpha} - \alpha^*||^2 = \int |(\hat{\phi} - \phi^*) \cdot U|^2 \leq \int |\hat{\phi} - \phi^*|^2 \int |U|^2$ and hence that

$$||\hat{\alpha} - \alpha^*||^2 = O_p(I/n).$$

This completes the proofs of (6.32) and (6.33).

(Proof of (6.34)) According to (6.33), Lemma 2 of Kooperberg et al. (1995b) and Condition 4, $||\hat{\alpha} - \alpha^*||_\infty = O_p(I/\sqrt{n}) = o_p(1)$. The desired result follows from (6.3), (6.33) and

$$\hat{\lambda} - \lambda^* = \exp \hat{\alpha} - \exp \alpha^* = (\hat{\alpha} - \alpha^*) \int_0^1 \exp(\alpha^* + u(\hat{\alpha} - \alpha^*))du. \quad (6.41)$$

(Proof of (6.35)) The proof of (6.35) requires a sequence of lemmas. Set $\lambda_{\min}(\phi) = \min[\exp h(\cdot; \phi)]$ and $\lambda_{\max}(\phi) = \max[\exp h(\cdot; \phi)]$. 

95
Let \( I = I(\phi) \) denote the \( I \times I \) information matrix, which has entries

\[
-E(\frac{\partial^2 l(h)}{\partial \theta_j \partial \theta_k}) = \sum_i E \left( \int_0^{\tau_i} B_j(u)B_k(u) \exp(h(u|x_i; \phi)) du \right), \quad j = 1, \ldots, J; \quad k = 1, \ldots, J,
\]

\[
-E(\frac{\partial^2 l(h)}{\partial \theta_j \partial \beta_p}) = \sum_i E \left( \int_0^{\tau_i} B_j(u)Q_p(u|x_i) \exp(h(u|x_i; \phi)) du \right), \quad j = 1, \ldots, J; \quad p = 1, \ldots, P,
\]

and

\[
-E(\frac{\partial^2 l(h)}{\partial \beta_p \partial \beta_q}) = \sum_i E \left( \int_0^{\tau_i} Q_p(u|x_i)Q_q(u|x_i) \exp(h(u|x_i; \phi)) du \right), \quad p = 1, \ldots, P; \quad q = 1, \ldots, P.
\]

**Lemma 2.** There is a positive constant \( M_2 \) such that

\[
M_2^{-1} \lambda_{\min}(\phi)nI^{-1}|\tau|^2 \leq \tau^T I(\phi)\tau \leq M_2 \lambda_{\max}(\phi)nI^{-1}|\tau|^2 \quad (6.42)
\]

for \( \phi, \tau \in \Theta \) and \( n \geq 1 \). Moreover,

\[
[M_2 \lambda_{\max}(\phi)]^{-1}n^{-1}I|\tau|^2 \leq \tau^T [I(\phi)]^{-1} \tau \leq M_2 \lambda_{\min}(\phi)^{-1}n^{-1}I|\tau|^2 \quad (6.43)
\]

and

\[
[M_2 \lambda_{\max}(\phi)]^{-1}n^{-1}I|\tau| \leq ||[I(\phi)]^{-1} \tau| \leq M_2 \lambda_{\min}(\phi)^{-1}n^{-1}I|\tau| \quad (6.44)
\]

for \( n \geq 1 \) and \( \phi, \tau \in \Phi \) such that \( \lambda_{\min}(\phi) > 0 \).

**Proof.** Since the boundedness of the polynomial functions \( Q_1, \ldots, Q_P \) on \( T \times X \), and the number \( P \) is fixed, we can find an upper bound \( N_Q \) so that \( \sum_p (Q_p/N_Q) \leq 1 \), then conclude from (6.1) that (6.45) holds.

Let dimension \( I = J + P \) and \( \phi = [\theta, \beta] \). There is a positive constant \( M_0 \) such that

\[
M_0^{-1} I^{-1} |\phi|^2 \leq |\phi \cdot U|^2 \leq M_0 I^{-1} |\phi|^2, \quad \phi \in \Phi := \mathbb{R}^I. \quad (6.45)
\]
From the definition of Fisher information matrix, we have

$$
\tau' I(\phi) \tau = \sum_i \int_0^1 h^2(u|x_i; \tau)(1 - F(u|x_i)) \exp h(u|x_i; \phi) du.
$$

We conclude from Condition 1, 3 and 4 that (6.42) holds. If \( \lambda_{\text{min}}(\phi) > 0 \), then it follows from (6.42) that there is a nonsingular symmetric matrix \( R(\phi) \) such that \( I(\phi) = R(\phi) R(\phi) \). Also,

$$
[M_2 \lambda_{\text{max}}(\phi)]^{-1} n^{-1} I \leq \frac{|\tau|^2}{\tau' I(\phi) \tau} \leq M_2 [\lambda_{\text{min}}(\phi)]^{-1} n^{-1} I, \quad \tau \in \Theta.
$$

Replacing \( \tau \) by \( [R(\phi)]^{-1} \tau \), we conclude that (6.43) is valid.

Similarly, it follows from (6.43) applied to \( \tau \) and \( [R(\phi)]^{-1} \tau \) that (6.44) is valid. \( \square \)

**Lemma 3.** Let \( \psi(\cdot) \) and \( s(\cdot) \) denote piecewise smooth functions on \([0, 1]\). Set

$$
w(y) = s(y) - \int_0^y s(u) \psi(u) du + \int_0^1 s(u) \psi(u) du = s(y) + \int_y^1 s(u) \psi(u) du.
$$

Then

$$
\int_0^1 s^2(y) dy = O \left( \int_0^1 w^2(y) dy \right).
$$

**Proof.** We have

$$
ds(y) - s(y) \psi(y) = dw(y),
$$

$$
d \left( s(y) \exp \int_y^1 \psi(u) du \right) = \left( \exp \int_y^1 \psi(u) du \right) dw(y),
$$

$$
s(1) - s(y) \exp \int_y^1 \psi(u) du = \int_y^1 \left( \exp \int_u^1 \psi(t) dt \right) dw(u),
$$

$$
w(1) - s(y) \exp \int_y^1 \psi(u) du = \int_y^1 \left( \exp \int_u^1 \psi(t) dt \right) dw(u).
$$
[Since \( w(1) = s(1) \).] Thus

\[
s(y) \exp \int_y^1 \psi(u) du = w(1) - \int_y^1 \left( \exp \int_u^1 \psi(t) dt \right) dw(u)
\]

\[= w(1) - \left. \left[ w(u) \exp \int_u^1 \psi(t) dt \right] \right|_y^1
\]

\[= w(y) \left( \exp \int_y^1 \psi(u) du \right)
\]

\[= w(y) \left( \psi(u) \left( \exp \int_u^1 \psi(t) dt \right) du \right).
\]

Set \( \Psi(u, y) = \psi(u) \exp \int_u^y \psi(t) dt \). Then

\[
s(y) = w(y) - \exp \left( -\int_y^1 \psi(u) du \right) \int_y^1 w(u) \psi(u) \left( \exp \int_u^1 \psi(t) dt \right) du
\]

\[= w(y) - \int_y^1 w(u) \Psi(u, y) du.
\]

Thus

\[
\int_0^1 s^2(y) dy \leq 2 \int_0^1 w^2(y) dy + 2 \int_0^1 \left( \int_y^1 w(u) \Psi(u, y) du \right)^2 dy
\]

\[\leq 2 \int_0^1 w^2(y) dy + 2 \int_0^1 w^2(u) du \int_0^1 \Psi^2(u, y) dudy
\]

\[= O \left( \int_0^1 w^2(y) dy \right).
\]

This completes the proof of Lemma (3). \( \square \)

Let \( \mathbf{G}^*(y, x) = \mathbf{G}(y, x; \phi^*) = [G_j(y, x; \phi^*)] \) denote the \( I \)-dimensional vectors with the \( j \)-th entry given by

\[
G_j(y, x; \phi^*) = B_j(y) - \int_0^y B_j(u) \exp h(u | x; \phi^*) du, \quad j = 1, \ldots, J,
\]
or the $p$-th entry given by

$$G_p(y, x; \phi^*) = Q_p(y, x) - \int_0^y Q_p(u, x) \exp h(u|x; \phi^*) du, \quad p = 1, \ldots, P,$$

where $x = (x_1, \ldots, x_M)$. It follows from (6.4) and the basis properties of B-spline that

$$\max_{1 \leq j \leq J} \sup_{y, x} \left| \int_0^y B_j(u) \exp h(u|x; \phi^*) du \right| = O(J^{-1}), \quad (6.46)$$

and so

$$\sum_j |B_j(y) - \int_0^y B_j(u) \exp h(u|x; \phi^*) du|^2 = O(1) \quad uniformly \quad in \quad (y, x) \in T \times \mathcal{X}; \quad (6.47)$$

Also, since

$$\left| \int_0^y Q_p(u, x) \exp h(u|x; \phi^*) du \right| \leq \max_{1 \leq p \leq P} Q_p(y, x) \left| \int_0^y \exp h(u|x; \phi^*) du \right| = M_3 \left| \int_0^y \exp h(u|x; \phi^*) du \right|, \quad (6.48)$$

so that

$$\sum_p |Q_p(y, x) - \int_0^y Q_p(u, x) \exp h(u|x; \phi^*) du|^2 = O(1) \quad uniformly \quad in \quad (y, x) \in T \times \mathcal{X}. \quad (6.49)$$

We conclude from (6.47) and (6.49) that

$$|G^*(y, x)| = O(1) \quad uniformly \quad in \quad (y, x) \in T \times \mathcal{X}. \quad (6.50)$$

Note that $S^* = \sum_i G^*(Y_i, x_i)$ and $E(S^*) = 0$. Let $VC(S^*)$ denote the variance-covariance matrix of $S^*$. 

99
Lemma 4. There is a positive constant $M_4$ such that

$$M_4^{-1}nI^{-1}|\tau|^2 \leq \tau'VC(S^*)\tau \leq M_3nI^{-1}|\tau|^2, \quad \tau \in \Theta, \quad n \geq 1.$$  

Proof. Note that

$$\tau'VC(S^*)\tau = \text{var}(\tau'S^*)$$

$$= \sum_i \text{var} \left( h(Y_i|x_i; \tau) - \int_0^{Y_i} h(u|x_i; \tau) \exp h(u|x_i; \phi^*) du \right).$$

By (6.1),

$$\sum_i E[h^2(Y_i|x_i; \tau)] = O(I^{-1}|\tau|^2) \quad (6.51)$$

and

$$\sum_i E \left[ \left( \int_0^{Y_i} h(u|x_i; \tau) \exp h(u|x_i; \phi^*) du \right)^2 \right] = O(I^{-1}|\tau|^2). \quad (6.52)$$

It follows from (6.51) and (6.52) that the upper bound holds.

Set $s(y|x) = h(y|x; \tau)$, $\psi(y|x) = h(y|x; \phi^*)$ and $\mu = E[s(Y|x) - \int_0^{Y} s(u|x)\psi(u|x)]du$,

we have

$$\text{var} \left( h(Y|x; \tau) - \int_0^{Y} h(u|x; \tau) \exp h(u|x; \phi^*) du \right)$$

$$= \text{var} \left( s(Y|x) - \int_0^{Y} s(u|x)\psi(u|x) du \right)$$

$$= E \left[ \left( s(Y|x) - \int_0^{Y} s(u|x)\psi(u|x) du - \mu \right)^2 ; Y < 1 \right]$$

$$+ E \left[ \left( - \int_0^{Y} s(u|x)\psi(u|x) du - \mu \right)^2 ; Y \geq 1 \right] \quad (6.53)$$

$$\geq \min_{0 \leq t \leq 1} f(t|x) \int_0^1 \left[ s(t) - \int_0^{t} s(u|x)\psi(u|x) du - \mu \right]^2 dt$$

$$+ \mathbb{1}(Y = 1) \int_0^1 \left[ - \int_0^{t} s(u|x)\psi(u|x du) - \mu \right]^2 dt.$$
By Condition 1 and 2, $P(Y = 1) > 0$. Set $c_1 = \min\{\min_{0 \leq t \leq 1} f(t|x), P(Y = 1)\}$. Then $c_1 > 0$. Hence by (6.53), Lemma 3 and Condition 1, there is a positive constant $c_2$ such that

$$\text{var} \left( h(Y|x; \tau) - \int_0^Y h(u|x; \tau) \exp h(u|x; \phi^*) \right)$$

$$\geq \frac{c_1}{2} \int_0^1 \left[ s(t|x) - \int_0^t s(u|x) \psi(u|x) du + \int_0^1 s(u|x) \psi(u|x) du \right]^2 dt$$

$$\geq c_2 \int_0^1 h^2(t|x; \tau) dt.$$

It follows from Condition 1 and (6.1) that there is a positive constant $c_3$ such that

$$\sum_i \text{var} \left( h(Y_i|x_i; \tau) - \int_0^{Y_i} h(u|x_i; \tau) \exp h(u|x_i; \phi^*) du \right) \geq c_3 n^{-1} |\tau|^2.$$

This completes the proof of Lemma 4.

\[\square\]

Set $I^* = I(\phi^*)$. Consider the approximation $\hat{\phi} = \hat{\phi}_n \in \Phi$ to $\hat{\phi} - \phi^*$ defined by $I^* \hat{\phi} = S^*$. Note that $\hat{\phi} = (I^*)^{-1} S^*$ and hence $E(\hat{\phi}) = 0$ and $[G^*(y,x)]' \hat{\phi} = [G^*(y,x)]'(I^*)^{-1} S^*$. It follows from (6.50), (6.4) and (6.44) that

$$|\tau'(I^*)^{-1} G^*(y,x)| = O(n^{-1}|\tau|) \text{ uniformly in } n, \tau, y, \text{ and } x. \quad (6.54)$$

**Lemma 5.** $\max_j |\hat{\phi}_j| = O_p(\sqrt{n^{-1} I \log I})$.

**Proof.** Since $(I^*)^{-1} VC(S^*)(I^*)^{-1}$ is the variance-covariance matrix of $\hat{\phi}$, it follows from (6.4), (6.44) and Lemma 4 that $\max_j \text{var}(\hat{\phi}_j) = O(I/n)$. Observe that

$$\hat{\phi}_j = \sum_i [(I^*)^{-1} G^*(Y_i,x_i)]_j.$$
By (6.54)
\[ \max_j \sup_{y,x} \left[ (\Gamma^*)^{-1} G^*(y,x) \right]_j = O(I/n). \]

The desired result follows from Condition 4 and Bernstein’s inequality. \(\square\)

(Proof of (6.35)) The proof is contained in the following lemma.

**Lemma 6.**
- \( \max_j |\hat{\phi}_j - \phi_j^*|^2 = O_p(n^{-1}I \log I). \)
- \( |\hat{\phi} - \phi^* - \hat{\varphi}| = O_p(n^{-2}I^3 \log I). \)

**Proof.** It follows from the maximum likelihood equation that

\[
\hat{\phi} - \phi^* - \hat{\varphi} = \hat{\phi} - \phi^* - (\Gamma^*)^{-1} S^*
\]
\[
= \hat{\phi} - \phi^* + (\Gamma^*)^{-1} D(\hat{\phi} - \phi^*)
\]
\[
= (\Gamma^*)^{-1}(\Gamma^* + D)(\hat{\phi} - \phi^*). \]

According to (6.4) and (6.44)
\[
|((\Gamma^*)^{-1}(\Gamma^* + D)(\hat{\phi} - \phi^*))|^2 = O(n^{-2}I^2|((\Gamma^* + D)(\hat{\phi} - \phi^*))|^2). \]

We claim that
\[
|((\Gamma^* + D)(\hat{\phi} - \phi^*))|^2 = O_p(n \max_j |\hat{\phi}_j - \phi_j^*|^2). \tag{6.55} \]

The proof of (6.55) will be given shortly. Therefore
\[
|\hat{\phi} - \phi^* - \hat{\varphi}|^2 = O_p(n^{-1}I^2 \max_j |\hat{\phi}_j - \phi_j^*|^2). \]

Consequently, by Lemma 5,
\[
\max_j (\hat{\phi}_j - \phi_j^*)^2 = O_p(n^{-1}I \log I + n^{-1}I^2 \max_j |\hat{\phi}_j - \phi_j^*|^2). \]
Thus by Condition 3,
\[ \max_j (\hat{\phi}_j - \phi_j^*)^2 = O_p(n^{-1} \log I). \]
which yields the desired results.

Proof of (6.55). Set \( N_i(t) = \text{ind}(Y_i \leq t) \) and \( Z_i(t) = \text{ind}(Y_i \geq t), i \leq i \leq n \). The log-likelihood function can be written as
\[
l(\phi) = \sum_i \left( \int g(u|x_i; \phi) dN_i(u) - \int Z_i(u) g(u|x_i; \phi) du \right).
\]
The \( j \)-th entry of the score function \( S(\phi) \) is given by
\[
\frac{\partial l(\phi)}{\partial \theta_j} = \sum_i \left( \int B_j(u|x_i) dN_i(u) - \int B_j(u|x_i) Z_i(u) \exp g(B_j(u|x_i)dN_i(u); \phi) du \right), \quad 1 \leq j \leq J,
\]
and the \( p \)-th entry is
\[
\frac{\partial l(\phi)}{\partial \beta_p} = \sum_i \left( \int Q_p(u|x_i) dN_i(u) - \int Q_p(u|x_i) Z_i(u) \exp g(B_j(u|x_i)dN_i(u); \phi) du \right), \quad 1 \leq p \leq P.
\]
The entry of the Hessian matrix of \( l(\phi) \) is given by
\[
l''_{jk}(\phi) = \frac{\partial^2 l(\phi)}{\partial \theta_j \partial \theta_k} = -\sum_i \int_0^1 B_j(u|x_i) B_k(u|x_i) Z_i(u) \exp g(u|x_i; \phi) du,
\]
\[
l''_{jp}(\phi) = \frac{\partial^2 l(\phi)}{\partial \theta_j \partial \beta_p} = -\sum_i \int_0^1 B_j(u|x_i) Q_p(u|x_i) Z_i(u) \exp g(u|x_i; \phi) du,
\]
\[
l''_{pq}(\phi) = \frac{\partial^2 l(\phi)}{\partial \beta_p \partial \beta_q} = -\sum_i \int_0^1 Q_p(u|x_i) Q_q(u|x_i) Z_i(u) \exp g(u|x_i; \phi) du.
\]
Set
\[
l'''_{vw}(\phi) = \frac{\partial^3 l(\phi)}{\partial \phi_v \partial \phi_w \partial \phi_m},
\]
where \( \phi_m \) could be \( \theta \) or \( \beta \), \( v, w, m \in [1, ..., J] \). Note that

\[
\int_0^1 [l''_{vw}(\phi^* + t(\hat{\phi} - \phi^*)) - l''_{vw}(\phi^*)] dt \\
= \int_0^1 \left( \int_0^1 \sum_m l'''_{vw}(\phi^* + u(\hat{\phi} - \phi^*))(\hat{\phi}_m - \phi^*_m) du \right) dt.
\]

The entry of \( \mathbf{I}' + \mathbf{D} \) can be written as

\[
\int_0^1 [l''_{vw}(\phi^* + t(\hat{\phi} - \phi^*)) - l''_{vw}(\phi^*)] + l''_{vw}(\phi^*) - E[l''_{vw}(\phi^*)] \\
= \sum_m A_{vwm}(\hat{\phi}_m - \phi^*_m) + l''_{vw}(\phi^*) - E[l''_{vw}(\phi^*)],
\]

where

\[
A_{vwm} = A_{nvwm} = \int_0^1 (1 - t)l'''_{vw}(\phi^* + t(\hat{\phi} - \phi^*)) dt.
\]

Thus the entry of \((\mathbf{I}' + \mathbf{D})(\hat{\phi} - \phi^*)\) is

\[
\sum_w \sum_m A_{vwm}(\hat{\phi}_w - \phi^*_w)(\hat{\phi}_m - \phi^*_m) + \sum_w \{l''_{vw}(\phi^*) - E[l''_{vw}(\phi^*)]\}(\hat{\phi}_w - \phi^*_w).
\]

We claim that

\[
\sum_v \left( \sum_w \sum_m A_{vwm}(\hat{\phi}_w - \phi^*_w)(\hat{\phi}_m - \phi^*_m) \right)^2 \\
= O_p \left( \max_v |\hat{\phi}_v - \phi^*_v|^2 n^2 I^{-2} |\hat{\phi} - \phi^*|^2 \right) \quad (6.56)
\]

\[
O_p \left( n \max_v |\hat{\phi}_v - \phi^*_v|^2 \right)
\]

and

\[
\sum_v \left( \sum_w \{l''_{vw}(\phi^*) - E[l''_{vw}(\phi^*)]\}(\hat{\phi}_w - \phi^*_w) \right)^2 \\
= O_p \left( n \max_v |\hat{\phi}_v - \phi^*_v|^2 |\hat{\phi} - \phi^*|^2 \right). \quad (6.57)
\]
The proofs of (6.56) and (6.57) will be given shortly. It follows from (6.56) and (6.57) that

$$\left| (I^* + D)(\hat{\varphi} - \varphi^*) \right|^2 = O_p \left( n \max_j (\hat{\varphi}_w - \varphi^*_w)^2 \right)$$

as desired.

Proof of (6.56). This follows from (6.32) and the following result.

**Lemma 7.** There is a positive constant $M_4$ such that

$$\sum_j \left( \sum_k \sum_{m \max_0 \leq t \leq 1} l''_{jkm}(\varphi^* + t(\hat{\varphi} - \varphi^*))||\tau_m|| \right)^2 \leq M_4 n^2 I^{-2}||\tau||^2, \quad \tau \in \Theta.$$  

**Proof.** Note that

$$\sum_j \left( \sum_k \sum_{m \max_0 \leq t \leq 1} l''_{jkm}(\varphi^* + t(\hat{\varphi} - \varphi^*))||\tau_m|| \right)^2 [\max_{0 \leq t \leq 1} ||\exp g(\cdot, \cdot, \varphi^* + t(\hat{\varphi} - \varphi^*))||_{\infty}]^2$$

$$\times \sum_j \left( \sum_k \sum_{m \max_0 \leq t \leq 1} \int U_j(u|x_i)U_m(u|x_i)du \right)^2 .$$

When $U_j U_m = B_j B_m$, there is a positive constant $J_0$ (not depending on $n$) such that

$$B_j(u) B_m(u) = 0 \quad \text{unless } |m - j| \leq J_0. \quad (6.58)$$

Thus, by (6.1) and the properties of B-splines,

$$\sum_i \sum_{m \max_0 \leq t \leq 1} \int B_j(u) B_m(u)du$$

$$\leq nI^{-1} \sum_{|m - j| \leq J_0} ||\tau_m||.$$
Hence, by the Schwarz inequality,

$$
\sum_j \left( \sum_i \sum_m \tau_m | \int B_j(u) B_m(u) du | \right)^2 \leq n^2 I^{-2} (2J_0 + 1) |\tau|^2.
$$

When $U_j U_m = Q_j B_m$,

$$
\sum_j \left( \sum_i \sum_m \tau_m | \int B_m(u) Q_j(u|x_i) du | \right)^2 \\
\leq P \max_{i,j} |Q_j(u|x_i)| \sum_i \sum_m \tau_m |B_m(u)Q_j(u|x_i) du \\
n^2 I^{-2} P |\tau|^2.
$$

The case when $U_j U_m = Q_j Q_m$ is trivial. The desired result follows from (6.4), (6.32) and Condition 4.

Proof of (6.57). Set

$$
V_{jk}(u) = \sum_i U_j(u|x_i) U_k(u|x_i) \{ Z_i(u) - E[Z_i(u)] \} \lambda^* u|x_i).
$$

Then $E[V_{jk}(u)] = 0$ for $0 \leq u \leq 1$ and

$$
l''_{jk}(\phi^*) - d[l''_{jk}(\phi^*]) = - \int V_{jk}(u) du.
$$

Thus (6.57) follows from the next lemma.

Lemma 8. Uniformly in $\tau \in \Phi$,

$$
\sum_j \left( \sum_k \tau_k \int V_{jk}(u) du \right)^2 = O_p \left( n \max_k \tau_k^2 \right).
$$
Proof. By (6.58) and the Schwarz inequality,

\[
\sum_j \left( \sum_k |\tau_k| \int V_{jk}(u) \, du \right)^2 \leq \max_k \tau_k^2 \sum_j \left( \sum_{|k-j| \leq J_0} \left| \int V_{jk}(u) \, du \right|^2 \right) \leq \max_k \tau_k^2 M_0 \times \sum_j \sum_{|k-j| \leq J_0} \int V_{jk}^2(u) \, du,
\]

where \( M_0 = \max((2J_0 + 1), P) \).

Since \( E[V_{jk}^2(u)] = \sum_i [U_j(u|x_i)U_k(u|x_i)\lambda^*(u|x_i)]^2 \text{var}(Z_i(u)) \) for \( 0 \leq u \leq 1 \), the desired result follows from (6.1) and (6.4).

This completes the proof of (6.52).

6.4 Asymptotic Distributions of the Estimates

Let \( I = I(\phi) \) denote the \( I \times I \) information matrix, which has entries

\[
-E\left( \frac{\partial^2 l(h)}{\partial \theta_j \partial \theta_k} \right) = \sum_i E \left( \int_0^{\tau_i} B_j(u)B_k(u) \exp h(u|x_i; \phi) \, du \right), \quad j = 1, \ldots, J; \quad k = 1, \ldots, J,
\]

\[
-E\left( \frac{\partial^2 l(h)}{\partial \theta_j \partial \beta_p} \right) = \sum_i E \left( \int_0^{\tau_i} B_j(u)Q_p(u|x_i) \exp h(u|x_i; \phi) \, du \right), \quad j = 1, \ldots, J; \quad p = 1, \ldots, P,
\]

and

\[
-E\left( \frac{\partial^2 l(h)}{\partial \beta_p \partial \beta_q} \right) = \sum_i E \left( \int_0^{\tau_i} Q_p(u|x_i)Q_q(u|x_i) \exp h(u|x_i; \phi) \, du \right), \quad p = 1, \ldots, P; \quad q = 1, \ldots, P.
\]

Let \( \varpi \) denote a real-valued parameter depending on \( \lambda^* \), so that \( \varpi = \Gamma(\phi^*) \) for some function \( \Gamma(\phi) \), \( \phi \in \Phi \). The maximum likelihood estimated of \( \varpi \) is given by
$\hat{\varpi} = \Gamma(\hat{\phi})$. Suppose $\Gamma$ is continuously differentiable on $\Phi$. Let $\nabla \Gamma(\phi)$ denote the gradient of $\Gamma$ at $\phi$. The asymptotic standard deviation (ASD) and standard error (SE) of $\hat{\varpi}$ are defined by

\[
ASD(\hat{\varpi}) = \sqrt{\nabla \Gamma(\phi^*)^T [I(\phi^*)]^{-1} \nabla \Gamma(\phi^*)}
\]

and

\[
SE(\hat{\varpi}) = \sqrt{\nabla \Gamma(\phi)^T [I(\phi)]^{-1} \nabla \Gamma(\phi)}.
\]

Recall that $\phi^* = [\theta^*, \beta]$ and $U = [B(t), Q(t|x)]$. Thus, for example,

\[
ASD(\hat{\alpha}(t|x)) = \sqrt{\nabla U(t|x)^T [I(\phi^*)]^{-1} \nabla U(t|x)}
\]

and

\[
SE(\hat{\varpi}) = \sqrt{\nabla U(t|x)^T [I(\phi)]^{-1} \nabla U(t|x)}.
\]

**Theorem 4.** Under Conditions 1-4, for $t \in T$ and $x \in \mathcal{X}$,

\[
\frac{\hat{\alpha}(t|x) - \alpha^*(t|x)}{ASD(\hat{\alpha}(t|x))} \xrightarrow{d} N(0, 1), \quad \frac{SE(\hat{\alpha}(t|x))}{ASD(\hat{\alpha}(t|x))} = 1 + o_p(1), \quad (6.59)
\]

\[
\frac{\hat{\lambda}(t|x) - \lambda^*(t|x)}{ASD(\hat{\lambda}(t|x))} \xrightarrow{d} N(0, 1), \quad \frac{SE(\hat{\lambda}(t|x))}{ASD(\hat{\lambda}(t|x))} = 1 + o_p(1), \quad (6.60)
\]

The proof of Theorem 3 is given below. Confidence intervals can be constructed using Theorem 3 in an obvious manner. Suppose $\varpi = \Gamma(\phi^*)$ is a parameter of interest. Then $\hat{\varpi} \pm z_{1-\alpha} SE(\hat{\varpi})$ is an asymptotic $100(1 - \alpha)$% confidence interval for $\varpi$.

**Proof of Theorem 4**

Throughout this section, we assume that Conditions 1-4 hold.

**Lemma 9.** $|\tau' V C(\mathbf{S}^*) \tau - \tau' \Gamma \tau| = O(nI^{-1}|\tau|^2 \rho), \quad \tau \in \Theta$
Recall that the log-likelihood function is given by

\[
l(\phi) = \sum_i \left( \int \nabla U_j(u|x_i) dN_i(u) - \int \nabla U_j(u|x_i) Z_i(u) \exp(h(u|x_i; \phi) du) \right)
\]

\[
= \sum_i \int \nabla U_j(u|x_i) dM_i(u|x_i), \quad 1 \leq j \leq J,
\]

where \(dM_i(u|x_i) = dN_i(u) - Z_i(u) \exp(h(u|x_i; \phi) du).\) Thus,

\[
\tau' VC(S^*) \tau = \text{var}(\tau' S^*) = \sum_i \text{var} \left( \int h(u|x_i; \tau) dM^*_i(u|x_i) \right)
\]

(6.61)

and

\[
\tau' I(\phi^*) \tau = \sum_i E \left( \int h^2(u|x_i; \tau) Z_i(u) \lambda^*(u|x_i) du \right)
\]

\[
= \sum_i \int h^2(u|x_i; \tau)(1 - F(u|x_i)) \lambda^*(u|x_i) du,
\]

(6.62)

where \(M^*_i(t) = N_i(t) - \int_0^t Z_i(u) \exp(h(u|x_i; \phi^*) du).\) Let \(E^*(\cdot)\) and \(\text{var}^*(\cdot)\) denote the expectation and variance functions taken with respect to \(f^*\). According to Theorem 2.5.4 of Fleming and Harrington (1991) (p.77), or Proposition II.4.1 of Andersen et al. (1993) (p.78), \(M^*_i(u|x_i)\) is a zero-mean martingale with \(\langle M^*_i, M^*_i \rangle^*(t) = \int_0^t Z_i(u) \exp(h(u|x_i; \phi^*) du).\) (Here it is necessary to use an alternative probability space with probability measure \(P^*\), under which the counting process \(N_i(t)\) has an intensity function \(\lambda^*; \langle \cdot, \cdot \rangle^*\) is the corresponding variation process.) Hence,

\[
\text{var}^*(\tau' S^*) = \sum_i E^* \left( \int h^2(u|x_i; \tau) Z_i(u) \lambda^*(u|x_i) du \right)
\]

\[
= \sum_i \int h^2(u|x_i; \tau)(1 - F(u|x_i)) \lambda^*(u|x_i) du.
\]

(6.63)
It follows from (6.32), (6.62), (6.63) and (6.1) that

\[ |\tau' \mathbf{I}(\phi^*) \tau - \text{var}(\tau' \mathbf{S}^*)| = O \left( n I^{-1} |\tau|^2 ||F - F^*||_\infty \right), \quad \tau \in \Theta. \tag{6.64} \]

Set

\[ U_i = \int h(u|x_i; \tau)dM^*_i(u|x_i) \]
\[ = h(Y_i|x_i; \tau) - \int h(u|x_i; \tau)\lambda^*(u|x_i)Z_i(u)du. \]

Then

\[ \text{var}(U_i) - \text{var}^*(U_i) = E(U_i^2) - [E(U_i)]^2 - E^*(U_i^2) + [E^*(U_i)]^2. \tag{6.65} \]

Write \( g = g(\cdot; \tau) \). Then

\[ E(U_i) = E(h(Y_i|x_i; \tau)) - E\left( \int h^* Z_i \right) \]
\[ = \int h(u|x_i)f(u|x_i) - \int h(u|x_i)\lambda^*(u|x_i)(1 - F(u|x_i)) \]

and

\[ E^*(U_i) = E^*(h(Y_i|x_i; \tau)) - E^*(\int h^* Z_i) \]
\[ = \int h(u|x_i)f^*(u|x_i) - \int h(u|x_i)\lambda^*(u|x_i)(1 - F^*(u|x_i)) \]
Thus,

\[
[E(U_i)]^2 - [E^*(U_i)]^2 = [E(U_i) - E^*(U_i)][E(U_i) + E^*(U_i)] \\
= \left( \int h(f - f^*) - \int h\lambda^*(F^* - F) \right) \\
\times \left( \int h(f + f^*) - \int h\lambda^*(F^* + F) \right)
\]

Hence, by (6.1), (6.4) and (6.5)

\[
\sum_i |[E(U_i)]^2 - [E^*(U_i)]^2| = O\left(nI^{-1}|\tau|^2||f - f^*||_\infty\right), \quad \tau \in \Theta. \tag{6.66}
\]

We claim that

\[
\sum_i |[E(U_i^2)] - [E^*(U_i^2)]| = O\left(nI^{-1}|\tau|^2||f - f^*||_\infty\right), \quad \tau \in \Theta. \tag{6.67}
\]

The proof of (6.67) will be given shortly. Hence, by (6.61) and (6.65) to (6.67),

\[
|\tau'VC(S^*)\tau - var^*(\tau'\tau)| = O\left(nI^{-1}|\tau|^2||f - f^*||_\infty\right), \quad \tau \in \Theta. \tag{6.68}
\]

The desired result follows from (6.5), (6.6), (6.64) and (6.68). To verify (6.67), we first note that,

\[
E(U_i^2) = E\left(h^2(Y_i|x_i; \tau)\right) + E\left[\left(\int_0^{Y_i} h(u|x_i; \tau)\lambda^*(u|x_i)du\right)^2\right] \\
- 2E\left(h(Y_i|x_i; \tau)\int_0^{Y_i} h(u|x_i; \tau)\lambda^*(u|x_i)du\right) \tag{6.69}
\]
\[ E^*(U_i^2) = E^* \left( h^2(Y_i|x_i; \tau) \right) + E^* \left( \left( \int_0^{Y_i} h(u|x_i; \tau) \lambda^*(u|x_i) du \right)^2 \right) \]

\[ - 2E^* \left( h(Y_i|x_i; \tau) \int_0^{Y_i} h(u|x_i; \tau) \lambda^*(u|x_i) du \right). \]

By (6.1)

\[ \sum_i \left| E \left( h^2(Y_i|x_i; \tau) \right) - E^* \left( h^2(Y_i|x_i; \tau) \right) \right| \]

\[ \leq \sum_i \int h^2(u|x_i) |f(u|x_i) - f^*(u|x_i)| du \]

\[ = O \left( nI^{-1} |\tau|^2 ||f - f^*||_\infty \right). \]

\[ E \left[ \left( \int_0^y h(u|x_i; \tau) \lambda^*(u|x_i) du \right)^2 \right] - E^* \left[ \left( \int_0^y h(u|x_i; \tau) \lambda^*(u|x_i) du \right)^2 \right] \]

\[ = \int \left( \int_0^y h(u|x_i; \tau) \lambda^*(u|x_i) du \right)^2 (f(y|x) - f^*(y|x)) dy. \]

Thus, by (6.4), (6.5) and (6.1),

\[ \sum_i \left| E \left[ \left( \int_0^y h(u|x_i; \tau) \lambda^*(u|x_i) du \right)^2 \right] \right| - E^* \left[ \left( \int_0^y h(u|x_i; \tau) \lambda^*(u|x_i) du \right)^2 \right] | = O \left( nI^{-1} |\tau|^2 ||f - f^*||_\infty \right). \]

Also

\[ E \left( h(Y|x; \tau) \int_0^Y h(u|x; \tau) \lambda^*(u|x) du \right) \]

\[ - E^* \left( h(Y|x; \tau) \int_0^Y h(u|x; \tau) \lambda^*(u|x) du \right) \]

\[ = \int h(t|x; \tau) [f(t|x) - f^* t|x] \times \left( \int_0^t h(u|x; \tau) \lambda^*(u|x) du \right) dt. \]
Thus, by the Schwarz inequality, (6.4) and (6.1),

\[
\sum_i |E \left( h(Y_i|x_i; \tau) \int_0^{Y_i} h(u|x_i; \tau) \lambda^*(u|x_i) du \right) - E^* \left( h(Y_i|x_i; \tau) \int_0^{Y_i} h(u|x_i; \tau) \lambda^*(u|x_i) du \right) | = O \left( nI^{-1}|\tau|^2 ||f - f^*||_\infty \right).
\]

(6.73)

It follows from (6.69) to (6.73) that (6.67) holds.

This completes the proof of Lemma 9. \(\square\)

(Proof of (6.59)) Set \(W_i = [\nabla K(t|x)]'(I^*)^{-1}G^*(Y_i|x_i), i = 1, 2, ... n\). Observe that

\[
[\nabla K(t|x)]'(\hat{\varphi}) = \sum_j \hat{\varphi}_j \nabla K_j(t|x) = \sum_i W_i.
\]

Also, \(E([\nabla K(t|x)]'(\hat{\varphi})) = 0\) and \(\text{var}([\nabla K(t|x)]'(\hat{\varphi})) = [\nabla K(t|x)]'(I^*)^{-1}VC(S^*)(I^*)^{-1}\nabla K(t|x)\).

**Lemma 10.** \(\text{var}([\nabla K(t|x)]'(\hat{\varphi})) \sim n^{-1}I\).

**Proof.** Since \(B_j \geq 0\) and \(\sum_j B_j = 1\), we have that

\[
|B(t|x)| \sim 1.
\]

(6.74)

Also, since there are fixed number of functions \(Q_p\) and it has boundaries for \(t \in [0, 1]\), by (6.74), (6.44) and (6.4),

\[
|(I^*)^{-1}\nabla K(t|x)| = O(n^{-1}I).
\]

(6.75)

By Lemma (9), (6.75) and Condition 4 \((\rho = o(1))\),

\[
|[\nabla K(t|x)]'(I^*)^{-1}VC(S^*)(I^*)^{-1}\nabla K(t|x) - [\nabla K(t|x)]'(I^*)^{-1}\nabla K(t|x)| = o(n^{-1}I).
\]

(6.76)
By (6.74) and (6.43),

\[ [\nabla \mathbf{K}(t|\mathbf{x})]'(\mathbf{I}^*)^{-1} \nabla \mathbf{K}(t|\mathbf{x}) \sim n^{-1} I. \tag{6.77} \]

It follows from (6.76) and (6.77) that

\[ \text{var}(\nabla \mathbf{K}(t|\mathbf{x})' \hat{\phi}) = (1 + o(1))[\nabla \mathbf{K}(t|\mathbf{x})]'(\mathbf{I}^*)^{-1} \nabla \mathbf{K}(t|\mathbf{x}) \sim n^{-1} I \]

as desired.

**Lemma 11.**

\[ \frac{\sum_j \hat{\phi}_j \nabla \mathbf{K}_j(t|\mathbf{x})}{SD([\nabla \mathbf{K}(t|\mathbf{x})]'\hat{\phi})} \xrightarrow{d} N(0, 1). \]

**Proof.** The random variables $W_1, ..., W_n$ are independent with mean zero. Moreover, by (6.75) and (6.50),

\[ |W_i|^2 = \|[[\nabla \mathbf{K}(t|\mathbf{x})]'(\mathbf{I}^*)^{-1} \mathbf{G}^*(Y_i|\mathbf{x}_i)]|^2 = O(I^2/n^2) \tag{6.78} \]

The desired result follows from Condition 4, Lemma 10 and the central limit theorem.

Now according to Lemma (1)(ii) and Condition 4,

\[ \|[[\nabla \mathbf{K}(t|\mathbf{x})]'(\hat{\phi} - \phi^* - \hat{\psi})\| = o_p(\sqrt{n^{-1}I}). \]

Since $\hat{\phi}(t|\mathbf{x}) - \phi^*(t|\mathbf{x}) = \sum_j(\hat{\theta}_j - \theta_j^*)B_j(t) + \sum_p(\hat{\beta}_p - \beta_p^*)Q_p(t|\mathbf{x})$, we now conclude from Lemma 10 and 11 that

\[ \frac{\hat{\phi}(t|\mathbf{x}) - \phi^*(t|\mathbf{x})}{SD([\nabla \mathbf{K}(t|\mathbf{x})]'\hat{\phi})} \xrightarrow{d} N(0, 1). \]

By (6.77),

\[ \text{AV}(\hat{\phi}(t|\mathbf{x})) = [\nabla \mathbf{K}(t|\mathbf{x})]'(\mathbf{I}^*)^{-1} \nabla \mathbf{K}(t|\mathbf{x}) \sim n^{-1} I. \]

114
Thus, by (6.76),
\[ \text{var}([\nabla K(t|x)]'(\hat{\phi})) \simeq AV(\hat{\phi}(t|x)). \]

Hence,
\[ \frac{\hat{\alpha}(t|x) - \alpha^*(t|x)}{\text{ASD}(\hat{\alpha}(t|x))} \overset{d}{\rightarrow} N(0, 1). \]

This completes the proof of the first part of Theorem 4. \( \square \)

Set \( \hat{I} = I(\hat{\phi}) \), The proof of the second part of Theorem 4 depends on the next two lemmas.

**Lemma 12.** Uniformly in \( \tau \in \Theta \),
\[ |(\hat{I} - I^*)\tau|^2 = O_p(n|\tau|^2I^{-1}\log I). \]

**Proof.** Observe that
\[
E[l''_{jk}(\phi)] - E[l''_{ji}(\phi^*)] = \max_m |\phi_m| \left[ \max_{0 \leq t \leq 1} ||\exp h(\cdot; \phi^* + t(\phi - \phi^*))||_{\infty} \right] \\
\times \sum_m \sum_i \int U_j(u|x_i)U_k(u|x_i)U_m(u|x_i)(1 - F(u|x_i))du,
\]

It follows the basic properties of B-splines and the boundness of functions \( Q \) as in the proof of Lemma (7) that uniformly in \( \phi, \tau \in \Theta \),
\[
\sum_j \left[ \sum_k (E[l''_{jk}(\phi)] - E[l''_{jk}(\phi^*)])\tau_k \right] ^2 \\
= O \left( n^2 \max_m (\phi_m - \phi_m^*)^2 \left[ \max_{0 \leq t \leq 1} ||\exp h(\cdot; \phi^* + t(\phi - \phi^*))||_{\infty} \right] ^2 I^{-2}|\tau|^2 \right).
\]

The desired conclusion follows from (6.4), (6.32), Condition 4 and (6.35). \( \square \)
Lemma 13. **Uniformly in** $\tau \in \Theta$,

$$\left| \left( \hat{I}^{-1} - (I)^{-1} \right) \tau \right|^2 = O_p \left( n^{-3} I^3 \log I |\tau|^2 \right).$$

*Proof.* Since $\hat{I}^{-1} - (I)^{-1} = (i^{*})^{-1}(I^{*} - \hat{I})\hat{I}^{-1}$, the desired result follows from (6.4), (6.44) with $\phi = \phi^{*}$ and $\phi = \hat{\phi}$ and Lemma 12.

The proof of the second part of (6.59) will now be given. Recall that

$$SE(\hat{\alpha}(t|x)) = \sqrt{[U(t|x)]'(\hat{I}^{-1}U(t|x))}$$

and

$$ASD(\hat{\alpha}(t|x)) = \sqrt{[U(t|x)]'(I^{*})^{-1}U(t|x))}.$$ 

By the Schwarz inequality, (6.74), Lemma 13 and Condition 4,

$$\left| [U(t|x)]'(\hat{I}^{-1}U(t|x)) - [U(t|x)]'(I^{*})^{-1}U(t|x)] \right|$$

$$= \left| [U(t|x)]'(\hat{I}^{-1} - (I)^{-1})U(t|x)) \right|$$

$$\leq \left| [U(t|x)]||\hat{I}^{-1} - (I)^{-1})U(t|x)| \right|$$

$$= O_p(\sqrt{n^{-3} I^3 \log I})$$

$$= o_p(n^{-1} I).$$

It now follows from (6.77) that the second part of (6.59) holds.

(Proof of the second part of Theorem 4, Equation (6.60))

It follows from (6.3) and (6.41) that

$$\hat{\lambda}(t|x) - \lambda^*(t|x) = [\hat{\alpha}(t|x) - \alpha^*(t|x)]\lambda^*(t|x) + o_p(\sqrt{n^{-1} I}).$$
By (6.4) and (6.77), \( ASD(\hat{\lambda}(t|x)) = \lambda^*(t|x)ASD(\hat{\alpha}(t|x)) \sim \sqrt{n^{-1}I} \). Thus the desired result follows from \( SE(\hat{\lambda}(t|x)) = \lambda^*(t|x)SE(\hat{\alpha}(t|x)) \), (6.55), (6.36) and (6.59).
APPENDIX

We provide two complete proofs in the appendix. The first one is the proof of discrete point process likelihood (see Section 2.2), and the second is the proof of time-rescaling theorem (see Section 2.3).

**Proposition 1.** Given $0 < \tau_1 < \tau_2 < \ldots < \tau_n < T$, a set of neural spike train points, the joint density function of these $n$ events in the time interval $(0, T]$ is

$$f(\tau_1, \tau_2 \ldots \tau_n \cap N(T) = n) = f(\tau_1, \tau_2 \ldots \tau_n \cap \tau_{n+1} > T) = f(\tau_1, \tau_2 \ldots \tau_n \cap N(\tau_n) = n) \cdot \Pr(\tau_{n+1} > T|\tau_1, \tau_2 \ldots \tau_n)$$

$$= \prod_{k=1}^{n} \lambda(\tau_k|H_{\tau_k}) \exp \left\{- \int_{\tau_{k-1}}^{\tau_k} \lambda(s|H_s)ds \right\} \exp \left\{ - \int_{\tau_n}^{T} \lambda(s|H_s)ds \right\}$$

$$= \prod_{k=1}^{n} \lambda(\tau_k|H_{\tau_k}) \exp \left\{ - \int_{0}^{T} \lambda(s|H_s)ds \right\}$$

$$= \exp \left\{ \int_{0}^{T} \log \lambda(s|H_s)dN(s) - \int_{0}^{T} \lambda(s|H_s)ds \right\} = \prod_{k=1}^{n} \lambda(\tau_k|H_{\tau_k}) \exp \left\{ - \int_{0}^{T} \lambda(s|H_s)ds \right\}$$

$$= \exp \left\{ \int_{0}^{T} \log \lambda(s|H_s)dN(s) - \int_{0}^{T} \lambda(s|H_s)ds \right\}$$

(6.79)

**Proof.** Let $\{t_k\}_{k=1}^{n}$ be a $K$ partition of the observation interval $(0, T]$. Take $\Delta_k = t_k - t_{k-1}$, where $t_0 = 0$. Assume that the partition is sufficiently fine so that there is at most one spike in any $(t_{k-1}, t_k]$. For a neural spike train choosing $\Delta_k \leq 1$ msec would suffice. We define $dN(k) = 1$ if there is a spike in $(t_{k-1}, t_k]$ and 0 otherwise, and the events

$$A_k = \text{one spike in } (t_{k-1}, t_k]|H_k), \quad E_k = \{A_k\}^{dN(k)}\{A_k^c\}^{1-dN(k)}, \quad H_k = \{ \bigcap_{j=1}^{k-1} E_j \},$$

(6.80)

for $k = 1, 2, \ldots, K$. 

118
In any interval \((t_{k-1}, t_k]\) we have

\[
\Pr(E_k) = \begin{cases} 
\lambda(t_k|H_k)\Delta_k + o(\Delta_k), & \text{if } dN(k) = 1 \\
1 - \lambda(t_k|H_k)\Delta_k + o(\Delta_k), & \text{if } dN(k) = 0 
\end{cases} \tag{6.81}
\]

By construction of the partition we must have \(\tau_j \in (t_{kj-1}, t_{kj}]\), \(j = 1, 2, ..., n\) for a subset of the intervals satisfying \(k_1 < k_2 < ... < k_n\). The remaining \((K - n)\) intervals have no spikes. The spike events form a sequence of correlated Bernoulli trials. Following Equation 6.81, the probability of exactly \(n\) events in \((0, T]\) may be computed as

\[
\Pr(\tau_j \in (t_{kj-1}, t_{kj}], j = 1, 2, ..., n \bigcap N(T) = n) \prod_{j=1}^{n} \Delta_{kj}
\]

\[
= \Pr(\bigcap_{k=1}^{K} E_k)
= \prod_{k=2}^{K} \Pr(E_k|\bigcap_{j=1}^{k-1} E_j) \Pr(E_1)
= \prod_{k=1}^{K} \left[\lambda(t_k|H_k)\Delta_k\right]^{dN(t_k)} \left[1 - \lambda(t_k|H_k)\Delta_k\right]^{1-dN(t_k)} + o(\Delta_*)
= \prod_{j=1}^{n} \left[\lambda(t_{kj}|H_{kj})\Delta_{kj}\right]^{dN(t_{kj})} \prod_{l \neq k_j} \left[1 - \lambda(t_l|H_l)\Delta_l\right]^{dN(t_l)} + o(\Delta_*)
= \prod_{j=1}^{n} \left[\lambda(t_{kj}|H_{kj})\Delta_{kj}\right]^{dN(t_{kj})} \prod_{l \neq k_j} \exp\left\{-\lambda(t_l|H_l)\Delta_l\right\} + o(\Delta_*)
= \exp\left\{\sum_{j=1}^{n} \log \lambda(t_{kj}|H_{kj})dN(t_{kj}) - \sum_{l \neq k_j} \lambda(t_l|H_l)\Delta_l\right\} \cdot \exp\left\{\sum_{j=1}^{n} \log \Delta_{kj}\right\} + o(\Delta(\ast)), \tag{6.82}
\]

where \(\Delta_* = \max_K \Delta_k\). Because \(\Delta_k\) is small for any \(k\), we have used the approximation \([1 - \lambda(k)\Delta_k] \approx \exp\left\{-\lambda(k)\Delta_k\right\}\). It follows that the probability density of exactly these
Proposition 2. Let $0 < u_1 < u_2 < \ldots < u_n < T$ be a realization from a point process with a conditional intensity function $\lambda(t|H_t)$ satisfying $0 < \lambda(t|H_t)$ for all $t \in (0,T]$. Define the transformation

$$\Lambda(u_k) = \int_0^{u_k} \lambda(u|H_u)du,$$

for $k = 1, 2, \ldots, n$, and assume $\Lambda(t) < \infty$ with probability one for all $t \in (0,T]$. Then the $\{\Lambda(u_k)\}$ are a Poisson process with unit rate.

Proof. Let $\eta_k = \Lambda(u_k) - \Lambda(u_{k-1})$ for $k = 1, 2, \ldots, n$ and set $\eta_T = \int_{u_n}^T \lambda(u|H_u)du$. To establish the result, it suffices to show that the $\tau_k$’s are independent and identically distributed exponential random variables with mean 1. Because the $\eta_k$ transformation is one-to-one and $\eta_{n+1} > \eta_T$ if and only if $u_{n+1} > T$, the joint probability density of the $\eta_k$’s is

$$f(\eta_1, \eta_2, \ldots, \eta_n \cap \eta_{n+1} > \eta_T) = f(\eta_1, \eta_2, \ldots, \eta_n) \cdot \Pr(\eta_{n+1} > \eta_T|\eta_1, \eta_2, \ldots, \eta_n)$$

(6.85)
We evaluate each of the two terms respectively,

\[
\Pr(\eta_{n+1} > \eta_T | \eta_1, \eta_2, ..., \eta_n) = \Pr(u_{n+1} > T | u_1, u_2, ..., u_n) \\
= \exp\left\{ - \int_{u_n}^{T} \lambda(u | H_{u_n}) \, du \right\} \\
= \exp\{-\tau_T\}, 
\]

(6.86)

and

\[
f(\eta_1, \eta_2, ..., \eta_n) = |J| f(u_1, u_2, ..., u_n \cap N(u_n) = n), 
\]

(6.87)

where \(|J|\) is the Jacobian of the transformation between \(u_j, j = 1, 2, ..., n\) and \(\eta_k, k = 1, 2, ..., n\). Because \(\eta_k\) is a function of \(u_1, u_2, ..., u_k\), \(J\) is a lower triangular matrix, and its determinant is the product of its diagonal elements defined as \(|J| = |\prod_{k=1}^{n} J_{kk}|\).

By assumption \(0 < \lambda(t | H_t)\) and the definition of \(\eta_k\), the mapping of \(u\) to \(\eta\) is one-to-one. Therefore, by the inverse differentiation theorem, the diagonal elements of \(J\) are

\[
J_{kk} = \frac{du_k}{d\eta_k} = \lambda(u_k | H_{u_k})^{-1}. 
\]

(6.88)

So, we have

\[
f(\eta_1, \eta_2, ..., \eta_n) = \prod_{k=1}^{n} \lambda(u_k | H_{u_k})^{-1} \cdot \prod_{k=1}^{n} \lambda(u_k | H_{u_k}) \exp\left\{ \int_{u_{k-1}}^{u_k} \lambda(u | H_u) \, du \right\} \\
= \prod_{k=1}^{n} \exp\{-[\Lambda(u_k) - \Lambda(u_{k-1})]\} \\
= \prod_{k=1}^{n} \exp\{-\eta_k\}. 
\]

(6.89)

Finally, substituting Equation (6.86) and Equation (6.89) yields

\[
f(\eta_1, \eta_2, ..., \eta_n, \cap \eta_{n+1} > \eta_T) = \prod_{k=1}^{n} \exp\{-\eta_k\} \exp\{-\tau_T\}, 
\]

(6.90)
which establishes the result.
Bibliography


