ABSTRACT

YU, LIN. On Estimation of Contagion-based Social Network Dependence with Event Time Data. (Under the direction of Dr. Wenbin Lu.)

As the advances of a wide variety of data collection methods and the emerging of a growing number of social networking services, the study of social networks has received much attention. In this dissertation, we analyze two different types of social network datasets and propose new methods to address them respectively. The first one refers to a social network with time-to-event data, and the second one considers a sequence of recurrent events data within a social network group. Our objective for both studies is to model the contagion-based social network dependence by analyzing how users behave through the influence of social events.

This dissertation contains three parts. Chapter 1 gives an overall introduction of social network analysis and overview of the underlying basis models for both studies. In Chapter 2, we extend generalized linear transformation model to study the social network influence with network-based time-to-event data. A time-varying covariate is proposed to incorporate network structure into the model and quantify the contagion-based social correlation. We further introduce a novel data generation procedure in simulations and establish the asymptotic properties of the proposed estimators. The numerical performances of the estimators are demonstrated via both simulation studies and a real-world application.

In Chapter 3, we propose a new way to model the contagion-based social network correlation with recurrent events data. It is well known that individuals are influenced through the network ties. In particular, the future actions of individuals depend not only on their own past behaviors, but also on their friends' past activities. Thus, we generalize the Hawkes self-exciting point process to model both self and mutually exciting influence in a social network dataset. A semi-parametric estimation method is considered for model flexibility. Consistency and asymptotic normality of the proposed estimators are further established. Both simulation studies and an analysis of an online social network dataset are provided to illustrate the empirical performance of the proposed method for the parameters estimation and the influential group detection.

On Estimation of Contagion-based Social Network Dependence with Event Time Data

by
Lin Yu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2018

APPROVED BY:

Dr. Soumendra Lahiri	Dr. Rui Song
Dr. Daowen Zhang	Dr. Wenbin Lu Chair of Advisory Committee

DEDICATION

To my parents.

BIOGRAPHY

Lin Yu was born in Shangrao, Jiangxi, China in November 1992. She graduated from Guangfeng High school in 2008 and entered Tianjin Polytechnic University afterwards. After obtaining her bachelor's degree in Information and Computation Science in 2012, she continued her study in the Department of Statistics at Columbia University in 2013, and got the Master's degree in June 2014 from Columbia University. In 2014, she was granted a full scholarship in Norch Carolina State University to pursue a Ph.D in Statistics. Under the direction of Dr. Wenbin Lu, her dissertation focuses on modeling and estimation of contagion-based social dependence in social network analysis. She will complete her Ph.D study in May, 2018.

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude and appreciation to my advisor Dr. Wenbin Lu, for his patience, encouragement and continuous support throughout my Ph.D study. His guidance, motivation, and immense knowledge help me in all the time of my research and writing of this thesis. I have been extremely lucky to have an advisor who cared so much of my research studies, and who responded my questions and queries so promptly. I learned a lot from Dr. Lu, not only on the statistical knowledge but also on how to conduct a research work. I could not have imagined having a better advisor and mentor for my graduate study.

Besides my advisor, I would like to thank the rest of my thesis committee members, Dr. Soumendra Lahiri, Dr. Rui Song and Dr. Daowen Zhang, for their insightful comments and suggestions, but also for the questions which incented me to widen my research from various perspective. A special thanks to Dr. Song for her valuable guidance and consistent support for one of my application projects in financial area. In addition, thanks to Dr. Lahiri and Dr. Zhang for their insight and useful advices for helping me to improve my interpersonal abilities and teaching skills when I was working as a graduate teaching assistant with them for different courses. I also want to thank Dr. Jeff Mielke, who agrees to serve as my graduate school representative in both my Ph.D oral preliminary exam and final defense out of his busy schedule.

I would like to take this opportunity to thank all the faculty members of the Department of Statistics at NCSU, who offered so many useful courses on various topics. I also want to thank all the staff members for their patience and being extremely helpful. Thanks to all my friends at NCSU, not only for their useful suggestions but also for being there to listen when I needed an ear.

My sincere thanks also goes to Dr. Ching-Ray Yu, who provided me an opportunity to join Pfizer as a statistician intern. The internship experience are extremely valuable to me, and I have gained a lot from my supervisor Dr. Ching-Ray Yu and all the colleagues, not only in choosing a career path in future, but also in learning strengths and weakness about myself.

Last but not the least, words cannot express the feelings I have for my parents for their constant unconditional love and support. A special thanks to my boyfriend, Shikai Luo, for all his love and always being there for me. Without their encouragement, I would not be able to achieve everthing I have today. They are and always will be my source of strength and inspiration.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Social Network Analysis

Over the past few decades, there are plenty of studies focusing on the network research in a wide range of areas, such as psychology, geography, economics, health care, online networking, treatment recommendation, etc. It is undoubtable that the study of social networks is a quickly widening multidisciplinary area and has caught a lot of attention. Social network data typically consists of a set of nodes representing people or other entities embedded in a social context, and a relational tie measured on each pair of nodes representing interaction, collaboration, or influence between nodes. Quantitative study on social networks was initially discussed by [Mor34], who developed a new technique called 'sociometry' to study the structure of groups and the positions of individuals within groups. Afterwards, the study of social networks has become quite popular in the sociological and behavioral sciences. Examples of the substantive concerns of recent social network studies include community detection, link prediction, dynamic social network analysis and many other important studies available for specific applications. Community detection aims at grouping individual nodes in accordance with the relationships among them to form strongly linked subgraph from the entire graph [Les08; KN11; Zha12; Xie13]; link prediction predicts missing links in current networks or new links in future networks, and is important for mining and analyzing the evolution of social networks [LNK07; MM07; Mil09]; dynamic social network analysis focuses on the statistical analysis of network data and the understanding of network dynamics via simulations [Cou03; Tan07; Xin10; Ngu11].

The field of social network analysis is fast growing, and the development of new approaches emerges constantly. [HL81] first develop a statistical tool named p_1 model to utilize the information about the attributes of individual nodes in social networks. Based on the framework of the p_1 model, [WW87] further propose a posteriori stochastic block model by incorporating the block information into the model when both the block structure and the individual nodes are of interest. The idea of stochastic block model is to divide the individual nodes into distinct blocks, where all nodes within the same block have the same pattern of connection to the nodes in other blocks, and such idea has been proved to be a useful statistical tool for social network analysis. [Hof02] establish a latent position method for studying the probability of a relational tie between nodes given the latent positions of individuals in an unobserved social space. This method has been further studied and generalized in a lot of literatures including [SM06; Han07; Air08; Roh11; Sus12] and etc.

Another important research question in social network analysis is to study the social correlation between individual nodes through the node covariates. In general, the social correlation might involve three factors: homophily, social contagion and external influence. The evidence of homophily can be referred as an increased rate of interactions among individuals sharing similar characteristics. For example, two students taking the photography class are more likely to become friends since they both like photography. The second one is social contagion, or social influence, which occurs when one's emotions, opinions or behaviors are triggered by his/her friends' recent actions. For instance, an individual goes to a new restaurant because of his/her friends' recommendation. The last one is external influence, where the external factors impact two individuals' behaviors or other measurable responses. Consider the following scenario. Two friends bought the same product recently simply because the item is on sale. However, sometimes these three sources might be confounded with each other. Researchers are interested in distinguishing and quantifying them respectively.

There has been some theoretical and empirical works on how a user's actions can be correlated to his/her friends' past behaviors via modeling the similarity between individuals' trait. [Li16] propose a regression model with network-based penalty on individual node effects to encourage similarity between predictions for linked nodes and to incorporate network cohesion in their model. Note that, they use the generic term "cohesion" to cover possibilities of both homophily and social contagion, however they can not distinguish the dependence resulting from homophily and from social contagion. Another popular approach for studying social network dependence is to consider a spatial autoregression model [Lee04; Lee10; Zho18], which characterizes the spatial correlation between different nodes through the processes known as assortative mixing on traits, or more simply as homophily. Once the spatial autocorrelation is estimated and the network structure is fixed, one can predict a node's behavior by inferring about his/her friends'.

Researchers are not only interested in modeling the social correlation resulting from the characteristic similarities, but also like to identify the situations where the social influence is the source of correlation, which is important in reality. For example, a marketing company can take advantage of this information to design viral marketing campaigns, or target a demographic group to send out coupons. Furthermore, the recent availability of massive network data sets generated by instant messages, emails, online posts, and smartphone communications enables novel investigations of the information diffusion and influence in networks. Related works can be found in, e.g. [Kem03; MR07; Che09a; Che09b; Cen10; Bak12; AW12; SDX13]. In applications like these works, the primary interests are to address the issue of identifying influential sets of individuals and modeling the information diffusion in networks.

1.2 Overview for Linear Transformation Model

The Cox proportional hazards model is one of the most important methods used in medical researches for investigating the association between the survival time of patients and one or more predictor variables. Let T be the 'failure time', and Z be a corresponding covariate vector. Denote $S_Z(\cdot)$ as the survival function of T given Z. The hazard function for the Cox proportional hazards model ([Cox72]) has the form

$$\lambda(t|Z) = \lambda_0(t)\exp(Z^T\beta), \tag{1.1}$$

where β is a $p \times 1$ vector of unknown regression parameters, and the cumulative hazard function is $\Lambda(t) = \int_0^t \lambda(s)ds$.

Based on the partial likelihood function, we can make inference about β in (1.1). The estimation of regression coefficients β can be obtained without a parametric assumption about the baseline hazard function $\lambda_0(t)$. However, in certain applications it may be violated. For example, sometimes it is more reasonable to assume that the baseline and subject-specific hazard functions become more similar with time. A useful alternative is the proportional odds model ([Ben83], [Pet84]), which assumes that the odds of survival is proportional and consequently the ratio of hazards approaches unity with time. There are certain restrictions of the proportional odds model, which may be violated when covariates have multiplicative effect on the odds of survival beyond time t. That is, the ratio of the hazards converges to unity as time t increases.

In this thesis, we consider a broader class of semiparametric linear transformation model ([DD88a]), which contains the proportional hazards and proportional odds models.

The cumulative hazard function is

$$\Lambda(t|Z) = G\{\Lambda_0(t)\exp(Z^T\beta)\} \tag{1.2}$$

where $G(\cdot)$ is a completely specified continuous increasing function. In [ZL07], the authors consider two functions for $G(\cdot)$ in (1.2). The first one is called BoxCox transformation, which has the form of

$$G(x) = \frac{1}{\rho}\{(1+x)^\rho - 1\},\ \rho \geq 0. \tag{1.3}$$

And the other one is called logarithmic transformation,

$$G(x) = \frac{1}{s}\log(1+sx),\ s \geq 0. \tag{1.4}$$

It is easy to see that $G(x) = x$ corresponds to the proportional hazards model with $\rho = 1$ in (1.3) or $s = 0$ in (1.4), and $G(x) = \log(1+x)$ corresponds to the proportional odds model with $\rho = 0$ in (1.3) or $s = 1$ in (1.4). Under mild conditions, the resulting estimators for β can be shown to be consistent, asymptotically normal, and asymptotically efficient based on the nonparametric maximum likelihood method ([ZL07]).

The linear transformation model in (1.2) can be extended to allow for time-varying covariates $Z(t)$ and to multiple and recurrent events ([ZL06]), which uses a counting process $N(t)$, recording the number of events occured by time t, to charaterize event history data. In chapter 2, we will discuss the extension in more details.

1.3 Overview for Self-exciting Point Process

A point process indexed by time is called a counting process when it counts the number of events occurring over time. Examples of events are waking up during night, having a child, resharing a post, and etc. While homogeneous Poisson processes assumes constant intensity over time, the assumptions of self-exciting point processes are that all the previous events influence the future evolution of the process. Self-exciting point processes are widely used to model "rich get richer" phenomena, such as financial transactions [Bow07; BH09; Emb11], video viewing activities [MC09; Mas13], and tweet popularities [She14; Zha15; KL16]. Furthermore, it is ideal for modeling information cascades in online social networks, since every new resharing not only increases its cumulative count of resharing by one, but also exposes new followers who may reshare the post in future.

In this section, we start by introducing the definition and basic form of the standard self exciting point process. Consider a point process $(N(t) : t \geq 0)$ with associated history $\mathscr{H}(t)$, the corresponding conditional intensity function is defined as

$$\begin{aligned} \lambda(t) = \lim_{h \to 0} \frac{E\{N(t+h) - N(t) | \mathscr{H}(t)\}}{h} &= \mu + \int_0^t g(t-s) dN(s) \\ &= \mu + \sum_{t_i < t} g(t - t_i), \end{aligned}$$

where $\mu > 0$ is the baseline intensity, and $g(t)$ is the excitation function. The term "self-exciting" indicates an increase of the conditional intensity because of an arrival, and since $g(t)$ is a monotone decreasing function, the impact dies off as time goes by. It was initially used to model the earthquake [Oga88]. As we know, the aftershock activities are more frequent after the occurrence of a major earthquake, and then decaying over time.

Depending on the form chosen for the excitation function g, the process may depend only on the recent history, if g decays rapidly, or may have longer term effects. Typically, due to the form of the hazard function $\lambda(t) \geq 0$, we require $g(t) \geq 0$ for $t \geq 0$ and $g(t) = 0$ for $t < 0$. An common choice of the excitation function is the exponential decay function

$$g(t) = ae^{-bt},\ t \geq 0,\ \text{with } a, b > 0.$$

which indicates that the self-exciting effect of an event decays in time t. However, in order to increase the model flexibility, the form of the excitation function is usually unspecified.

1.4 Summary of Our Works

In chapter 2, we model and estimate the contagion-based social network dependence based on time-to-event data. A generalized linear transformation model is proposed for the conditional survival probability at each observed event time, which uses a time-varying covariate to incorporate the network structure and quantify the contagion-based social correlation. We develop the nonparametric maximum likelihood estimation for the proposed model. The consistency and asymptotic normality of the resulting estimators for the regression parameters are established. Simulations are conducted to evaluate the empirical performance of the proposed estimators. We further apply the proposed method to analyze a time-to-event data about playing a popular mobile game from Tencent and find that there is a significant contagion-based social correlation in times to play the game.

Furthermore, we extend our method to address multiple events problems in social networks, e.g. network-based recurrent events data. In chapter 3, we develop an extended self-exciting point process to model the network of social influence with recurrent events data. A contagion-based social exciting term has been incorporated into the model to measure the influence of social actions between individuals. Based on the monotone B-splines, we establish an efficient estimation procedure for the model parameters. The resulting estimators are shown to be consistent, asymptotically normal and semiparametric efficient. To evaluate the performance of the proposed estimators, we experiment with both simulations and real-world datasets, and find that the proposed method can discover and calculate the social influence accurately.

CHAPTER

2

MODELING AND ESTIMATION OF CONTAGION-BASED SOCIAL NETWORK DEPENDENCE WITH TIME-TO-EVENT DATA

2.1 Introduction

Social network data consists of social ties, node characteristics and behaviors over time. It is known that people who are close to each other in a social network are more likely to behave in a similar way. One of the reasons they act similarly is due to the peer influence and social contagion that acts along the network ties. A primary interest of social network data analysis is to identify the contagion-based social correlation.

In the literature, [Ana08] proposed a way to model the contagion-based social correlation and developed statistical tests for the existence of such a correlation. Specifically, they modeled a specified action of users in a social network by a logistic regression with the number of âĂIJactiveâĂİ friends included as a covariate. Here, an active friend of a user is the one who took the action in the past. The regression coefficient associated with this covariate measures the magnitude of the social influence. In their method, time is discretized and a logistic regression is built on each discrete time point in an ad hoc fashion.

In this work, we are interested in modeling and estimation of the contagion-based social network dependence with time-to-event data. Our work is motivated by a study of the initial playing times of a popular mobile game from Tencent. Due to the confidentiality, we cannot disclose the name of the game here. This study involves 966 Tencent QQ users. Tencent QQ is a chatting application widely used in China. The players can send messages to their friends asking them to join the game. The endpoint of interest is the time at which a user begins to play the game since it was launched. In addition, some characteristics of these users, such as age, gender, location and QQ level, and their friend network are recorded. We would like to test whether an individual begins to play this game because his/her friends have started to played it, and estimate how much the influence will be. To do this, we utilize a generalized linear transformation model [DD88b; Che95] for the conditional survival probability at each observed event time and we also use a time-varying covariate for the number of active friends to model the contagion-based social network dependence. We develop an efficient estimation procedure for the model parameters based on the nonparametric maximum likelihood.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the proposed generalized linear transformation model for network-based time-to-event data and its associated data generation procedure and describe our methodology for parameters estimation. The asymptotic properties of the proposed estimators are studied in Section 2.3. In Section 2.4, simulation studies are conducted to evaluate the empirical performance of the proposed method. In Section 2.5, we further illustrate our method with an application to a data set for initial times of playing a mobile game from Tencent. Section 2.6 concludes with discussions. All proofs are contained in the Appendix.

2.2 Methodology

Consider a social network with n individuals and the adjacency matrix W, where $W_{i,j} = 1$ means individual i and j are friends, and $W_{i,j} = 0$ otherwise. By convention, all the diagonal entries of W are assumed to be zero. To model the contagion-based social network dependence for time-to-event data, we consider a new data generating mechanism. Specifically, let $T_{(k)}$ denote the time to the kth event in the network. In our motivating example, this represents the kth smallest time that a user started to play the mobile game since it was launched. During a fixed study period, we totally observe M_n event times, that is, $0 < T_{(1)} < \cdots < T_{(M_n)} \leq \tau$, where τ is the total study duration. Let i_k denote the index of the user who experienced the event at time $T_{(k)}$, $k = 1, \dots, M_n$. Here, the adjacency matrix W is assumed to be static over the study period. One of the possible future researches would be considering a dynamic social network, where the adjacency matrix is a time-dependent covariate.

In order to incorporate the network structure and quantify the contagion-based social correlation, we introduce a time-varying covariate $a_{j,k}$ for $j = 1, \ldots, n$ and $k = 1, \ldots, M_n$, which is defined as the number of active friends of individual j up to time $T_{(k)}$. Here, active friends of individual j up to time $T_{(k)}$ refer to those friends of individual j who had experienced the event before $T_{(k)}$. Let Z_j denote the p-dimensional baseline covariates of individual j. For simplicity, we define the covariates of individual j up to time $T_{(k)}$ as $X_{j,k} = (Z_j^T, g(a_{j,k}))^T$, where $g(\cdot)$ is a known non-decreasing function with $g(0) = 0$. For example, we can take $g(a) = \log(a+1)$. Let N_k denote the index set of individuals who are at risk up to time $T_{(k)}$. Note that $i_k \in N_k$. Then, the observed data can be summarized as

$$\{i_k, T_{(k)}, (X_{j,k}, j \in N_k);\ k = 1, \ldots, M_n\}.$$

Different from the classical survival data, the above data representation is not recorded based on individuals but according to sequential event times. Such a representation can facilitate the modeling of the contagion-based social correlation. In addition, it is assumed that the censoring can only occur at the end of the study, which is generally true for the social network study, for example, as in the considered mobile game application.

2.2.1 Proposed Model

In classical survival model, individual failure times can be generated independently. However, in social network study, the event time of individual i may depend on the status of his or her friends. Hence, we generate $T_{(1)}, \ldots, T_{(M_n)}$ sequentially based on the following conditional survival model. Specifically, suppose we have generated the first $(k-1)$ event times: $T_{(1)}, \ldots, T_{(M_n)}$, $k \geq 1$. Then, we know $(i_1, ..., i_{k-1})$ and the covariates $(X_{jk}, j \in N_k)$ on the interval $(T_{(k1)}, T_{(k)}]$ for those individuals who are at risk for the kth event. Note that $T_{(0)} = 0$ and $N_1 = (1, \ldots, n)$. At the baseline, there are no active nodes in the network, that is, $a_{j,1} \equiv 0$ for all j. Therefore, $X_{j,1} = (Z_j^T, 0)^T$. To generate $T_{(k)}$, we introduce a latent event time $T_{j,k}$ at which individual j first plays the game after $T_{(k-1)}$. To be specific, all the $T_{j,k}$ for $j \in N_k$ are not observed, they are only used as latent event time to characterize the kth observed event time $T_{(k)}$. Here, $T_{j,k}$ is generated from the following conditional survival model

$$
\begin{aligned}
&P(T_{j,k} > t | T_{j,k} > T_{(k-1)}, X_{j,k}) \\
&\qquad = \exp\Big(-\Big[G\left\{\Lambda(t)e^{\theta^T X_{j,k}}\right\} - G\left\{\Lambda(T_{(k-1)})e^{\theta^T X_{j,k}}\right\}\Big]\Big),
\end{aligned}
\tag{2.1}
$$

for $t > T_{k-1}$, where $\theta = (\beta^T, \beta_a)^T$ is the $(p+1)$-dimensional parameters of interest, $\Lambda(t)$ is an unspecified monotone increasing function with $\Lambda(0) = 0$ and $G(\cdot)$ is a specified monotone increasing transformation function, for example, a class of logarithmic transformation

$$
G(x) = \begin{cases} \frac{1}{s}\log(1 + sx), & s > 0 \\ x. & s = 0 \end{cases}
\tag{2.2}
$$

It can be seen that the above model is a generalization of the linear transformation model of [ZL06] for the conditional survival probability. Note that the parameter β_a measures the magnitude of the contagion-based social correlation.

Then, we define

$$
T_{(k)} = \min_{j \in N_k} T_{j,k}, \quad i_k = \arg\min_{j \in N_k} T_{j,k}.
$$

In addition, the numbers of active friends are updated by $a_{j,k+1} = a_{j,k} + W_{j,i_k}$ for $j \in N_{k+1} = N_k \backslash \{i_k\}$, which stay the same on the interval $(T_{(k)}, T_{(k+1)}]$. Repeat the above step until all the event times are generated. Based on the proposed data generating mechanism, the observed log-likelihood is given by

$$\ell_n(\theta,\Lambda) = \sum_{k=1}^{M_n} \Big(\log\lambda(T_{(k)}) + \theta^T X_{i_k,k} + \log \dot{G}\left\{\Lambda(T_{(k)}-)e^{\theta^T X_{i_k,k}}\right\} \\ - \sum_{j\in N_k}\Big[G\left\{\Lambda(T_{(k)})e^{\theta^T X_{j,k}}\right\} - G\left\{\Lambda(T_{(k-1)})e^{\theta^T X_{j,k}}\right\}\Big]\Big), \tag{2.3}$$

where $\lambda(t) = d\Lambda(t)/dt$ and $\dot{G}(u) = dG(u)/du$.

2.2.2 Nonparametric Maximum Likelihood Estimation

Here, we derive the nonparametric maximum likelihood estimation based on the likelihood function (2.3). The maximum of (2.3) does not exist if $\Lambda(\cdot)$ is restricted to be absolutely continuous. As widely studied in the literature for the nonparametric maximum likelihood estimation, we assume that $\Lambda(\cdot)$ is a non-decreasing step function with jumps only at observed event times $T_{(1)},\ldots,T_{(M_n)}$. Let $\Lambda\{T_{(k)}\}$ be the jump size at time $T_{(k)}$. Then, we have $\lambda(T_{(k)}) = \Lambda\{T_{(k)}\}$ for $k = 1,\ldots,M_n$.

To simplify the estimation, we consider a reparameterization. Define $\gamma_k = \log\Lambda\{T_{(k)}\}$. We have $\Lambda(T_{(k)}) = \sum_{\ell=1}^{k} e^{\gamma_\ell} = e^{\gamma_k} + \Lambda(T_{(k-1)})$, $k = 1,\ldots,M_n$. Thus, the log-likelihood function can be rewritten as

$$\ell_n(\theta,\gamma) = \sum_{k=1}^{M_n} \left(\gamma_k + \theta^T X_{i_k,k} + \log \dot{G}\left\{ (\sum_{\ell=1}^{k-1} e^{\gamma_\ell}) e^{\theta^T X_{i_k,k}} \right\} \right. \\ \left. - \sum_{j\in N_k}\left[G\left\{ (\sum_{\ell=1}^{k} e^{\gamma_\ell}) e^{\theta^T X_{j,k}} \right\} - G\left\{ (\sum_{\ell=1}^{k-1} e^{\gamma_\ell}) e^{\theta^T X_{j,k}} \right\} \right] \right). \tag{2.4}$$

In the following, we consider the logarithmic transformation function (2.2) and present the estimation of the parameters (θ,γ) in two cases: $s = 0$ and $s > 0$. However, the proposed estimation method can be easily extended to other specified transformation functions.

First, consider the case $s = 0$ with $G(x) = x$ and $\dot{G}(x) \equiv 1$. Then, the observed log-likelihood is reduced to

$$\ell_n(\theta,\gamma) = \sum_{k=1}^{M_n} \left(\gamma_k + \theta^T X_{i_k,k} - e^{\gamma_k} \sum_{j \in N_k} e^{\theta^T X_{j,k}} \right). \tag{2.5}$$

Taking the derivative of (2.5) with respect to γ_k and setting them equal to 0, we can obtain an explicit solution for γ_k as $\hat{\gamma}_k(\theta) = -\log\left(\sum_{j \in N_k} e^{\theta^T X_{j,k}}\right)$, for $k = 1, \ldots, M_n$. Then, plugging $\hat{\gamma}_k(\theta)$ back into model (2.5), we obtain the profile log-likelihood for θ

$$p\ell_n(\theta) = \sum_{k=1}^{M_n} \left\{ \theta^T X_{i_k,k} - \log\left(\sum_{j \in N_k} e^{\theta^T X_{j,k}} \right) \right\}, \tag{2.6}$$

which is similar to the log partial likelihood function for the proportional hazards model. Let $\hat{\theta}_n$ denote the resulting maximizer of θ. The asymptotic variance-covariance matrix of $\hat{\theta}_n$ can be estimated by $\mathrm{I}^{-1}(\hat{\theta}_n)$, where $\mathrm{I}(\hat{\theta}_n)$ is the negative of the second derivative of $p\ell_n(\theta)$ with respect to θ.

Next, we consider the case $s > 0$. The log-likelihood function (2.4) reduces to

$$\begin{aligned} \ell_n(\theta,\gamma) = \sum_{k=1}^{M_n} \Bigg(& \gamma_k + \theta^T X_{i_k,k} - \log\left\{ 1 + s \left(\sum_{\ell=1}^{k-1} e^{\gamma_\ell} \right) e^{\theta^T X_{i_k,k}} \right\} \\ & - \frac{1}{s} \sum_{j \in N_k} \left[\log\left\{ 1 + s \left(\sum_{\ell=1}^{k} e^{\gamma_\ell} \right) e^{\theta^T X_{j,k}} \right\} - \log\left\{ 1 + s \left(\sum_{\ell=1}^{k-1} e^{\gamma_\ell} \right) e^{\theta^T X_{j,k}} \right\} \right] \Bigg). \end{aligned} \tag{2.7}$$

Then, the estimates of θ and γ_k's can be obtained by the following procedure.

Step 1. Choose the initial estimator $\theta^{(0)}$, for example, $\theta^{(0)} = 0$;

Step 2. Given $\theta^{(0)}$, we solve $\gamma_k^{(1)}$ sequentially by maximizing $\ell_n(\theta^{(0)}, \gamma_1, \ldots, \gamma_{M_n})$ using a coordinate descent algorithm;

Step 3. Given $\{\gamma_k^{(1)}, k = 1, \ldots, M_n\}$, we update θ by

$$\theta^{(1)} = \arg\max_{\theta} \ell_n(\theta, \gamma^{(1)}, \ldots, \gamma_{M_n}^{(1)});$$

Step 4. Iterate *Step* 2 and *Step* 3 until a convergence criterion is met.

Let $\hat{\theta}_n$ and $\hat{\Lambda}_n$ denote the resulting estimators of θ and Λ, respectively, at convergence. The asymptotic variance-covariance matrix of $\hat{\theta}_n$ can be obtained by the following numerical differentiation method. For a small value $\delta > 0$, let $\hat{\gamma}_{n,j}^{+}$ and $\hat{\gamma}_{n,j}^{-}$ denote the solutions for γ obtained by maximizing $\ell_n(\theta, \gamma)$ with θ fixed at $\hat{\theta}_n + \delta e_j$ and $\hat{\theta}_n - \delta e_j$, respectively, where e_j is a $(p+1)$-vector with the jth component as 1 and others as 0, $j = 1, \ldots, p+1$. Let $\ell_{n,k}(\theta, \gamma)$ denote the kth summand in $\ell_n(\theta, \gamma)$. Define $S_{k,j}(\hat{\theta}_n) = \{\ell_{n,k}(\hat{\theta}_n + \delta e_j, \hat{\gamma}_{n,j}^{+}) - \ell_{n,k}(\hat{\theta}_n - \delta e_j, \hat{\gamma}_{n,j}^{-})\}/(2\delta)$ and $S_k(\hat{\theta}_n) = \{S_{k,1}(\hat{\theta}_n), \ldots, S_{k,p+1}(\hat{\theta}_n)\}^T$. Then, the observed information matrix is $I(\hat{\theta}_n) = \sum_{k=1}^{M_n} S_k(\hat{\theta}_n) S_k(\hat{\theta}_n)^T$ and the asymptotic variance-covariance matrix of $\hat{\theta}_n$ can be estimated by $\{I(\hat{\theta}_n)\}^{-1}$.

2.3 Asymptotic Properties

Denote the true values of θ and Λ by θ_0 and Λ_0. To establish the asymptotic properties of the proposed estimators, we assume the following conditions:

Condition 1. The function $\Lambda_0(t)$ is strictly increasing and continuously differentiable with $\Lambda_0(\tau) < \infty$, and the parameters θ_0 lie in the interior of a compact set $\mathscr{C}$.

Condition 2. The covariates vectors $X_{j,k}$ are bounded in the sense that $P(|X_{j,k}| < m) = 1$

for some positive constant m, for any j, k. In addition, if there exists a vector γ and a deterministic function $A(t)$ such that $A(t) + \gamma^T X_{j,k} = 0$ with probability one, then $\gamma = 0$ and $A(t) = 0$.

Condition 3. The information matrix $I(\theta_0)$ defined in the Appendix is finite and positive definite.

Note that Conditions 1-3 are commonly assumed in the literature for establishing the asymptotic properties of nonparametric maximum likelihood estimators in survival models (e.g. [ZL06]). In particular, the boundedness of covariates assumed in Condition 2 is satisfied when the following two conditions hold: (i) the baseline covariates Z are bounded; (ii) the number of friends of each node is bounded by a constant as the number of nodes goes to infinity, i.e. the social network is sparse in some sense.

Theorem 2.3.1 (Consistency) *Assumptions conditions 1-2 hold. We have*

$$\sup_{t\in[0,\tau]} |\hat{\Lambda}_n(t) - \Lambda_0(t)| \to 0 \textit{ a.s.} \quad \textit{and } ||\hat{\theta}_n - \theta_0||_2 \to 0 \textit{ a.s.}$$

Theorem 2.3.2 (Asymptotic Normality) *Assume conditions 1-3 hold. We have* $n^{1/2}(\hat{\theta}_n - \theta_0)$ *converges in distribution to a multivariate normal with mean* 0 *and variance* $\{I(\theta_0)\}^{-1}$.

2.4 Simulation Studies

In simulations, we consider a social network with $n = 1000$ subjects. The adjacency matrix W is generated by $P(W_{i,j} = 1) = 0.1$ for $i \neq j$. Event times $T_{(k)}$'s are generated sequentially following the descriptions given in Section 2.1. Here, we consider a single baseline covariate Z generated from a standard normal distribution and a logarithm transformation of the time-varying covariate, $g(a) = \log(a+1)$. We choose the regression parameters as $\beta = 0.5$

and $\beta_a = 0, 0.01, 0.05$ or 0.1. Here, β_a measures the magnitude of the social influence. In addition, we set $\Lambda(t) = \lambda t$ with $\lambda = 0.01$. We consider the link function $G(x) = \frac{1}{s}\log(1 + sx)$ with $s = (0, 0.5, 1)$. The study duration τ is chosen to yield the total number of events $M_n = 600$ or 800.

Table 2.1 summaries the results based on 1000 Monte Carlo replicates for each setting. We observe that in all settings the proposed estimators are nearly unbiased, the standard error estimators are close to the standard deviations of the estimators, and the empirical coverage probabilities of the 95% Wald-type confidence intervals are close to the nominal level.

2.5 Analysis of Mobile Game Data

We apply our method to analyze a time-to-event data about playing a popular mobile game from Tencent. The study involves 966 individuals over the 77 days duration. The friendship connections between individuals are known, which can be represented as the adjacency matrix W. The time at which each individual began to play the mobile game since it was launched are recorded. In addition, the baseline information, such as age, gender, location, and QQ level, are recorded. The friends network of these individuals is shown in Figure 2.1. We notice that there are some isolated nodes in the network. In addition, we divide individuals into five groups based on the number of active friends at the end of study and show them in different color in Figure 2.2. There are 241 individuals who have zero active friends over the study duration. The majority of individuals belong to the second group with the number of active friends greater than 0 and less than or equal to 10. Note that there is one individual with the number of active friends greater than 100, denoted by the yellow dot in Figure 2.2.

Table 2.1 Simulation results. *SE, mean of estimated standard errors; SD, standard deviations of the estimates; CP, empirical coverage probability of 95% Wald-type confidence intervals.*

		$M_n = 600$				$M_n = 800$			
$s = 0$	Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
β	0.5	0.500	0.043	0.044	0.950	0.500	0.038	0.039	0.943
β_a	0	0.002	0.179	0.185	0.946	0.002	0.172	0.178	0.942
β	0.5	0.500	0.043	0.044	0.947	0.500	0.038	0.039	0.946
β_a	0.01	0.012	0.180	0.185	0.945	0.012	0.173	0.178	0.943
β	0.5	0.500	0.043	0.044	0.949	0.500	0.038	0.039	0.941
β_a	0.05	0.051	0.180	0.186	0.945	0.050	0.173	0.178	0.946
β	0.5	0.500	0.043	0.044	0.946	0.500	0.038	0.039	0.946
β_a	0.1	0.101	0.181	0.187	0.944	0.100	0.174	0.179	0.945
$s = 0.5$	Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
β	0.5	0.502	0.053	0.053	0.954	0.502	0.050	0.050	0.958
β_a	0	0.005	0.196	0.151	0.970	0.008	0.191	0.148	0.965
β	0.5	0.502	0.054	0.053	0.957	0.502	0.051	0.050	0.958
β_a	0.01	0.016	0.197	0.153	0.970	0.019	0.191	0.149	0.962
β	0.5	0.502	0.054	0.054	0.956	0.502	0.051	0.051	0.959
β_a	0.05	0.058	0.198	0.154	0.968	0.060	0.193	0.151	0.965
β	0.5	0.503	0.055	0.056	0.954	0.503	0.052	0.052	0.956
β_a	0.1	0.110	0.200	0.160	0.964	0.111	0.194	0.156	0.961
$s = 1$	Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
β	0.5	0.506	0.063	0.067	0.948	0.505	0.062	0.064	0.954
β_a	0	0.017	0.207	0.207	0.935	0.015	0.203	0.197	0.942
β	0.5	0.506	0.064	0.067	0.949	0.505	0.062	0.064	0.955
β_a	0.01	0.027	0.208	0.208	0.936	0.026	0.204	0.199	0.939
β	0.5	0.506	0.064	0.068	0.948	0.506	0.063	0.065	0.955
β_a	0.05	0.068	0.209	0.211	0.936	0.068	0.205	0.201	0.936
β	0.5	0.507	0.066	0.070	0.951	0.506	0.064	0.067	0.955
β_a	0.1	0.120	0.212	0.214	0.927	0.119	0.207	0.204	0.936

We fit the proposed models with age and gender included as baseline covariates. As in simulations, we consider the logarithm transformation of the time-varying covariate, $g(a) = \log(a+1)$, and the link function $G(x) = \frac{1}{s}\log(1+sx)$ with $s = (0, 0.5, 1)$. The estimation results of the fitted models are given in Table 2.2. We also report the log likelihood value of the fitted models. The results indicate that the contagion-based social correlation is positive and significant in all models. This implies that individuals are influenced by their friends' behavior. As the number of active friends increases, an individual is more likely to start playing the game soon. In addition, based on the fitted log likelihood values, the models with $s = 0.5$ and 1 have comparable fit and are better than that with $s = 0$.

Table 2.2 Analysis of mobile game data. *log-LH, log likelihood value of the fitted model.*

		age	gender	$g(a)$	log-LH
	Estimation	3.5×10^{-3}	5.6×10^{-2}	1.1×10^{-1}	
$s = 0$	Stad. Err.	6.9×10^{-3}	7.7×10^{-2}	3.9×10^{-2}	-6639.3
	Z statistics	0.506	0.727	2.796	
	Estimation	-1.2×10^{-2}	7.7×10^{-2}	3.4×10^{-1}	
$s = 0.5$	Stad. Err.	9.7×10^{-3}	1.3×10^{-1}	8.5×10^{-2}	-6632.3
	Z statistics	-1.245	0.574	3.941	
	Estimation	-6.1×10^{-3}	1.2×10^{-1}	3.3×10^{-1}	
$s = 1$	Stad. Err.	9.8×10^{-3}	1.3×10^{-1}	8.6×10^{-2}	**-6631.9**
	Z statistics	-0.622	0.878	3.910	

2.6 Discussion

In this chapter, we propose a new way of modeling and estimation for contagion-based social dependence with time-to-event data. It can be extended to accommodate multiple events, such as network-based recurrent events data, by incorporating both self-exciting

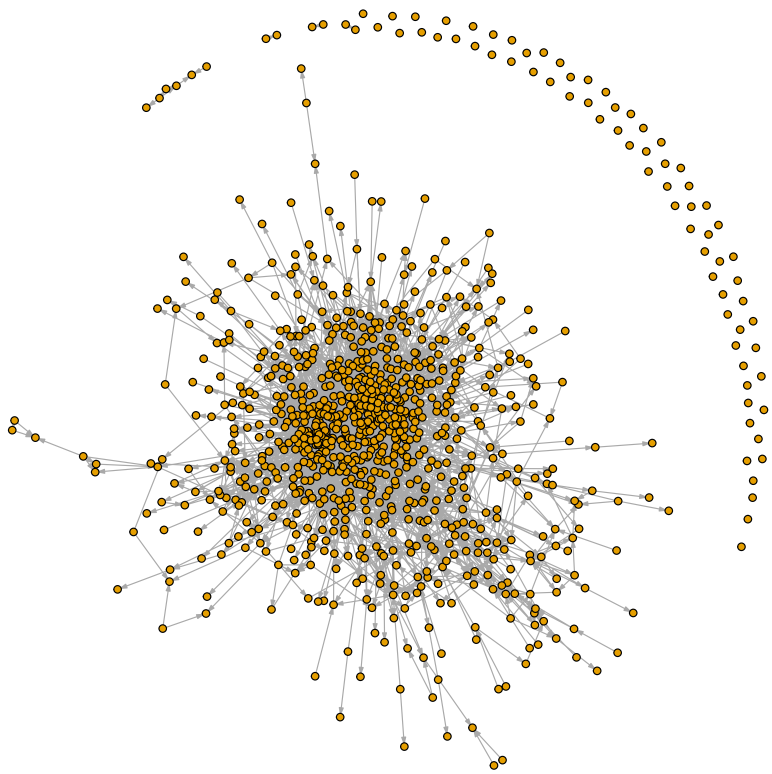

Figure 2.1 Network visualization

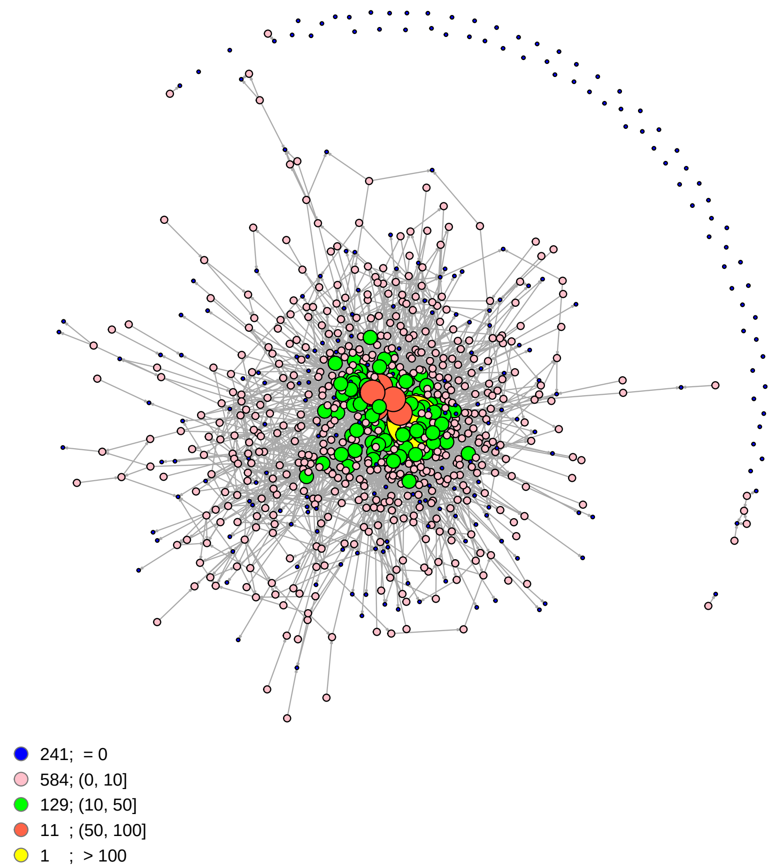

Figure 2.2 Number of active-friends at the end of study

and contagion-based social exciting processes, which will be discussed in chapter 3.

Another interesting direction for future research would be extending our method to dynamic social networks with evolving friendship connections. For example, individuals could become friends with others during the study period, which will change the value of the adjacency matrix. Furthermore, our technique only considers the case when the transformation parameter s is known. It would be very interesting to treat s as unknown and try to estimate s along with other parameters in our model.

CHAPTER

3

SELF AND MUTUALLY EXCITING POINT PROCESS FOR RECURRENT EVENTS DATA IN NETWORKS

3.1 Introduction

Social network modeling has long been of interest to researchers. Recently it has begun to evolve rapidly and attracted growing interest both within statistics and more widely.

Social network data typically consists of a set of nodes with node characteristics, social ties measured on each pair of nodes and behaviors over time. One of the primary interests in this area is to understand the driving forces behind the evolution of social networks, such as, epidemic diffusion [Mor93; MN00; Lei06], peer influence [Bur07; Bra09; AW12; Ye12], the teams formation [Gui05; CC06; Ana12], and a wide range of other interesting studies. In other words, it is known that individuals are influenced through the network ties, and it is also known that individuals take into account others' attributes when deciding to take certain actions.

Recent explosion of massive online social networking websites (Facebook, Twitter, LinkedIn, etc.) has enabled individuals to connect with their friends more conveniently. For example, users can update their status, share interesting posts, and initiate a conversation with their friends via Facebook. People are also more likely to be influenced by their friends on social networks. If many of one's friends are active users on Facebook, he will tend to be more active as well, by resharing the same post or updating his own status, etc. Furthermore, such influences through social networks are usually recurrent. For instance, if one's post is liked by a lot of his friends, he will be more willing to update his status and share posts more frequently on Facebook. The stimulus of such social actions comes from two sources; one is self, and the other is friends' influence. One interesting research question that has caught a lot of attention is to study how users behave through the influence of social actions. It can help evaluate the performance of existing systems, and lead to better site design and advertisement placement policies. For example, Facebook can take advantage of such information to target the influential group with specific ads. It is also helpful in viral marketing that the viral marketers could exploit the models of user interaction to quickly and widely spread their promotions. However, due to the large size of social network datasets and complex interaction between individuals, it is still a challenging problem, and the potential benefits have attracted a lot of researchers.

Previous studies on examining the influence of social actions using online network datasets include sending online messages to other users [Gol07; Chu08; Hub08; Wil09], making the decision to use an online social technologies [LH07; CL10], and assessing the third party applications [Naz08; Gjo08; Sch09]. All of these studies analyze the social influence of user actions through certain "visible" artifacts like messages, posts and comments. In this paper, our method is also based on the analysis of such kind of "visible" artifacts to estimate the social influence of online network datasets. But our method can also be applied to other social network datasets, provided that there are certain social actions observed within networks over time. To do this, we generalize the standard self-exciting point process to characterize the stimulus of social actions, resulting from both self and friends' influence.

It is widely accepted that the self-exciting point process [Haw71] is a useful technique to model recurrent events data with the occurrence of one event increasing the possibility of future events. Over the past decades, it has been applied in a wide range of areas, such as seismology, criminology, finance, social science, etc. [Bai15] propose a semi-parametric Hawkes self-exciting process regression model for modeling recurrent events data with temporal clustering feature. Based on the monotone B-spline approximation of the excitation function, they develop estimators for both parametric and nonparametric components of the model and further establish their asymptotic results. There have also been some works devoted to study social network analysis. By taking account of the sparsity and low-rank structure of the social network, [Zho13] establish a regularized convex optimization approach to infer the social network influence from the observed recurrent events based on the mutually-exciting multi-dimensional Hawkes model. The resulting optimization problem is shown to be efficient and accurate both on simulation studies and real world datasets. [Zha15] develop a flexible framework for modeling information cascades and predicting the final size of an information cascade spreading through a network. Based on self-exciting point process, their approach provides a theoretical framework to explain temporal patterns of information cascades and accurately predict its final size.

In this study, we focus on modeling the social network influence with recurrent events data. An extension of the standard self-exciting point model has been used to incorporate both self and mutually exciting parts into the model. Moreover, we develop an efficient estimation procedure for the model parameters based on the monotone B-splines. The estimated parameters are shown to be consistent, asymptotically normal, and semi-parametric efficient. Furthermore, we evaluate the performance of the proposed method with both simulation and an online social network dataset, consisted of private messages sent on an online social networking website at the University of California, Irvine [1].

The remainder of this chapter is organized as follows. In Section 3.2, we review the Hawkes self-exciting process, and extend it to model network-based recurrent time-to-event data by incorporating a mutually exciting influence into the model. A semi-parametric estimation method is proposed in Section 3.3. Then, we establish the asymptotic properties of the proposed estimators in Section 3.4. In Section 3.5, the empirical performance of the proposed estimators are evaluated using simulations, and the proposed method is further illustrated with an application to a network data set with recurrent time-to-event data in Section 3.6. Section 3.7, concludes with discussions. All proofs are given in the supplementary material.

---

[1] https://snap.stanford.edu

3.2 Proposed Model

Consider a social network with n individuals and the adjacency matrix W, where $W_{i,j} = 1$ means individuals i and j are friends, and $W_{i,j} = 0$ otherwise. By convention, we assume all the diagonal entries of W to be zero. Let $T_{(k)}$ denote the time to the kth event in the network, and the event times are generated sequentially. During a fixed study period τ, the observed event times are $0 < T_{(1)} < T_{(2)} < \cdots < T_{(M)} \leq \tau$ with the corresponding subject indexes $\{i_{(1)}, i_{(2)}, \ldots, i_{(M)}\}$, where M is the total number of observed events. In addition, let $N_i(t)$, $i = 1, \cdots, n$, denote the counting process for the events taken on individual i. Specifically, the event times of individual i are denoted by $T_{(i_1)} < T_{(i_2)} < \cdots < T_{(i_{n_i})}$, where $1 \leq i_1 < \cdots < i_{n_i} \leq M$ and $n_i = N_i(\tau)$. Let X_i denote the baseline covariates for individual i. Then, the observed data are

$$W \text{ and } \{(X_i, n_i, t_{i1}, \ldots, t_{in_i}); i = 1, \ldots, n\}.$$

As generally assumed in social network analysis, we only allow the censoring at the end of the study in this work. An example under this scenario is given in the real-world dataset. The time of all the private messages has been recorded up to a given time span. Next, based on the classical self-exciting point process, we incorporate a mutually exciting part to model social influence between individuals.

3.2.1 Self and Mutually Exciting Point Process

It can be seen that in social network study the event time of individual i not only depends on his own past activities, but also the event path of his friends. Hence, all the time points $T_{(1)}, \ldots, T_{(M_n)}$ are generated sequentially. For $t > T_{(k-1)}$ we introduce a latent event time $T_{j,k}$

for $j = 1, \ldots, n$, which is obtained by solving the following conditional intensity

$$\lambda_j(t) = \beta^T X_j + \rho \sum_{\ell=1}^{n} \int_0^t g(t-s) W_{j,\ell} dN_\ell(s) + \int_0^t g(t-s) dN_j(s), \tag{3.1}$$

where $\beta \in \mathbb{R}^p$ is the vector of regression coefficients for the baseline covariates, $g(\cdot)$ is assumed to be an unknown smooth and bounded decreasing function, and $\rho \in \mathbb{R}$ is the measure of social influence. The second term in (3.1) refers to the mutually-exciting part from the network effects, and the third term refers to the self-exciting part. Then, $T_{(k)}$ is defined as

$$T_{(k)} = \min_{j \in \{1,\ldots,n\}} T_{j,k} \text{ and } i_{(k)} = \arg \min_{j \in \{1,\ldots,n\}} T_{j,k}.$$

Let $\delta_{k,i}$ be a binary variable at time $T_{(k)}$ with $\delta_{k,i} = 1$ refering to $i_{(k)} = i$ (i.e. individual i took action at time $T_{(k)}$), and the rest $\delta_{k,j} = 0$ for $j = \{1, \ldots, n\} \backslash \{i_k\}$. Then, δ is a $M \times n$ matrix with row sums equal to 1, that is, $\sum_{j=1}^n \delta_{k,j} = 1$ for all $k = 1, \ldots, M_n$. For numerical computation, an alternative form of the conditional intensity model at time $T_{(k)}$ for individual j is

$$\lambda_j(t) = \beta^T X_j + \rho \sum_{\ell=1}^{k-1} W_{i_{(\ell)},j} g(t - T_{(\ell)}) + \sum_{\ell=1}^{k-1} \delta_{\ell,j} g(t - T_{(\ell)}). \tag{3.2}$$

Then, the log-likelihood for the parameters $\theta = (\beta^T, \rho, g)^T$ is

$$\begin{aligned}\ell_n(\theta) =& \frac{1}{nM}\sum_{i=1}^{n}\Big\{\sum_{k=1}^{M}\delta_{k,i}\log\lambda_i(T_{(k)}) - \int_0^{\tau}\lambda_i(t)dt\Big\} \\ =& \frac{1}{nM}\sum_{i=1}^{n}\sum_{k=1}^{M}\delta_{k,i}\log\Big\{\beta^T X_i + \sum_{\ell=1}^{k-1}\big(\rho W_{i_{(\ell)},i} + \delta_{\ell,i}\big)g(T_{(k)} - T_{(\ell)})\Big\} - \frac{\tau}{nM}\sum_{i=1}^{n}\beta^T X_i \\ & - \frac{\rho}{nM}\sum_{i=1}^{n}\sum_{\ell=1}^{n}W_{i,\ell}\sum_{k=1}^{M}\delta_{k,\ell}\int_0^{\tau - T_{(k)}} g(s)ds - \frac{1}{nM}\sum_{i=1}^{n}\sum_{k=1}^{M}\delta_{k,i}\int_0^{\tau - T_{(k)}} g(s)ds. \end{aligned} \tag{3.3}$$

3.3 A Monotone B-Spline Based Estimator

Here we derive a semi-parametric estimation method for this model based on monotone B-splines. Let ξ^n be a sequence of length $k_n + d$ such that $0 = \xi_1 = \cdots = \xi_d < \xi_{d+1} < \cdots < \xi_{k_n+1} = \cdots = \xi_{k_n+d} = \tau$, where $d > r$ is the order of B-spline basis functions and $k_n > d$ is the number of B-spline basis functions, depending on the sample size n, such that $k_n \to \infty$ as $n \to \infty$.

Then, we use the order d B-splines $\sum_{j=1}^{k_n}\gamma_j B_j^d(t)$ to approximate the excitation function $g(t)$, where $B_j^d(t)$ are the order d B-spline basis functions for $j = 1, \ldots, k_n$, defined recursively as

$$B_j^k(t) = \frac{t - \xi_j}{\xi_{j+k-1} - \xi_j}B_j^{k-1}(t) + \frac{\xi_{j+k} - t}{\xi_{j+k} - \xi_{j+1}}B_{j+1}^{k-1}(t), k = d, d-1, \ldots, 2,$$

and

$$B_j^1(t) = \begin{cases} I\big(\xi_j \le t < \xi_{j+1}\big), j \ne k_n, j \in \{1, \ldots, k_n + d - 1\}, \\ I\big(\xi_j \le t \le \xi_{j+1}\big), j = k_n. \end{cases}$$

To make sure the estimated excitation function $\hat{g}(t) = B(t)^T \hat{\gamma}$ is positive and decreasing, we can reparamtrize γ in terms of the logarithms of their successive differences, $\gamma_i^* = \log(\gamma_i - \gamma_{i+1})$ with $\gamma_{k_n+1} = 0$. Furthermore, the term $\int_0^{\tau - T_{(k)}} g(t)dt$ in (3.3) can be approximated by an order $d+1$ B-splines as follows

$$\int_0^{\tau - T_{(k)}} \sum_{j=1}^{k_n} \gamma_j B_j^d(t) dt = \sum_{j=1}^{k_n} \Big\{ \sum_{m=1}^{j} \gamma_m (\xi_{m+d} - \xi_m)/d \Big\} B_j^{d+1}(\tau - T_{(k)}).$$

Therefore, for the parameter $\theta^* = (\beta^T, \rho, \gamma^{*T})^T$ the reparametrized log-likelihood is

$$\begin{aligned}
\ell_n(\theta^*) =& \frac{1}{nM} \sum_{i=1}^{n} \sum_{k=1}^{M} \delta_{k,i} \log \Bigg\{ \beta^T X_i + \sum_{\ell=1}^{k-1} \big(\rho W_{i_{(\ell)},i} + \delta_{\ell,i}\big) \sum_{j=1}^{k_n} \Bigg(\sum_{q=j}^{k_n} e^{\gamma_q^*} \Bigg) B_j^d(T_{(k)} - T_{(\ell)}) \Bigg\} \\
& - \frac{\tau}{nM} \sum_{i=1}^{n} \beta^T X_i \\
& - \frac{\rho}{nM} \sum_{i=1}^{n} \sum_{\ell=1}^{n} W_{i,\ell} \sum_{k=1}^{M} \delta_{k,\ell} \times \sum_{j=1}^{k_n} \Bigg\{ \sum_{m=1}^{j} \Bigg(\sum_{q=m}^{k_n} e^{\gamma_q^*} \Bigg) (\xi_{m+d} - \xi_m)/d \Bigg\} B_j^{d+1}(\tau - T_{(k)}) \\
& - \frac{1}{nM} \sum_{i=1}^{n} \sum_{k=1}^{M} \delta_{k,i} \times \sum_{j=1}^{k_n} \Bigg\{ \sum_{m=1}^{j} \Bigg(\sum_{q=m}^{k_n} e^{\gamma_q^*} \Bigg) (\xi_{m+d} - \xi_m)/d \Bigg\} B_j^{d+1}(\tau - T_{(k)}).
\end{aligned} \tag{3.4}$$

For any fixed γ^*, $\ell_n(\theta^*)$ is a covex function of the parameters (β^T, ρ). Then, given a small $\epsilon > 0$, the estimates of θ^* can be found by the following procedure,

Step 1. Initialize $\gamma^{*(0)}$;

Step 2. Given $\gamma^{*(k-1)}$, $(\beta^T, \rho)^{(k)}$ can be solved simultaneously by maximizing $\ell_n(\beta^T, \rho | \gamma^{*(k-1)})$;

Step 3. Then, fix the value $(\beta^T, \rho)^{(k)}$, we update $\gamma^{*(k)}$ by maximizing $\ell_n(\gamma^* | (\beta^T, \rho)^{(k)})$;

Step 4. Iterate *Step 2* and *Step 3* until convergence is met,

$$\|(\beta^T, \rho)^{(k)} - (\beta^T, \rho)^{(k-1)}\| \leq \epsilon \left(1 + \|(\beta^T, \rho)^{(k)}\|\right),$$
$$\|\gamma^{*(k)} - \gamma^{*(k-1)}\| \leq \epsilon \left(1 + \|\gamma^{*(k)}\|\right).$$

Let $\hat{\theta}_n = (\hat{\beta}_n^T, \hat{\rho}_n, \hat{\gamma}_n^*)^T$ denote the resulting estimators at convergence. The observed information matrix $I(\hat{\theta}_n)$ can be calculated from Hessian matrix of $\hat{\theta}_n$. Then, the asymptotic variance-covariance matrix of $(\hat{\beta}_n^T, \hat{\rho}_n)$ can be estimated by $AI^{-1}(\hat{\theta}_n)A^T$, where $A = A_{(p+1)\times(p+1+k_n)} = (I_{p+1}, 0)$ is the identity matrix of size $p+1$ padded with zeros.

3.4 Asymptotic Properties

In this section we consider asymptotic properties of our estimators. We assume that the parameters of interest (β^T, ρ) belong to a bounded convex set in $\mathbb{R}^{p+1}$, denoted as $\mathscr{B}$. Furthermore, for some positve K and r, $g(t)$ is a decreasing function with $g(t) \leq g(0) \leq K$ for any $t \geq 0$, and r times continuously differentiable. Therefore, the parameter space for g is

$$\begin{aligned}\mathscr{F}_r =& \{h(t) : h(t) \text{ is } r \text{ times continuously differentiable and decreasing function },\\ & 0 \leq t \leq \tau \leq K\}.\end{aligned}$$

Then, the full parameter space for $\theta = (\beta^T, \rho, g)^T$ is $\Theta = \mathscr{B} \times \mathscr{F}_r$.

Let $\theta_0 = (\beta_0^T, \rho_0, g_0)^T$ denote the true value of the parameter θ, and $\|\cdot\|$ denote the Euclidean norm.

Condition 1. (β_0^T, ρ_0) is an interior point of $\mathscr{B}$.

Condition 2. For any $\beta \in \mathscr{B}$, there exists an $\epsilon > 0$ such that $\beta^T X \geq \epsilon$ almost surely.

Condition 3. There exists a constant $Q^* > 0$ such that $\|X\| \leq Q^*$ almost surely.

Condition 4. If for any $(\beta^{1T}, \rho^1)^T, (\beta^{2T}, \rho^2)^T \in \mathscr{B}$, $g^1, g^2 \in \mathscr{F}_r$ and $j = 1, \ldots, n$,

$$\begin{aligned}\beta^{1T} X_j + \rho^1 \sum_{\ell=1}^{n} \int_0^t g^1(t-s) W_{j,\ell} dN_\ell(s) + \int_0^t g^1(t-s) dN_j(s) \\ = \beta^{2T} X_j + \rho^2 \sum_{\ell=1}^{n} \int_0^t g^2(t-s) W_{j,\ell} dN_\ell(s) + \int_0^t g^2(t-s) dN_j(s)\end{aligned}$$

almost surely, then $(\beta^{1T}, \rho^1)^T = (\beta^{2T}, \rho^2)^T$ and $g^1 = g^2$.

Condition 5. For $\Delta(\xi) = \max_{d \leq i \leq k_n} |\xi_{i+1} - \xi_i|$ and $\delta(\xi) = \min_{d \leq i \leq k_n} |\xi_{i+1} - \xi_i|$, the sequence of knots ξ^n satisfies $\Delta(\xi^n) = O(n^{-q})$ for some $q \in (0, 1/2)$ and $\Delta(\xi^n)/\delta(\xi^n)$ is bounded.

Condition 1 is commonly assumed in the semi-parametric literature. Condition 2 guarantees the positivity of the baseline intensity. Condition 3 is typically satisfied in social network data set. Condition 4 makes sure the identifiablity of parameters. Similar as [Lu09] and [Zho98], Condition 5 is used to balance the model bias induced by the finite-dimensional approximation to the infinite-dimensional parameter when studying the asymptotic properties of B-spline based estimators.

Theorem 3.4.1 (Consistency) *Assume conditions 1-5 hold, the estimator $\hat{\theta}_n$ is consistent,*

$$\|\hat{\beta}_n - \beta_0\| + |\hat{\rho}_n - \rho_0| + \int_0^\tau |\hat{g}_n(s) - g_0(s)| ds \xrightarrow{P} 0.$$

Theorem 3.4.2 (Rate of Convergence) *Assume conditions 1-5 hold and the condition that*

$k_n \longrightarrow \infty$ as $n \longrightarrow \infty$, suppose that $\lim_{n\longrightarrow\infty} k_n/n = 0$. *Then*

$$\|\hat{\beta}_n - \beta_0\| + |\hat{\rho}_n - \rho_0| + \int_0^\tau |\hat{g}_n(s) - g_0(s)| ds = O_p\left((\frac{k_n}{n})^{1/2} + k_n^{-r} \right).$$

Theorem 3.4.3 (Asymptotic Normality) *In addition to the conditions of Theorem 2, and assume that* $\lim_{n\longrightarrow\infty} k_n^2/n = 0$ *and* $\lim_{n\longrightarrow\infty} n k_n^{-4r} = 0$. *Then, we have*

$$\sqrt{n}\left(\begin{pmatrix} \hat{\beta}_n \\ \hat{\rho}_n \end{pmatrix} - \begin{pmatrix} \beta_0 \\ \rho_0 \end{pmatrix} \right) \xrightarrow{D} N\left(\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_\beta & \Sigma_{\beta\rho} \\ \Sigma_{\rho\beta} & \Sigma_\rho \end{pmatrix} \right).$$

3.5 Simulation Studies

In this section, we illustrate the performance of the proposed estimators under several settings. Here, we consider a social network of size $n = 250$, or 500, with corresponding total number of events $M = 500$, or 750. The study duration τ is chosen to reach M. For each sample, we choose two different network structures for adjacency matrix W:

1. Random Graph (RG): $P(W_{i,j} = 1) = 0.01$ for $i \neq j$;

2. Stochastic Block Model (SBM): 3 blocks with $P(W_{i,j}|\text{ within block }) = 0.01$ and $P(W_{i,j}|\text{ between block }) = 0.001$, for $i \neq j$.

In particular, under stochastic block model the sample sizes of each block are given by $(200, 200, 100)$ and $(100, 100, 50)$ for $n = 500$ and $n = 250$, respectively. Moreover, for each setting, we consider two scenarios for covariate structure:

1. $\beta = (\beta_0, \beta_1)^T = (1, 1)^T$ and $X_i = (1, X_{i1})^T$, where $X_{i1} \overset{i.i.d}{\sim}$ Uniform[0, 1];

2. $\beta = (\beta_0, \beta_1, \beta_2)^T = (1,1,1)^T$ and $X_i = (1, X_{i1}, X_{i2})^T$, where $X_{i1} \overset{i.i.d}{\sim}$ Uniform[0,1] and $X_{i2} \overset{i.i.d}{\sim}$ Bernoulli(0.5).

Moreover, the excitation function is chosen as $g(t) = 4e^{-8t}$, $0 \le t \le \tau$ in all the cases. The event times $T_{(k)}$'s are generated sequentially following the descriptions given in Section 2.2. The order d of the B-spline was set to 3. The interior knots of the B-spline were evenly place in $[0,\tau]$: $(\xi_d, \ldots, \xi_k, \ldots, \xi_{k_n+1}) = (0, \tau/m, \ldots, k\tau/n, \ldots, \tau)$, where $m = k_n - d + 1$. Generally, a small number of B-spline basis function is good enough to approximate a given continuous function. In our settings, we choose $m = 5$. We found our simulation results are not very sensitive to the choice of m.

We conduct 500 simulation runs for each setting. Table 3.1 and Table 3.2 summarize the results under Scenarios 1 and 2, respectively. The results show that in all settings the proposed estimators are nearly unbiased, the standard error estimators are close to the standard deviations of the estimators, and the empirical coverage probabilities of the 95% Wald-type confidence intervals are close to the nominal level. In addition, the running time per simulation for of our proposed estimation method in Python for a social network with size $(n = 250, M = 500)$ and $(n = 500, M = 750)$ are around 15 seconds and 1 minute, respectively.

Figure 3.1 and Figure 3.2 show the point-wise 95% confidence interval and the median of the estimates of the excitation function $g(t)$, using the monotone B-spline method with $d = 3$ and $m = 5$, under Scenario 2 with the adjacency matrix W generated from RG and SBM, respectively. The figures indicate that the monotone B-spline estimators are nearly unbiased in all cases, and the true excitation function $g(t)$ is covered by the 95% point-wise confidence bands entirely, which suggests that the performance of the monotone B-spline method for estimating the excitation function is satisfactory. The estimation results of the

Table 3.1 $X = (1, X_1)^T$. SE, mean of estimated standard errors; SD, standard deviations of the estimates; CP, empirical coverage probability of 95% Wald-type confidence intervals.

			$n = 250$ and $M = 500$				$n = 500$ and $M = 750$			
		Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
RG	β_0	1	0.996	0.203	0.203	0.952	0.993	0.176	0.166	0.952
	β_1	1	1.011	0.365	0.360	0.946	0.984	0.302	0.303	0.952
	ρ	0.05	0.050	0.035	0.032	0.948	0.051	0.022	0.018	0.976
	β_0	1	0.994	0.195	0.194	0.948	0.994	0.167	0.153	0.960
	β_1	1	1.010	0.352	0.341	0.958	0.982	0.285	0.281	0.940
	ρ	0.025	0.026	0.031	0.029	0.968	0.026	0.019	0.016	0.986
	β_0	1	0.996	0.189	0.184	0.952	0.997	0.156	0.146	0.97
	β_1	1	1.007	0.339	0.320	0.972	0.983	0.267	0.268	0.932
	ρ	0	-0.000	0.028	0.027	0.954	0.001	0.016	0.015	0.968
		Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
SBM	β_0	1	0.991	0.186	0.179	0.956	0.998	0.154	0.149	0.952
	β_1	1	1.010	0.349	0.331	0.968	0.981	0.281	0.277	0.946
	ρ	0.05	0.047	0.048	0.046	0.942	0.051	0.030	0.028	0.956
	β_0	1	0.987	0.183	0.176	0.950	0.998	0.151	0.148	0.954
	β_1	1	1.014	0.344	0.326	0.962	0.982	0.274	0.274	0.948
	ρ	0.025	0.023	0.045	0.044	0.944	0.027	0.027	0.025	0.954
	β_0	1	0.990	0.181	0.175	0.946	0.998	0.147	0.143	0.956
	β_1	1	1.009	0.339	0.318	0.970	0.984	0.267	0.265	0.952
	ρ	0	-0.000	0.042	0.042	0.938	0.002	0.025	0.023	0.964

excitation function for other settings are similar and omitted here.

3.6 Real-World Data

We further evaluate our method on a college message temporal network from SNAP [2]. This dataset involves 1899 individuals over 193 days duration, and consists of private messages sent on an online social networking website at the University of California, Irvine. Individuals could search the network and then initiate a conversation with others. In

[2]https://snap.stanford.edu

Table 3.2 $X = (1, X_1, X_2)^T$. SE, mean of estimated standard errors; SD, standard deviations of the estimates; CP, empirical coverage probability of 95% Wald-type confidence intervals.

			$n = 250$ and $M = 500$				$n = 500$ and $M = 750$			
		Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
RG	β_0	1	0.977	0.278	0.272	0.948	0.982	0.247	0.248	0.928
	β_1	1	1.035	0.459	0.441	0.968	0.999	0.376	0.409	0.916
	β_2	1	1.001	0.277	0.258	0.970	0.988	0.226	0.228	0.938
	ρ	0.05	0.051	0.042	0.037	0.966	0.050	0.030	0.021	0.968
	β_0	1	0.979	0.266	0.260	0.932	0.981	0.229	0.236	0.934
	β_1	1	1.032	0.444	0.429	0.964	0.996	0.356	0.376	0.926
	β_2	1	1.000	0.270	0.245	0.960	0.988	0.214	0.217	0.950
	ρ	0.025	0.026	0.037	0.034	0.966	0.026	0.024	0.019	0.978
	β_0	1	0.987	0.260	0.254	0.942	0.977	0.213	0.222	0.940
	β_1	1	1.022	0.429	0.412	0.966	0.998	0.335	0.355	0.922
	β_2	1	0.998	0.260	0.237	0.968	0.993	0.203	0.207	0.936
	ρ	0	-0.001	0.034	0.031	0.960	0.002	0.020	0.018	0.966
		Trues	Estimates	SE	SD	CP	Estimates	SE	SD	CP
SBM	β_0	1	1.001	0.255	0.261	0.936	0.990	0.209	0.194	0.964
	β_1	1	1.004	0.442	0.458	0.948	0.971	0.353	0.335	0.958
	β_2	1	0.999	0.268	0.273	0.958	1.009	0.213	0.223	0.940
	ρ	0.05	0.051	0.056	0.056	0.942	0.053	0.035	0.032	0.966
	β_0	1	1.005	0.253	0.256	0.930	0.989	0.205	0.189	0.964
	β_1	1	1.000	0.436	0.446	0.950	0.975	0.345	0.322	0.964
	β_2	1	0.998	0.264	0.270	0.954	1.010	0.208	0.214	0.936
	ρ	0.025	0.023	0.053	0.052	0.942	0.028	0.032	0.030	0.970
	β_0	1	1.007	0.249	0.254	0.940	0.988	0.199	0.186	0.966
	β_1	1	0.998	0.430	0.440	0.944	0.976	0.336	0.319	0.960
	β_2	1	1.000	0.261	0.262	0.956	1.010	0.203	0.211	0.942
	ρ	0	-0.004	0.050	0.050	0.948	0.004	0.030	0.028	0.954

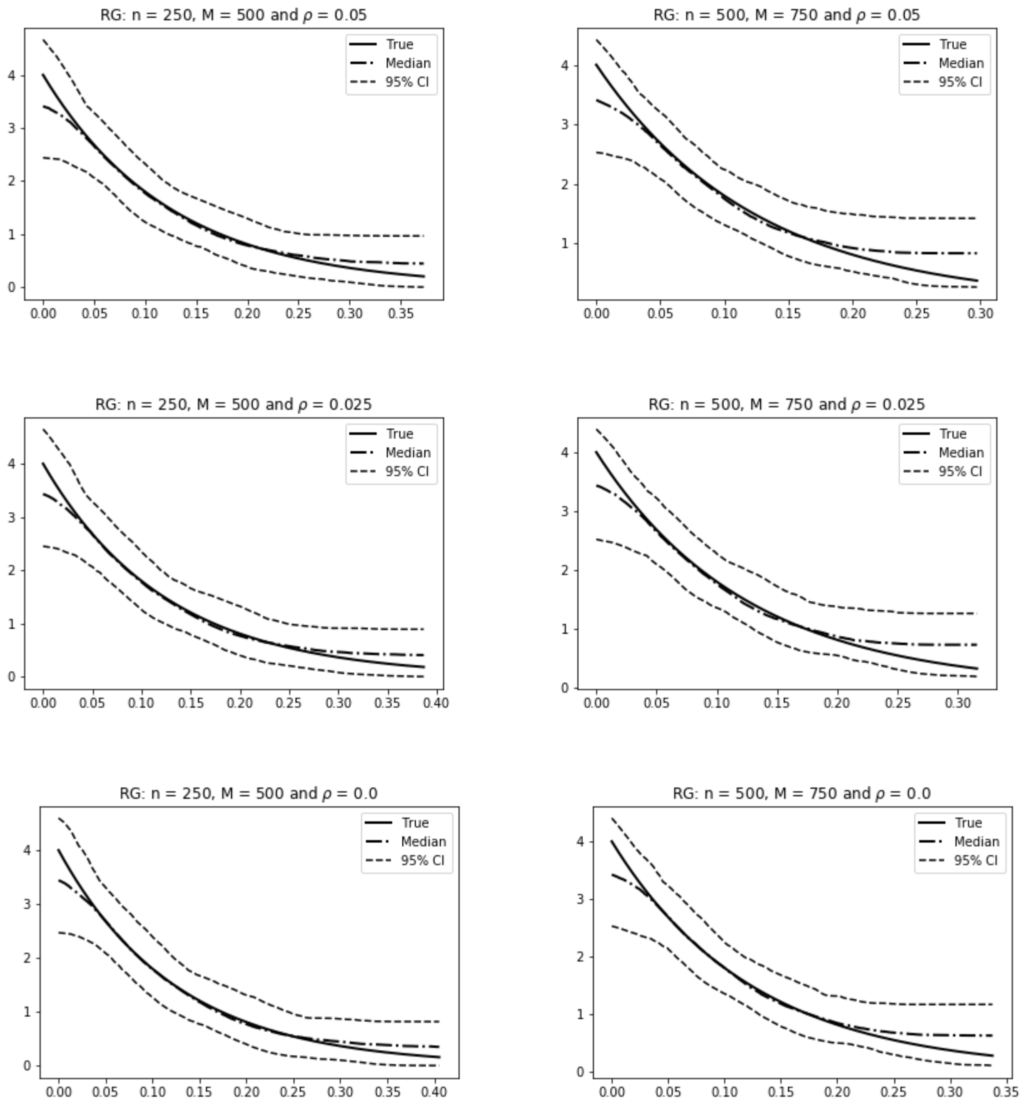

Figure 3.1 Plots of the true excitation function (solid curve), the point-wise 2.5th (dashed), 50th (dash-dotted) and 97.5th (dashed) percentiles of the 500 estimated excitation functions with the RG network structure: $(n, M) = (250, 500)$ (left panel) and $(n, M) = (500, 750)$ (right panel).

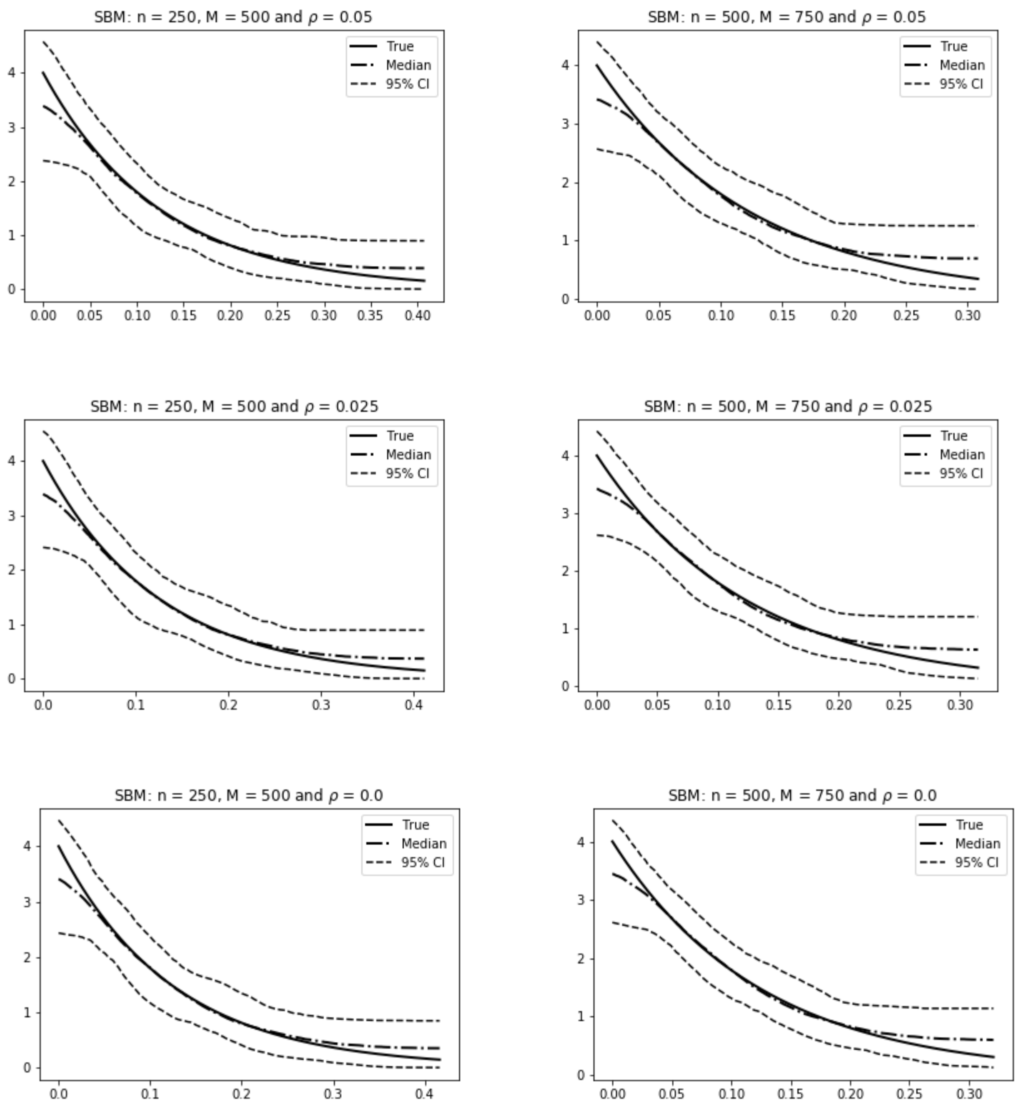

Figure 3.2 Plots of the true excitation function (solid curve), the point-wise 2.5th (dashed), 50th (dash-dotted) and 97.5th (dashed) percentiles of the 500 estimated excitation functions with the SBM network structure: $(n, M) = (250, 500)$ (left panel) and $(n, M) = (500, 750)$ (right panel).

this dataset, an temporal edge (u, v, t) means that individual u sent a private message to individual v at time t, which corresponds to the recurrent event in our model. There are 59835 temporal edges, or recurrent events, with 20296 edges in static graph over the study duration. We refer the 20296 static edges as the friendship connection between individuals. In this dataset, there is no baseline information for each individual. Hence, we only include the intercept for the baseline intensity in our model.

Due to the limited memory storage, we only run our model on subsets of the entire network. To be specific, we randomly select 2500 temporal edges, among which the static edges are used as friendship connection between individuals. The friendship connection of one randomly sampled sub-network is shown in Figure 3.3. It can be seen that there are some isolated nodes in the network. The rests are closely connected with each other.

We fit the proposed model to the college message network. As in simulations, we choose $d = 3$ and equally spaced interior knots $i/m \times 193$, $i = 0, \ldots, m$, with $m = 5$ for estimating the semi-parametric model with the proposed monotone B-spline estimator. We scale the events times from 0 to 1 when fitting our model. Based on 10 randomly selected sub-networks, the estimation results of model parameters β and ρ are given in Table 3.3. It can be seen that for all 10 randomly sampled sub-networks, the estimation results are similar. In particular, the estimates of the social influence parameter ρ are positive and significant. This implies the existence of mutually exciting influence coming from friends' behaviors. Specifically, the likelihood of an individual's future participation in an event is increased if many of his or her friends have taken actions before.

In addition, Figure 3.4 shows the monotone B-spline estimate of the excitation function $g(t)$ of one randomly sampled sub-network. The estimated excitation functions of other randomly selected sub-networks are almost the same as the one shown in Figure 3.4. From this figure we note that the excitation effect associated with the previous recurrent events

Table 3.3 Summary of estimation results from 10 randomly selected sub-networks.

	EST ×100	SE ×100	Z STAT
β_0	9.7 (0.84)	0.90 (0.037)	10.827 (0.503)
ρ	9.7 (0.58)	0.73 (0.082)	13.392 (0.927)

EST, mean of estimated parameters; SE, mean of estimated standard errors; Z STAT, mean of Z-statistics. Standard deviations are reported in brackets.

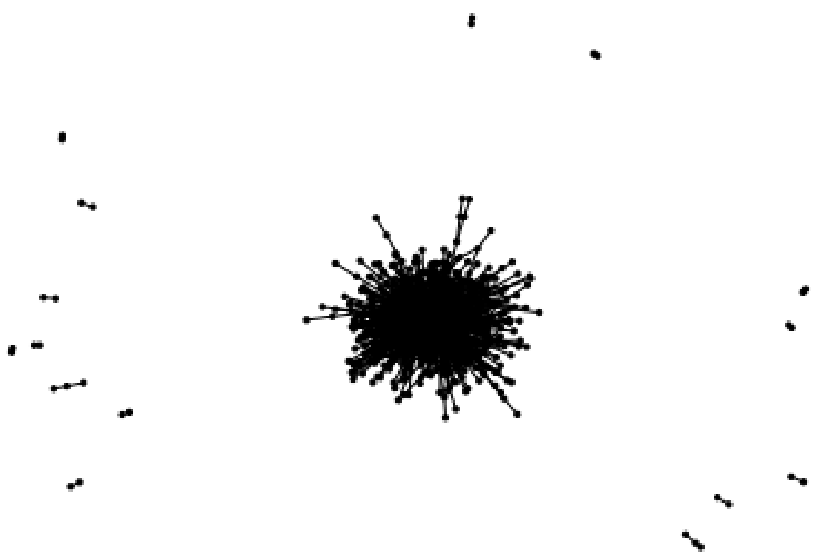

Figure 3.3 College Message Network Visualization for randomly selected 2500 temporal edges

decays quickly over time, and eventually approaches zero after almost one month.

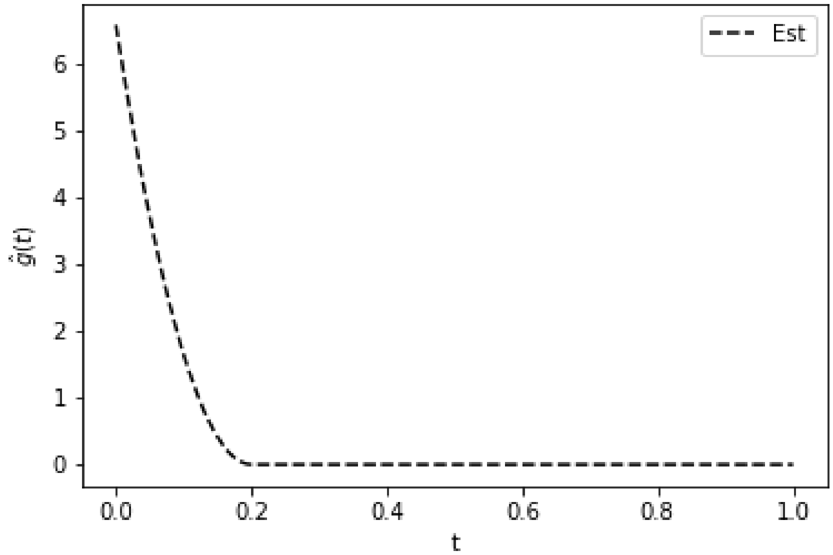

Figure 3.4 The estimated excitation functions for one randomly selected sub-network.

3.7 Discussion

In this paper, we propose a novel method to model the influence of social actions with recurrent events data based on the self and mutually exciting point process. The asymptotic properties of the resulting estimators are established. In our current work, the self-excitation function and mutual excitation function share the same form, but the mutual excitation effect from friends' actions is multiplied by a social influence parameter ρ. In general, we can consider different monotone nonincreasing functions for the self-excitation and mutual excitation. Moreover, in our current work, the friendship network is treated as fixed. In some follow-up studies, it may vary over time. It would be desirable to treat the adjacency matrix as a time-dependent covariate and generalize our model and method for dynamic social networks. These are interesting research topics that need further investigation.

BIBLIOGRAPHY

[Air08] Airoldi, E. M. et al. "Mixed membership stochastic blockmodels". *Journal of Machine Learning Research* **9**.Sep (2008), pp. 1981–2014.

[Ana08] Anagnostopoulos, A. et al. "Influence and correlation in social networks". *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2008, pp. 7–15.

[Ana12] Anagnostopoulos, A. et al. "Online team formation in social networks". *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 839–848.

[AW12] Aral, S. & Walker, D. "Identifying influential and susceptible members of social networks". *Science* (2012), p. 1215842.

[Bai15] Bai, F. et al. "SEMIPARAMETRIC ESTIMATION OF A SELF-EXCITING REGRESSION MODEL WITH AN APPLICATION IN RECURRENT EVENT DATA ANALYSIS". *Statistica Sinica* (2015), pp. 1503–1526.

[Bak12] Bakshy, E. et al. "The role of social networks in information diffusion". *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 519–528.

[BH09] Bauwens, L. & Hautsch, N. "Modelling financial high frequency data using point processes". *Handbook of financial time series*. Springer, 2009, pp. 953–979.

[Ben83] Bennett, S. "Analysis of survival data by the proportional odds model". *Statistics in medicine* **2**.2 (1983), pp. 273–277.

[Bow07] Bowsher, C. G. "Modelling security market events in continuous time: Intensity based, multivariate point process models". *Journal of Econometrics* **141**.2 (2007), pp. 876–912.

[Bra09] Bramoullé, Y. et al. "Identification of peer effects through social networks". *Journal of econometrics* **150**.1 (2009), pp. 41–55.

[Bur07] Burk, W. J. et al. "Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors". *International journal of behavioral development* **31**.4 (2007), pp. 397–404.

[Cen10] Centola, D. "The spread of behavior in an online social network experiment". *science* **329**.5996 (2010), pp. 1194–1197.

[CC06] Cheatham, M. & Cleereman, K. “Application of social network analysis to collaborative team formation”. *Collaborative Technologies and Systems, 2006. CTS 2006. International Symposium on*. IEEE. 2006, pp. 306–311.

[Che09a] Chen, N. “On the approximability of influence in social networks”. *SIAM Journal on Discrete Mathematics* **23**.3 (2009), pp. 1400–1415.

[Che09b] Chen, W. et al. “Efficient influence maximization in social networks”. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2009, pp. 199–208.

[Che95] Cheng, S. et al. “Analysis of transformation models with censored data”. *Biometrika* **82**.4 (1995), pp. 835–845.

[CL10] Cheung, C. M. & Lee, M. K. “A theoretical model of intentional social action in online social networks”. *Decision support systems* **49**.1 (2010), pp. 24–30.

[Chu08] Chun, H. et al. “Comparison of online social relations in volume vs interaction: a case study of cyworld”. *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*. ACM. 2008, pp. 57–70.

[Cou03] Council, N. R. et al. *Dynamic social network modeling and analysis: Workshop summary and papers*. National Academies Press, 2003.

[Cox72] Cox, D. R. “Regression Models and Life-Tables”. *Journal of the Royal Statistical Society. Series B (Methodological)* **34**.2 (1972), pp. 187–220.

[DD88a] Dabrowska, D. M. & Doksum, K. A. “Estimation and testing in a two-sample generalized odds-rate model”. *Journal of the American Statistical Association* **83**.403 (1988), pp. 744–749.

[DD88b] Dabrowska, D. M. & Doksum, K. A. “Partial likelihood in transformation models with censored data”. *Scandinavian journal of statistics* **15** (1988), pp. 1–23.

[Emb11] Embrechts, P. et al. “Multivariate Hawkes processes: an application to financial data”. *Journal of Applied Probability* **48**.A (2011), pp. 367–378.

[Gjo08] Gjoka, M. et al. “Poking facebook: characterization of osn applications”. *Proceedings of the first workshop on Online social networks*. ACM. 2008, pp. 31–36.

[Gol07] Golder, S. A. et al. “Rhythms of social interaction: Messaging within a massive online network”. *Communities and technologies 2007* (2007), pp. 41–66.

[Gui05] Guimera, R. et al. “Team assembly mechanisms determine collaboration network structure and team performance”. *Science* **308**.5722 (2005), pp. 697–702.

[Han07] Handcock, M. S. et al. “Model-based clustering for social networks”. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**.2 (2007), pp. 301–354.

[Haw71] Hawkes, A. G. “Spectra of some self-exciting and mutually exciting point processes”. *Biometrika* (1971), pp. 83–90.

[Hof02] Hoff, P. D. et al. “Latent space approaches to social network analysis”. *Journal of the american Statistical association* **97**.460 (2002), pp. 1090–1098.

[HL81] Holland, P. W. & Leinhardt, S. “An exponential family of probability distributions for directed graphs”. *Journal of the american Statistical association* **76**.373 (1981), pp. 33–50.

[Hub08] Huberman, B. A. et al. “Social networks that matter: Twitter under the microscope” (2008).

[KN11] Karrer, B. & Newman, M. E. “Stochastic blockmodels and community structure in networks”. *Physical Review E* **83**.1 (2011), p. 016107.

[Kem03] Kempe, D. et al. “Maximizing the spread of influence through a social network”. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, pp. 137–146.

[KL16] Kobayashi, R. & Lambiotte, R. “TiDeH: Time-Dependent Hawkes Process for Predicting Retweet Dynamics.” *ICWSM*. 2016, pp. 191–200.

[Lee04] Lee, L.-F. “Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models”. *Econometrica* **72**.6 (2004), pp. 1899–1925.

[Lee10] Lee, L.-f. et al. “Specification and estimation of social interaction models with network structures”. *The Econometrics Journal* **13**.2 (2010), pp. 145–176.

[Lei06] Leibnitz, K. et al. “Modeling of epidemic diffusion in peer-to-peer file-sharing networks”. *International Workshop on Biologically Inspired Approaches to Advanced Information Technology*. Springer. 2006, pp. 322–329.

[Les08] Leskovec, J. et al. “Statistical properties of community structure in large social and information networks”. *Proceedings of the 17th international conference on World Wide Web*. ACM. 2008, pp. 695–704.

[Li16] Li, T. et al. “Prediction models for network-linked data”. *arXiv preprint arXiv:1602.01192* (2016).

[LNK07] Liben-Nowell, D. & Kleinberg, J. “The link-prediction problem for social networks”. *journal of the Association for Information Science and Technology* **58**.7 (2007), pp. 1019–1031.

[LH07] Lu, H.-P. & Hsiao, K.-L. “Understanding intention to continuously share information on weblogs”. *Internet research* **17**.4 (2007), pp. 345–361.

[Lu07] Lu, M. et al. “Estimation of the mean function with panel count data using monotone polynomial splines”. *Biometrika* **94**.3 (2007), pp. 705–718.

[Lu09] Lu, M. et al. “Semiparametric estimation methods for panel count data using monotone B-splines”. *Journal of the American Statistical Association* **104**.487 (2009), pp. 1060–1070.

[Lu08] Lu, W. “Maximum likelihood estimation in the proportional hazards cure model”. *Annals of the Institute of Statistical Mathematics* **60**.3 (2008), pp. 545–574.

[Mas13] Masuda, N. et al. “Self-exciting point process modeling of conversation event sequences”. *Temporal Networks*. Springer, 2013, pp. 245–264.

[Mil09] Miller, K. et al. “Nonparametric latent feature models for link prediction”. *Advances in Neural Information Processing Systems 22*. 2009, pp. 1276–1284.

[MC09] Mitchell, L. & Cates, M. E. “Hawkes process as a model of social interactions: a view on video dynamics”. *Journal of Physics A: Mathematical and Theoretical* **43**.4 (2009), p. 045101.

[MN00] Moore, C. & Newman, M. E. “Epidemics and percolation in small-world networks”. *Physical Review E* **61**.5 (2000), p. 5678.

[Mor34] Moreno, J. L. “Who shall survive?: A new approach to the problem of human interrelations.” (1934).

[Mor93] Morris, M. “Epidemiology and social networks: Modeling structured diffusion”. *Sociological methods & research* **22**.1 (1993), pp. 99–126.

[MR07] Mossel, E. & Roch, S. "On the submodularity of influence in social networks". *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM. 2007, pp. 128–134.

[MM07] Murata, T. & Moriyasu, S. "Link prediction of social networks based on weighted proximity measures". *Proceedings of the IEEE/WIC/ACM international conference on web intelligence*. IEEE Computer Society. 2007, pp. 85–88.

[Mur95] Murphy, S. "Asymptotic theory for the frailty model". *The Annals of Statistics* **23**.1 (1995), pp. 182–198.

[Naz08] Nazir, A. et al. "Unveiling facebook: a measurement study of social network based applications". *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*. ACM. 2008, pp. 43–56.

[Ngu11] Nguyen, N. P. et al. "Adaptive algorithms for detecting community structure in dynamic social networks". *INFOCOM, 2011 Proceedings IEEE*. IEEE. 2011, pp. 2282–2290.

[Oga88] Ogata, Y. "Statistical models for earthquake occurrences and residual analysis for point processes". *Journal of the American Statistical association* **83**.401 (1988), pp. 9–27.

[Pet84] Pettitt, A. "Proportional odds models for survival data and estimates using ranks". *Applied Statistics* (1984), pp. 169–175.

[Roh11] Rohe, K. et al. "Spectral clustering and the high-dimensional stochastic blockmodel". *The Annals of Statistics* **39**.4 (2011), pp. 1878–1915.

[SM06] Sarkar, P. & Moore, A. W. "Dynamic social network analysis using latent space models". *Advances in Neural Information Processing Systems 18*. MIT Press, 2006, pp. 1145–1152.

[Sch98] Scharfstein, D. O. et al. "Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data". *Lifetime data analysis* **4**.4 (1998), pp. 355–391.

[Sch09] Schneider, F. et al. "Understanding online social network usage from a network perspective". *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*. ACM. 2009, pp. 35–48.

[She14] Shen, H.-W. et al. "Modeling and Predicting Popularity Dynamics via Reinforced Poisson Processes." *AAAI*. Vol. 14. 2014, pp. 291–297.

[SDX13] Stieglitz, S. & Dang-Xuan, L. “Emotions and information diffusion in social mediaâĂŤsentiment of microblogs and sharing behavior”. *Journal of management information systems* **29**.4 (2013), pp. 217–248.

[Sus12] Sussman, D. L. et al. “A consistent adjacency spectral embedding for stochastic blockmodel graphs”. *Journal of the American Statistical Association* **107**.499 (2012), pp. 1119–1128.

[Tan07] Tantipathananandh, C. et al. “A framework for community identification in dynamic social networks”. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2007, pp. 717–726.

[VW96] Vaart, A. W. van der & Wellner, J. A. “Springer series in statistics”. *Weak convergence and empirical processesSpringer, New York Google Scholar* (1996).

[WW87] Wang, Y. J. & Wong, G. Y. “Stochastic blockmodels for directed graphs”. *Journal of the American Statistical Association* **82**.397 (1987), pp. 8–19.

[Wil09] Wilson, C. et al. “User interactions in social networks and their implications”. *Proceedings of the 4th ACM European conference on Computer systems*. Acm. 2009, pp. 205–218.

[Xie13] Xie, J. et al. “Overlapping community detection in networks: The state-of-the-art and comparative study”. *Acm computing surveys (csur)* **45**.4 (2013), p. 43.

[Xin10] Xing, E. P. et al. “A state-space mixed membership blockmodel for dynamic network tomography”. *The Annals of Applied Statistics* **4**.2 (2010), pp. 535–566.

[Ye12] Ye, M. et al. “Exploring social influence for recommendation: a generative model approach”. *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 671–680.

[ZL07] Zeng, D & Lin, D. “Maximum likelihood estimation in semiparametric regression models with censored data”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**.4 (2007), pp. 507–564.

[ZL06] Zeng, D. & Lin, D. “Efficient estimation of semiparametric transformation models for counting processes”. *Biometrika* **93**.3 (2006), pp. 627–640.

[Zha15] Zhao, Q. et al. “Seismic: A self-exciting point process model for predicting tweet popularity”. *Proceedings of the 21th ACM SIGKDD International Con-*

ference on Knowledge Discovery and Data Mining. ACM. 2015, pp. 1513–1522.

[Zha12] Zhao, Y. et al. "Consistency of community detection in networks under degree-corrected stochastic block models". *The Annals of Statistics* **40**.4 (2012), pp. 2266–2292.

[Zho18] Zhou, J. et al. "Estimating Spatial Autocorrelation with Sampled Network Data". *Journal of Business & Economic Statistics* in press (2018).

[Zho13] Zhou, K. et al. "Learning Social Infectivity in Sparse Low-rank Networks Using Multi-dimensional Hawkes Processes." *AISTATS*. Vol. 31. 2013, pp. 641–649.

[Zho98] Zhou, S et al. "Local asymptotics for regression splines and confidence regions". *The annals of statistics* **26**.5 (1998), pp. 1760–1782.

APPENDICES

APPENDIX

A

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

Our proofs follow similar steps as [Mur95]; [Sch98]; [Lu08]. However, the difference lies on the fact that the event times of individuals are no longer independent and identically distributed as in classical survival data. We need to represent the data sequentially based on the ordered event times as in the data generation process and define the associated martingale processes. Specifically, define the martingale process as

$$M_{j,k}(t) = N_{j,k}(t) - \int_0^t Y_{j,k}(u)\dot{G}\left\{\Lambda(u)e^{\theta^T X_{j,k}}\right\} e^{\theta^T X_{j,k}} d\Lambda(u),$$

where $N_{j,k}(t) = \mathrm{I}\,(T_{(k-1)} < T_{j,k} \le t)$ and $Y_{j,k}(t) = \mathrm{I}\,(T_{j,k} \ge t > T_{(k-1)})$ for individual $j \in N_k$. For simplicity, we only consider the link function $G(x) = \frac{1}{s}\log(1 + sx)$ in our proofs.

We can rewrite the log-likelihood as

$$\begin{aligned}
\ell_n(\theta, \Lambda) &= \sum_{k=1}^{M_n} \sum_{j \in N_k} \ell_{j,k}(\theta, \Lambda) \\
&= \sum_{k=1}^{M_n} \sum_{j \in N_k} \Bigg(\int_0^{T_{(k)}} \log\Big[\lambda(t)e^{\theta^T X_{j,k}} \dot{G}\left\{\Lambda(t)e^{\theta^T X_{j,k}}\right\}\Big] dN_{j,k}(t) \\
&\qquad - \int_0^{T_{(k)}} Y_{j,k}(t)e^{\theta^T X_{j,k}} \dot{G}\left\{\Lambda(t)e^{\theta^T X_{j,k}}\right\} d\Lambda(t)\Bigg).
\end{aligned} \tag{A.1}$$

Next, consider one-dimensional submodel $\Lambda_d(t) = \int_0^t \{1 + dh_1(u)\} d\hat{\Lambda}_n(u)$ and $\theta_d = dh_2 + \hat{\theta}_n$, where h_1 is a function and h_2 is a $(p+1)$-dimensional vector. Let $S_n(\hat{\Lambda}_n, \hat{\theta}_n)(h_1, h_2)$ denote the first derivative of $\ell_n(\theta_d, \Lambda_d)$ with respect to d and evaluated at $d = 0$. Then, we have $S_n(\hat{\Lambda}_n, \hat{\theta}_n)(h_1, h_2) = 0$ for all (h_1, h_2), since $(\hat{\Lambda}_n, \hat{\theta}_n)$ maximizes $\ell_n(\theta, \Lambda)$. In addition, S_n can be written as $S_n = S_{n_1} + S_{n_2}$, where

$$S_{n_1}(\hat{\Lambda}_n, \hat{\theta}_n)(h_1) = \sum_{k=1}^{M_n} \sum_{j \in N_k} \int_0^{T_{(k)}} \left[h_1(t) + \frac{\ddot{G}\left\{\hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}}\right\}}{\dot{G}\left\{\hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}}\right\}} e^{\hat{\theta}_n^T X_{j,k}} \int_0^t h_1(v) d\hat{\Lambda}_n(v) \right]$$
$$\times \left[dN_{j,k}(t) - Y_{j,k}(t) e^{\hat{\theta}_n^T X_{j,k}} \dot{G}\left\{\hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}}\right\} d\hat{\Lambda}_n(t) \right], \tag{A.2}$$
$$S_{n_2}(\hat{\Lambda}_n, \hat{\theta}_n)(h_2) = \sum_{k=1}^{M_n} \sum_{j \in N_k} \int_0^{T_{(k)}} \left[h_2^T X_{j,k} + \frac{\ddot{G}\left\{\hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}}\right\}}{\dot{G}\left\{\hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}}\right\}} \hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}} h_2^T X_{j,k} \right]$$
$$\times \left[dN_{j,k}(t) - Y_{j,k}(t) e^{\hat{\theta}_n^T X_{j,k}} \dot{G}\left\{\hat{\Lambda}_n(t) e^{\hat{\theta}_n^T X_{j,k}}\right\} d\hat{\Lambda}_n(t) \right], \tag{A.3}$$

where $\ddot{G}(u) = d\dot{G}(u)/du$.

After some calculations, we can show that the efficient score for θ is given by

$$S_{\text{eff}} = \sum_{k=1}^{M_n} \sum_{j \in N_k} \int_0^{T_{(k)}} \left\{ \frac{X_{j,k}}{1 + s\Lambda_0(t) e^{\theta_0^T X_{j,k}}} - w_{\text{eff}}(t) + \frac{s \int_0^t \lambda_0(v) e^{\theta_0^T X_{j,k}} w(v) dv}{1 + s\Lambda_0(t) e^{\theta_0^T X_{j,k}}} \right\} dM_{j,k}(t)$$
$$\equiv \sum_{k=1}^{M_n} S_{\text{eff},k},$$

where $w_{\text{eff}}(t)$ is a solution to the following integral equation

$$w(t) - \int_0^\tau Q(t, v) w(v) d\Lambda_0(v) = f(t), \quad t \in [0, \tau],$$

and

$$
\begin{aligned}
Q(t,v)=&\left[E\left\{\sum_{k=1}^{M_n}\sum_{j\in N_k}\mathrm{I}(t\le T_{(k)})Y_{j,k}(t)\frac{e^{\theta_0 X_{j,k}}}{1+s\Lambda_0(t)e^{\theta_0 X_{j,k}}}\right\}\right]^{-1}\\
&\times\Bigg(E\left[\sum_{k=1}^{M_n}\sum_{j\in N_k}\frac{s\mathrm{I}(v\vee t\le T_{(k)})Y_{j,k}(t)e^{2\theta_0^T X_{j,k}}}{(1+s\Lambda_0(t)e^{\theta_0^T X_{j,k}})^2}\right]\\
&\quad -E\left[\sum_{k=1}^{M_n}\sum_{j\in N_k}\int_{v\vee t}^{T_{(k)}}\frac{s^2 Y_{j,k}(u)e^{3\theta_0^T X_{j,k}}}{(1+s\Lambda_0(u)e^{\theta_0^T X_{j,k}})^3}d\Lambda_0(u)\right]\Bigg),\\
f(t)=&\left[E\left\{\sum_{k=1}^{M_n}\sum_{j\in N_k}\mathrm{I}(t\le T_{(k)})Y_{j,k}(t)\frac{e^{\theta_0 X_{j,k}}}{1+s\Lambda_0(t)e^{\theta_0 X_{j,k}}}\right\}\right]^{-1}\\
&\times\Bigg(E\left[\sum_{k=1}^{M_n}\sum_{j\in N_k}\frac{X_{j,k}\mathrm{I}(t\le T_{(k)})Y_{j,k}(t)e^{\theta_0^T X_{j,k}}}{(1+s\Lambda_0(t)e^{\theta_0^T X_{j,k}})^2}\right]\\
&\quad -E\left[\sum_{k=1}^{M_n}\sum_{j\in N_k}\int_{t}^{T_{(k)}}\frac{sX_{j,k}Y_{j,k}(u)e^{2\theta_0^T X_{j,k}}}{(1+s\Lambda_0(u)e^{\theta_0^T X_{j,k}})^3}d\Lambda_0(u)\right]\Bigg).
\end{aligned}
$$

Here, $v\vee t=\max(v,t)$. Note that the terms $S_{\mathrm{eff},k}$ and $S_{\mathrm{eff},k}^T$ are uncorrelated for any $k\neq k'$. Then, the information matrix for θ_0 can be defined by $\mathrm{I}(\theta_0)=\lim_{n\to\infty}n^{-1}\sum_{k=1}^{M_n}E(S_{\mathrm{eff},k},S_{\mathrm{eff},k}^T)$.

A.1 Proof of Consistency

Proof of Theorem 2.3.1. The proof of consistency consists of three steps: first we show that the nonparametric maximum likelihood estimators $\hat{\Lambda}_n$ and $\hat{\theta}_n$ exist or that the jump sizes of $\hat{\Lambda}_n$ are finite; next we show that $\hat{\Lambda}_n$ is bounded almost surely so that, along a subsequence, $\hat{\Lambda}_{n_m}(t)\to\Lambda^*(t)$ for all $t\in[0,\tau]$ and $\hat{\theta}_{n_m}\to\theta^*$; finally we show that $\Lambda^*=\Lambda_0$ and $\beta^*=\beta_0$.

Step 1. By Condition 1, $\hat{\theta}_n$ is finite, and we have $\sup_{\theta\in\mathscr{C}}|\theta^T X_{j,k}|\le M_0$ for some constant

$M_0 > 0$ for all j and k. Therefore, the log-likelihood (2.3) is bounded above by

$$\ell_n(\theta,\Lambda) < \sum_{k=1}^{M_n}\Bigg(\log\Lambda\{T_{(k)}\}+M_0-\log\left\{1+s\sum_{\ell=1}^{k-1}\Lambda\{T_{(\ell)}\}e^{-M_0}\right\}$$
$$-|N_k|\left[G\left\{\sum_{\ell=1}^{k}\Lambda(T_{(\ell)})e^{-M_0}\right\}-G\left\{\sum_{\ell=1}^{k-1}\Lambda(T_{(\ell)})e^{-M_0}\right\}\right]\Bigg), \qquad \text{(A.4)}$$

where $|N_k|$ is the number of elements in N_k. The right-hand side in (A.4) diverges to $-\infty$ if $\Lambda\{T_{(k)}\}$ goes to infinity for some k, which contradicts to the property of log-likelihood.

Step 2. By the property $n^{-1}\{\ell(\hat{\Lambda}_n,\hat{\theta}_n)-\ell(\bar{\Lambda}_n,\hat{\theta}_n)\}\geq 0$ with $\bar{\Lambda}_n=\hat{\Lambda}_n/\hat{\Lambda}_n(\tau)$, and following the similar steps as [ZL06], we can show that $\sup_n\hat{\Lambda}_n(\tau)<\infty$.

Step 3. Define the following quanity

$$\tilde{\Lambda}_n(t)=\int_0^t\sum_{k=1}^n\frac{\mathrm{I}\,(T_{(k)}\leq u)\sum_{j\in N_k}dN_{j,k}(u)}{\sum_{j=1}^n\mathrm{I}\,(T_{(k)}\geq u)Y_{j,k}(u)e^{\theta_0^TX_{j,k}}/\{1+s\Lambda_0(u-)e^{\theta_0^TX_{j,k}}\}}, \qquad \text{(A.5)}$$

which is a step function with jumps at $T_{(k)}$'s and converges uniformly to Λ_0 by uniform weak law of large numbers.

By Helly's theorem, we know that there exists convergent subsequences $\{\hat{\theta}_{n_m}\}$ and $\{\hat{\Lambda}_{n_m}\}$ such that $\hat{\theta}_{n_m}\to\theta^*$ and $\hat{\Lambda}_{n_m}(t)\to\Lambda^*(t)$ for all $t\in[0,\tau]$. Furthermore, we have $n^{-1}\{\ell(\hat{\Lambda}_{n_m},\hat{\theta}_{n_m})-\ell(\tilde{\Lambda}_{n_m},\theta_0)\}\geq 0$. By taking limits on both sides we obtain $E\{\ell(\Lambda^*,\theta^*)\}=E\{\ell(\Lambda_0,\theta_0)\}$, since the Kullback-Leibler information is negative.

Recall in term (A.1), we have

$$\ell_{j,k}(\Lambda,\theta)=\int_{T_{(k-1)}}^{T_{(k)}}\log\left\{\frac{\lambda(t)e^{\theta X_{j,k}}}{1+s\Lambda(t-)e^{\theta X_{j,k}}}\right\}dN_{j,k}(t)-\int_{T_{(k-1)}}^{T_{(k)}}Y_{j,k}(t)\frac{e^{\theta X_{j,k}}}{1+s\Lambda(t)e^{\theta X_{j,k}}}d\Lambda(t).$$

Then, the above equality holds if and only if $E\{\ell_{j,k}(\Lambda^*,\theta^*)\}=E\{\ell_{j,k}(\Lambda_0,\theta_0)\}$ for all j and k. Next, for $k=1,\ldots,M_n$ consider two cases: (1) $N_{j,k}(T_{(k)})=0,\ Y_{j,k}(T_{(k)})=1$ for some $j\in N_k$, and (2) $N_{j,k}(T_{(k)})=1$, $N_{j,k}(t-)=0$, and $Y_{j,k}(T_{(k)})=1$ for some $j\in N_k$ and t is between time $T_{(k-1)}$ and $T_{(k)}$. By taking difference between the equalities from two cases above, for all $t\in[0,\tau]$ we conclude that

$$\frac{\lambda^*(t)e^{\theta^{*T}X_{j,k}}}{1+s\Lambda^*(t)e^{\theta^{*T}X_{j,k}}}=\frac{\lambda_0(t)e^{\theta_0^TX_{j,k}}}{1+s\Lambda_0(t)e^{\theta_0^TX_{j,k}}}.$$

Then, integrating from 0 to t on both sides of above equality and by some simple algebra, we have

$$\Lambda^*(t)/\Lambda_0(t)=e^{(\theta_0-\theta^*)^TX_{j,k}},\quad \text{for all } t\in[T_{(k-1)},T_{(k)}] \text{ and } k=1,\ldots,M_n$$

By Condition 2, we have that $E\{\ell(\Lambda^*,\theta^*)\}=E\{\ell(\Lambda_0,\theta_0)\}$ if and only if $\Lambda^*=\Lambda_0$ and $\theta^*=\theta_0$. Therefore, we show that the subsequences $(\hat{\Lambda}_{n_m},\hat{\theta}_{n_m})\to(\Lambda_0,\theta_0)$. By Helly's theorem, we know that $(\hat{\Lambda}_n,\hat{\theta}_n)$ must also converge to (Λ_0,θ_0) almost surely. Since $\hat{\Lambda}_0$ and Λ_0 are bounded monotone function, the pointwise convergence can be strengthened to uniform convergence on $[0,\tau]$.

A.2 Proof of Asymptotic Normality

Here, we give an outline of the proof. Define $\psi_0=(\Lambda_0,\theta_0)$, $\psi=(\Lambda,\theta)$ and $h=(h_1,h_2)$. Assume that the class of h belongs to the space $H=B\otimes R^{p+1}$, where B is the space of bounded variation functions defined on $[0,\tau]$. Define the norm $||h||_H=||h_1||_v+|h_2|_1$, where $||h_1||_v$ is the total variation norm on $[0,\tau]$ and $|h_2|_1$ is the L_1-norm. In addition, define $H_m=\{h\in H:||h||_H\le m\}$. Assume $\psi\subset\ell^\infty(H_m)$, where $\ell^\infty(H_m)$ is the space of bounded real-valued functions on H_m under the supremum norm $||A(h)||=\sup_{h\in H_m}|A(h)|$. First, by the martingale central limit theorem, we can show that $n^{-1/2}S_n(\psi_0)(h)$ converges weakly to a tight Gaussian process G on $\ell^{\infty}(H_m)$. Define $S(\psi)(h) = \lim_{n\to\infty} n^{-1} S_n(\psi)(h)$. We have $S(\psi_0)(h) = 0$. Then, following similar arguments in [Sch98] and [Lu08], we can show that $S(\psi)(h)$ is Fréchet differentiable, and its derivative $\dot{S}(\psi)(h)$ is a continuous linear operator and continuously invertible on its range. Finally, by the maximal inequality for martingales (Theorem 2.3. Nishiyama, 1999), we have

$$||n^{-1/2}\{(S_n - S)(\psi_n) - (S_n - S)(\psi_0)\}|| = o_{p^*}(1).$$

for any $||\psi_n - \psi_0|| = O_p(n^{-1/2})$. Therefore, $n^{-1/2}(\hat{\psi}_n - \psi_0)(h)$ converges weakly to the Gaussian process $-\{\dot{S}(\psi)\}^{-1}G$. Then, following similar arguments in [Lu08], we can show that $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges in distribution to a multivariate normal with mean 0 and variance $\{\mathrm{I}(\theta_0)\}^{-1}$.

APPENDIX

B

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

The proofs follow similar steps as [Bai15]. However, in our model the individuals are not independent and identically distributed by the conditional intensity, which is differenct as in classcial self-exciting point process. Here, we need to take into account all the data points in the likelihood function. The details are shown below.

In this study, the parameters of interest are $\theta = (\beta^T, \rho, g)^T$. For any $\theta_1, \theta_2 \in \Theta$, define a semi-metric $D(\theta_1, \theta_2)$ as

$$D(\theta_1, \theta_2) = \|\beta_1 - \beta_2\| + |\rho_1 - \rho_2| + \int_0^\tau |g_1(s) - g_2(s)| ds. \tag{B.1}$$

Lemma 1. *Assume $\mathscr{F}$ is the set of all monotone polynomial splines with order d and is a q-dimensional linear space. Then for any $\eta > 0$ and $\epsilon < \eta$,*

$$\log N_{[\,]}(\epsilon, \mathscr{F}, D) \lesssim q \log(\frac{\eta}{\epsilon}),$$

where $N_{[\,]}(\epsilon, \mathscr{F}, D)$ is the bracketing number and covering number with respect to $D(\cdot,\cdot)$ of a function class $\mathscr{F}$.

Lemma 2. *Suppose f is a monotone nonincreasing function with bounded r-th derivative. Then there exists a monotone nonincreasing spline function f_n with order $d \geq r+1$ and knot sequence $0 = \xi_1 = \cdots = \xi_d < \xi_{d+1} < \cdots < \xi_{k_n} < \xi_{k_n+1} = \cdots = \xi_{k_n+d} = \tau$, such that*

$$\|f - f_n\| = O(k_n^{-r}).$$

Proof of Lemma 1 and Lemma 2 follow the similar steps of [Lu07] and [Lu09].

Proof of Theorem 1. Then, to prove the consistency of the estimated parameters is equivalent to show $D(\hat{\theta}, \theta_0) \xrightarrow{P} 0$. By Theorem 2.7.5 in [VW96], we have

$$\log N_{[\,]}(\epsilon, G, D_1) \lesssim \frac{1}{\epsilon}, \tag{B.2}$$

where $G = \{g(t) : g(t)$ is a nonincreasing and bounded function.$\}$, and $D_1(g_1, g_2) = \int_0^\tau |g_1(s)-$

$g_2(s)|ds$, for any $g_1, g_2 \in G$. Define two function classes

$$\mathscr{F}_1 = \left\{\ell_n(\beta,\rho,g); (\beta,\rho) \in \mathscr{B}, \text{ for any fixed } g \in G\right\},$$

$$\mathscr{F}_2 = \left\{\ell_n(\beta,\rho,g); g \in G, \text{ for any fixed } (\beta,\rho) \in \mathscr{B}\right\}.$$

Then, following similar steps in [Bai15], we can show that both $\mathscr{F}_1$ and $\mathscr{F}_2$ have integrable envelope functions. In addition, we have $N_{[\,]}(\epsilon, \mathscr{F}_1, \|\cdot\|) \lesssim (\frac{1}{\epsilon})^p$ and $\log N_{[\,]}(\epsilon, \mathscr{F}_2, D_1) \lesssim \frac{1}{\epsilon}$, because $\ell_n(\beta,\rho,g)$ is Lipschitz with respect to (β,ρ) for any fixed g.

Therefore, we have $\log N_{[\,]}(\epsilon, \mathscr{F}^*, D) \lesssim \frac{1}{\epsilon}$ for the function class $\mathscr{F}^* = \{\ell_n(\theta) : \theta \in \Theta\}$. Furthermore, by the Glivenko-Cantelli theorem we have

$$\sup_{\theta\in\Theta} |\ell_n(\theta) - E\ell_n(\theta)| \xrightarrow{P} 0. \tag{B.3}$$

By Lemma 2, there exists a $g_{0n} \in \mathscr{F}_r^n$ such that $\sup_{t\in[0,\tau]} |g_0(t) - g_{0n}(t)| = O(k_n^{-r})$, where $g_0 \in \mathscr{F}_r$. Then, we have

$$D(\theta_0, \theta_{0n}) \longrightarrow 0, \tag{B.4}$$

where define $\theta_{0n} = (\beta_0^T, \rho_0, g_{0n})$. Θ_n is compact with respect to $D(\cdot,\cdot)$ and $\ell_n(\theta)$ is continuous in $\theta \in \Theta_n \subset \Theta$. Moreover, θ_{0n} is the unique maximizer of $E\ell_n(\theta)$ on Θ_n, because θ_0 uniquely maximizes $E\ell_n(\theta)$ on Θ. Combining with the fact $\sup_{\theta\in\Theta_n} |\ell_n(\theta) - E\ell_n(\theta)| \xrightarrow{P} 0$, we have $D(\hat{\theta}_n, \theta_{0n}) \xrightarrow{P} 0$. Furthermore, we can conclude that $D(\hat{\theta}_n, \theta_0) \xrightarrow{P} 0$.

Proof of Theorem 2. The proof consists of two steps. First, by the proof of Theorem 1 it is clear that $D(\theta_0, \theta_{0n}) = D_1(g_0, g_{0n}) = O(k_n^{-r})$.

Next, it suffices to show that $D(\hat{\theta}_n, \theta_{0n}) = O_p((\frac{k_n}{n})^{\frac{1}{2}})$. For fixed $\delta > 0$, define two function classes:

$$\mathcal{M}_1 = \left\{\ell_n(\beta,\rho,g_{0n}) - \ell_n(\beta_0,\rho_0,g_{0n}); \|\beta-\beta_0\| + |\rho-\rho_0| \le \delta, \beta \in \mathscr{B}\right\},$$
$$\mathcal{M}_2 = \left\{\ell_n(\beta_0,\rho_0,g) - \ell_n(\beta_0,\rho_0,g_{0n}); \int_0^\tau |g(t)-g_{0n}|dt \le \delta, g \in \mathscr{F}_r^n\right\}.$$

Because $\ell_n(\beta,\rho,g)$ is Lipschitz with respect to (β^T,ρ), we have $N_{[\,]}(\epsilon,\mathcal{M}_1,\|\cdot\|) \lesssim (\frac{\delta}{\epsilon})^p$ for any fixed g_{0n}. Similarly, with any fixed $(\beta^T,\rho) \in \mathscr{B}$ we have $N_{[\,]}(\epsilon,\mathcal{M}_2,D_1) \lesssim (\frac{\delta}{\epsilon})^{k_n}$ for $\mathcal{M}_2$. Then, for the function class $\mathcal{M}_{n,\delta} = \{\ell_n(\theta) - \ell_n(\theta_{0n}); D(\theta,\theta_{0n}) \le \delta, \theta \in \Theta_n\}$, we obtain

$$N_{[\,]}(\epsilon,\mathcal{M}_{n,\delta},D) \le N_{[\,]}(\frac{\epsilon}{2},\mathcal{M}_1,\|\cdot\|)N_{[\,]}(\frac{\epsilon}{2},\mathcal{M}_2,D_1).$$

Furthermore, the entropy of the funciton class $\mathcal{M}_{n,\delta}$ satisfies $\log N_{[\,]}(\epsilon,\mathcal{M}_{n,\delta},D) \lesssim k_n \log(\frac{\delta}{\epsilon})$. For the bracketing integral $J_{[\,]}(\delta,\mathcal{M}_{n,\delta},D)$ defined in [VW96], it follows that $J_{[\,]}(\delta,\mathcal{M}_{n,\delta},D) \lesssim k_n^{1/2}\delta$. By Lemma 3.4.2 in [VW96], we obtain that

$$\begin{aligned}
&E\left\{\sup_{\frac{\delta}{2}<D(\theta,\theta_{0n})\le\delta} |(\ell_n(\theta)-\ell_n(\theta_{0n})) - E(\ell_n(\theta)-\ell_n(\theta_{0n}))|\right\} \\
&\le \frac{1}{\sqrt{n}} J_{[\,]}(\delta,\mathcal{M}_{n,\delta},D)(1 + \frac{J_{[\,]}(\delta,\mathcal{M}_{n,\delta},D)}{\delta^2\sqrt{n}}A_3) \\
&\lesssim \frac{1}{\sqrt{n}} k_n^{\frac{1}{2}}\delta(1 + k_n^{\frac{1}{2}}\delta/\delta^2\sqrt{n}A_3) = O((\frac{k_n}{n})^{\frac{1}{2}}\delta).
\end{aligned}$$

In addition, by Taylor's expansion it follows that $\sup_{\delta/2<D(\theta,\theta_{0n})\le\delta,\theta\in\Theta_n} E(\ell_n(\theta)) - E(\ell_n(\theta_{0n})) \lesssim -\delta^2$. Finally, together with the facts $\phi_n(\delta) = \delta k_n^{1/2}$, $\delta_n \equiv 0$ and $r_n = (n/k_n)^{1/2}$ in [VW96], we obtain $D(\hat{\theta}_n,\theta_{0n}) = O((\frac{k_n}{n})^{1/2})$.

Combining with the two steps above, we can conclude that $D(\hat{\theta}_n,\theta_0) = O_p((\frac{k_n}{n})^{1/2} + k_n^{-r})$.

Proof of Theorem 3. Let $\varphi = (\beta, \rho)$. The score functions for $\theta = (\beta, \rho, g)$ are respectively,

$$
\begin{aligned}
\ell_{n,\beta} &= \frac{1}{nM}\sum_{i=1}^{n}\sum_{k=1}^{M}\delta_{k,i}\frac{X_i}{\beta^T X_i + \sum_{\ell=1}^{k-1}\left(\rho W_{i_{(\ell)},i} + \delta_{\ell,i}\right) g(T_{(k)} - T_{(\ell)})} - \frac{\tau}{nM}\sum_{i=1}^{n} X_i, \\
\ell_{n,\rho} &= \frac{1}{nM}\sum_{i=1}^{n}\sum_{k=1}^{M}\delta_{k,i}\frac{\sum_{\ell=1}^{k-1} W_{i_{(\ell)},i}\, g(T_{(k)} - T_{(\ell)})}{\beta^T X_i + \sum_{\ell=1}^{k-1}\left(\rho W_{i_{(\ell)},i} + \delta_{\ell,i}\right) g(T_{(k)} - T_{(\ell)})} \\
&\quad - \frac{1}{nM}\sum_{i=1}^{n}\sum_{\ell=1}^{n} W_{i,\ell}\sum_{k=1}^{M}\delta_{k,\ell}\int_0^{\tau - T_{(k)}} g(s)\,ds, \\
\ell_{n,g}[h] &= \frac{1}{nM}\sum_{i=1}^{n}\sum_{k=1}^{M}\delta_{k,i}\frac{\sum_{\ell=1}^{k-1}\left(\rho W_{i_{(\ell)},i} + \delta_{\ell,i}\right) h(T_{(k)} - T_{(\ell)})}{\beta^T X_i + \sum_{\ell=1}^{k-1}\left(\rho W_{i_{(\ell)},i} + \delta_{\ell,i}\right) g(T_{(k)} - T_{(\ell)})} \\
&\quad - \frac{\rho}{nM}\sum_{i=1}^{n}\sum_{\ell=1}^{n} W_{i,\ell}\sum_{k=1}^{M}\delta_{k,\ell}\int_0^{\tau - T_{(k)}} h(s)\,ds - \frac{1}{nM}\sum_{i=1}^{n}\sum_{k=1}^{M}\delta_{k,i}\int_0^{\tau - T_{(k)}} h(s)\,ds,
\end{aligned}
$$

and $\ell_{n,\varphi} = (\ell_{n,\beta}, \ell_{n,\rho})$.

An outline of the proof of the asymptotic normality is given as follows. First, following the proof of [Bai15], the least favorable direction $h^*(t)$ for φ exists. Denote $\widehat{h}_n$ as the spline approximation for the least favorable function $h^*(t) \in \mathscr{F}_r$, s.t. $\|\widehat{h}(t) - h^*(t)\| = O(\kappa_n^{-r})$. Since the estimator $(\widehat{\varphi}_n, \widehat{g}_n)$ maximizes the log likelihood along the submodel $(\widehat{\varphi}_n + \epsilon b, \widehat{g}_n(t) + \epsilon \widehat{h}_n(t))$, we have

$$
\ell_{n,\varphi}(\widehat{\varphi}_n, \widehat{g}_n) + \ell_{n,g}(\widehat{\varphi}_n, \widehat{g}_n)[\widehat{h}_n] = 0.
$$

By the maximal inequalities for martingales, we further have

$$
\sqrt{n}\left(\ell_{n,\varphi}(\varphi_0, g_0) + \ell_{n,g}(\varphi_0, g_0)[h^*]\right) + o_p(1) = -\sqrt{n}E\left(\ell_{n,\varphi}(\widehat{\varphi}_n, \hat{g}) + \ell_{n,g}(\widehat{\varphi}_n, \hat{g})[\hat{h}_n]\right).
$$

By the properties of least favorable direction and our assumptions, Taylor expansion of the right hand side of the last equation further leads to

$$\sqrt{n}E\left(\ell_{n,\varphi\varphi}(\varphi_0,g_0)+\ell_{n,\varphi g}(\varphi_0,g_0)[h^*]\right)(\hat{\varphi}_n-\varphi_0)=-\sqrt{n}\left(\ell_{n,\varphi}(\varphi_0,g_0)+\ell_g(\varphi_0,g_0)[h^*]\right)+o_p(1).$$

Following the similar arguments as [Bai15], the left-hand side of the last equation is nonsingular. Therefore, we have that

$$\sqrt{n}(\hat{\varphi}_n-\varphi_0)=-\left\{E\left(\ell_{n,\varphi\varphi}(\varphi_0,g_0)+\ell_{n,\varphi g}(\varphi_0,g_0)[h^*]\right)\right\}^{-1}\sqrt{n}\left(\ell_{n,\varphi}(\varphi_0,g_0)+\ell_{n,g}(\varphi_0,g_0)[h^*]\right)$$
$$+o_p(1).$$

Hence, $\sqrt{n}(\hat{\varphi}_n-\varphi_0)$ converges to a normal distribution.