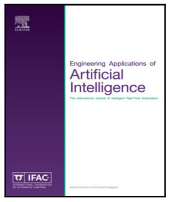
Engineering Applications of Artificial Intelligence 93 (2020) 103709

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Survival analysis of failures based on Hawkes process with Weibull base intensity

Lu-ning Zhang, Jian-wei Liu [*,1], Xin Zuo

Department of Automation, College of Information Science and Engineering, China University of Petroleum, Beijing Campus (CUP), Beijing, China

ARTICLE INFO

Keywords:
Survival analysis
Hawkes process
Conditional intensity function
Base intensity
Exponential distribution
Weibull distribution
Granger causality

ABSTRACT

In this paper, we construct a Hawkes process with time-varying base intensity to model the sequence of failure, i.e., failure events of the compressor station, and we combine survival analysis and point process model on various failure events of the compressor station based on Hawkes process. To our best knowledge, until now, nearly all relevant literature of the Hawkes point processes assumes that the base intensity of the conditional intensity function is time-invariant. This assumption is apparently too harsh to be verified. For example, in the practical application, including financial analysis, reliability analysis, survival analysis and social network analysis, the truth variation of the base intensity of the failure occurrence over time is not constant. The constant base intensity will not reflect the base intensity trend of the failure occurring over time. Thus, in order to solve this problem, in this paper, we propose a new time-varying base intensity, e.g. which is treated as obeying Weibull distribution. First, we introduce the base intensity into a Hawkes process that obeys the Weibull distribution, and then we propose an effective learning algorithm based on the maximum likelihood estimator. Experiments on the constant base intensity synthetic data, time-varying base intensity synthetic data, and real-world data show that our method can learn the triggering patterns of the Hawkes processes and the time-varying base intensity simultaneously and robustly. Experiments on real-world data also reveal the Granger causality of different types of failures and the base probability of failure varying over time. We put forward some suggestions for practical production based on the experimental results.

1. Introduction

Learning point processes, especially the Hawkes processes from irregular and asynchronous sequential data observed in continuous time, is a challenging task. Meanwhile, point processes can be applied to many fields, such as ad serving, disease prediction, and TV shows recommendation. All of these sequential data can be modeled by a point process, watched TV shows, clicked ads, and acquired illnesses all can be seen as random events. We can use the Hawkes processes to model the triggering patterns between the different types of events, for instance, which TV show audiences are watching, which ad users click and patient's disease belongs to which subject. More specifically, TV audiences will be more inclined to watch related programs after watching a program. When a user clicks on an ad, they are likely to click on another ad with the content associated with the previous ad. Patients with certain diseases are more likely to have related complications. Based on these phenomenons, we want to understand the trigger pattern between various faults in the compressor station, so we applied the point process model in the survival analysis of the compressor station system failures.

Because of the research of the causality relationship between different types of the events of the point process (Didelez, 2008), there are a lot of research focusing on Hawkes process, which is proposed by Hawkes (1971), and widely applied in many fields. Recently, an effective method is proposed to learn multi-dimensional Hawkes processes and Granger causality of different kinds of the events by learning the impact function of Hawkes processes and the causality relationship with different event types from the sequential data (Xu et al., 2016), Xu et al. (2016) propose an effective method to learn Hawkes processes and then use them to deduce the Granger causality hiding in the impact function, and apply it into IPTV data, to reflect the trigger patterns of users watching preference. From the perspective of the graphical model, learning causality relationship with different event types is equivalent to learning Granger causality graph. In Granger causality graph, the line with an arrow connecting two types of events nodes indicates that the event corresponding to the destination node depends on the historical events of the event corresponding to the source node.

However, before this, learning Granger causality for general multi-dimensional point processes with irregular and asynchronous event

* Corresponding author.
E-mail addresses: luning_zhang@foxmail.com (L.-n. Zhang), liujw@cup.edu.cn (J.-w. Liu), zuox@cup.edu.cn (X. Zuo).

[1] This work is supported by the National Key Research Program of the Ministry of Science and Technology (2016YFC0303703-03).

https://doi.org/10.1016/j.engappai.2020.103709
Received 15 November 2019; Received in revised form 14 March 2020; Accepted 9 May 2020
Available online 26 May 2020

sequences is hard to accomplish. Lian et al. (2015) propose a method that learns Granger causality by constructing features from history and selecting them. Unfortunately, this method relies heavily on feature constructing processes, may result in poor robustness. Han and Liu (2013) propose a vector auto-regressive (VAR) model to learn Granger causality from discrete time-lagged variables. Learning Granger causality from point processes is difficult because the event sequence of point processes is continuously evolving over time and does not have the fixed time-lag. So, it is hard to come up with an effective method to capture the Granger causality between different types of the events, and existing works nearly all focus on learning Granger causality from time series (Arnold et al., 2007; Eichler, 2012; Basu et al., 2015).

In this paper, we devise a new time-varying Hawkes process and make use of it to analyze the failure events sequences of the compressor station. First, we utilize the Hawkes process to analyze the causal relationship between different types of failures. We also devise the approach to get the impact of the possibility of other failures after a certain failure occurs. While analyzing the causal relationship between failures of the compressor station, we also want to identify the base intensity of each kind of failures occurred over time. However, this idea is difficult to implement based on the existing Hawkes process model. The reason is that nearly all the Hawkes processes models researcher proposed, assume the base intensity is constant. This assumption has a lot of restriction, for instance, in disease prediction, the base intensity of patients getting disease should not be constant, it must be time-varying, as the patient gets older, the risk of the getting sick should be bigger, and in TV shows recommendation, the attractive of TV show is also varying over time, should not be constant.

On the other hand, failure event sequences on the compressor station have some characteristics that other event sequences do not have, Fig. 1 depicts these characteristics of failure event sequences. During the normal operation of the compressor station, the probability of failure is relatively low, however, once a failure occurs, the likelihood of derivative failures will increase relatively. Thus, on the time axis, the distribution of failure events presents a sparse-clustering feature. We call a sequence of events with such characteristics a sparse-clustering sequence, the first failure in a clustering, we called as source failure. We assume that the source failure did not receive the impact of the previous failure history. When the device works normally, the failure occurs sparsely, and after one failure occurs, due to the trigger pattern between the source failure and derivative failures, there is a phenomenon of clustering between them.

Thus, in order to better guide production, we do not only learning the trigger patterns between different kinds of failures but also have to figures out the probability of occurrence of source failures changing with time, moreover, we also have to come up with a new framework of the time base intensity of failures in Hawkes process.

Based on the above discussion, in this paper, we introduce a kind of time-varying base intensity in Hawkes process, we treated the base intensity as random distribution, whose parameters obey Weibull distribution, Weibull distribution is widely used in many fields, such as survival analysis, reliability engineering and weather forecasting, it can express the trend of the base intensity over time. Introducing Weibull base intensity brings a new parameter into the model, we propose a new effective algorithm to learn it. After estimating the parameter, we can obtain the trend of the instantaneous occurrence rate of all kinds of failures, which can help the maintenance staff to develop a plan of equipment overhaul and maintenance to reduce the occurrence of compressor failures and improve production efficiency, which can bring greater economic benefits and safety benefits.

In Section 2, we discuss the related work about Survival analysis, Weibull distribution, Point process and Granger causality. Then, we introduce the basic concepts about point process, Weibull distribution and exponential distribution. Next, we theoretically deduce the Weibull–Hawkes model and propose a learning algorithm for model parameters. At last, we validate our model on two different synthetic datasets, demonstrating that our model is valid for both time-varying base intensity and constant base intensity, then, we validate our model on real-world data, it proves that our model is more effective on the real-world data, and obtains the base trend with time and the causal relationship of all kinds of failures.

2. Related work

Survival analysis. Survival analysis is a branch of statistics for analyzing the law of event occurrence as time goes by, such as death of infectious patients and failure in mechanical systems. One of the most important research content of survival analysis is time-to-event analysis. Time-to-event analysis is to model and predict when certain events, such as death, failure and disease attack, will occur. In addition, the time-to-event data are including the occurring events, its time-stamps, and the corresponding features of each individual. These characteristics of time-to-event data are especially similar to the data characteristics of the stochastic point process. The current research on survival analysis on time-to-event data has produced many gratifying results. Li, Kan et al. adopt survival analysis method on time-to-event data to predict the conversion from cognitive impairment to Alzheimer's Disease, which is critically important for prevention Alzheimer's Disease and targeted treatment (Li et al., 2017). Leger, S. et al. combine survival analysis and machine learning to model radiomics risk to predict patients' cancer risk on time-to-event data (Leger et al., 2017). Tierney, Jayne F., et al. propose a method to convert the time-to-event data to a less statistical and more practical guidance and use it more appropriately in meta-analysis (Tierney et al., 2007).

In engineering, survival analysis is called as reliability theory or reliability analysis, in economics, it is called duration analysis or duration modeling, and in sociology it is called event history analysis. Duchateau and Janssen (2007) systematically introduces the survival analysis and propose the new frailty model. In the field of reliability analysis, there have been many research results, such as Christodoulou (2011), Peña and Hollander (2004) and Pereira et al. (2018). Based on the extensive application of time-to-event analysis, reliability analysis, and their high similarity to the point process model, we improved our point process model with the knowledge of time-to-event analysis, to analyze the reliability of the compressor station subsystems.

Weibull distribution. Weibull distribution is first identified by Maurice Fréchet in 1927 and is first applied to describe a particle size distribution by Paul Rosin in 1933. However, Swedish mathematician Waloddi Weibull, describe it in detail in 1951, so this distribution is named as Weibull distribution. Weibull distribution is still widely applied in so many areas, such as lifetime prediction (Ali et al., 2015), reliability analysis (Bain, 2017), survival analysis (Cox, 2018), weather forecasting and the wind power industry to describe wind speed distributions (Mohammadi et al., 2016).

Point processes. The temporal point process is an event sequence randomly located on time space where the knowledge related to survival analysis can be applied. For instance, hazard function in survival analysis is similar to the conditional intensity function in the temporal point process. Thus, we will model the failure sequences of the compressor station by Hawkes processes combine with survival analysis. First, we need to introduce the most basic point process model, i.e., Poisson process. Poisson processes (Vere-Jones, 2003) are the one of simplest point process model, the difference between the other point processes and Poisson processes is that the current event of Poisson process is independent of the influence of historical events. The Poisson process has the following conditional intensity function:

$$\lambda_c(t) = h_c(t)$$

Generally speaking, Poisson process can be divided into two categories, Poisson process with time-varying $h_c(t)$, is commonly called

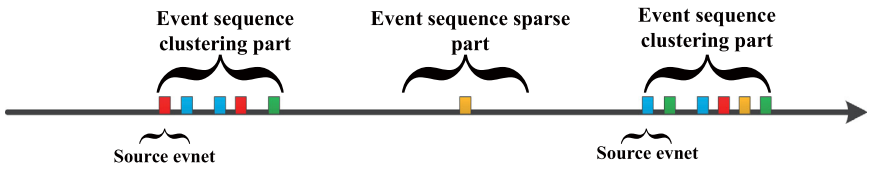


Fig. 1. The schematic diagram of sparse-clustering sequence.

the nonhomogeneous or inhomogeneous Poisson process; Poisson process with constant $h_c(t)$ is commonly called the homogeneous Poisson process.

However, pure Weibull distribution, survival analysis or Poisson process do not consider the trigger mode between various faults, so we need to combine these theories with the Hawkes process and Granger causality, then analyze the sequences of faults more effectively and comprehensively. Hawkes processes (Hawkes, 1971) are proposed to model the complicated event sequences that events occurred in the past will affect the occurrence of future events. Hawkes processes is an important kind of mutually exciting point processes which applies to many practical fields, e.g., financial analysis (Embrechts et al., 2011; Bacry et al., 2015), social network modeling (Zhou et al., 2013a), and seismic analysis (Daley and Vere-Jones, 2007).

The multidimensional Hawkes process is described by its conditional intensity function (similar as subhazard function in survival analysis) which has following form:

$$\lambda_c(t) = h_c(t) + \sum_{c'=1}^{C} \int_0^t \phi_{cc'}(s) d N_{c'}(t-s)$$
$$= h_c(t) + \sum_{c'=1}^{C} \int_0^t \phi_{cc'}(t-s) d N_{c'}(s)$$

$h_c(t)$ in the equation of conditional intensity function is the base intensity, which is independent of the history information. Generally speaking, the Hawkes processes research introduced in this section usually assumed that the base intensity is constant $h_c(t) = \mu$ and time invariant. Laub et al. (2015) give a detailed overview of the Hawkes process, introduce the mathematical model of the Hawkes process, parameter estimation methods, simulation methods and so on. In the last few years, the most of the research works which is related to Hawkes processes use predefined impact function with fixed parameters, for example, the power-law function is used in Zhao et al. (2015), and the exponential function is used in Zhou et al. (2013a), Hall and Willett (2016), Farajtabar et al. (2014), Yan et al. (2015) and Lewis and Mohler (2011), however, these methods oversimplify the impact function, thus may not learn the Hawkes processes well. There are also some models that have learned the impact function. The main idea of these approaches is to use a non-parametric model to enhance flexibility. The 1D Hawkes process with the nonparametric model is first proposed in Lewis and Mohler (2011), which is based on the ordinary differential equation, then this method is extended to multi-dimensional case in Zhou et al. (2013b) and Luo et al. (2015). Bacry et al. (2012) estimate the nonparametric Hawkes processes via solving the Wiener–Hopf equation. Reynaud-Bouret et al. (2010), Hansen et al. (2015) and Eichler et al. (2017) use the contrast function-based estimation method, minimize the estimation error of the conditional intensity function and transfer the problem to a Least-Squares(LS) problem. Du et al. (2012) and Lemonnier and Vayatis (2014) decompose impact functions into basis functions to avoid discretization.

The Gaussian process-based methods have been reported to successfully estimate the more general point processes (Lian et al., 2015; Adams et al., 2009; Lloyd et al., 2015; Samo and Roberts, 2015). However, these methods are too complicated to learn so many parameters, resulting in loss of model performance. Eichler et al. (2017) first demonstrate the relationship between the impact function of the Hawkes process and Granger causality. Xu et al. (2016) propose a nonparametric model to learn the Hawkes processes by an effective learning algorithm combining a maximum likelihood estimator (MLE) with a sparse-group-lasso (SGL) regularizer, this method can learn the impact function and Granger causality robustly, then they propose an adaptive procedure to select basis functions. These Hawkes-based methods nearly all set the base intensity to a constant, introducing unnecessary restrictions for model assumption on the survival analysis of faults. Du et al. (2016) propose the Recurrent Marked Temporal Point Process (RMTPP) to capture which type and when the events occur, they treat the intensity function of the temporal point process as a nonlinear function, and utilize recurrent neural networks to learn the representation of the impact from the event history automatically. Based on this idea, Mei and Eisner (2017) and Xiao et al. (2017) propose to model streams of discrete events in continuous time, creatively construct a neural self-modulating multivariate point process, which is called the neural Hawkes process, however, these two methods based on recurrent neural networks are black box models, which can be inconvenient to reveal the trend of the possibility of failure events themselves, and the trigger mode between faults.

And to our best knowledge, no one has ever introduced Weibull base intensity into Hawkes process, we will utilize the Weibull distribution for failure survival analysis to evaluate the system reliability of the compressor station, and our research will inject new impetus into the Hawkes process.

Granger causality. There are a lot of research efforts in Granger causality of point processes (Meek, 2014), for more general random processes, Chwialkowski and Gretton (2014) come up with a kernel independence test method. Ahmed and Xing (2009) apply lasso and its variants to reveal the inner-structure of nodes, which is a common way to learn Granger causality. Gunawardana et al. (2011) propose a model can identify temporal dependencies between different types of events. Basu et al. (2015) take an inherent grouping structure into considered when leaning Granger causality from a discrete transition process. Based on parametric cascade generative process, Daneshmand et al. (2014) put forward inference model for continuous time diffusion network to learn Granger causality. Didelez (2008) propose a theorem that reflects the relationship between the impact function and Granger causality, the most effective method is learning Granger causality from the impact function in the conditional intensity of point processes. And Eichler et al. (2017) specialize in the work for Hawkes processes and reveals the relationship between impact function and Granger causality. Xu et al. (2016) come up with a nonparametric model to learn the Hawkes processes, the Granger causality graph, and the triggering patterns of the Hawkes processes simultaneously. Since these methods do not consider the effects of time-varying base intensity, the results of causal analysis have received a negative impact, resulting in the Granger relationship obtained by these methods are not the most sparse and most accurate. In this paper, we will utilize a similar algorithm to analyze the causal relationship between the failures of the various subsystems in the compressor station in order to prevent the occurrence of secondary failures.

3. Basic concepts

3.1. Point process

Definition 1 (*Counting Process*)**.** A counting process is a stochastic process $(N(t) : t \geq T_b)$, T_b is the beginning of the observation window.

A counting process is almost surely finite, and is a right-continuous step function which the size of increments is +1. Further, denoting $H(t)$, $(t > T_b)$, $H(t)$ is the history of the arrivals up to time t. (Strictly speaking, $H(\cdot)$ is a filtration, an increasing sequence of σ-algebras.)

Definition 2 (*Temporal Point Process*)**.** A Point process is a collection of mathematical points randomly distributed in a mathematical space, such as time and real space. The point process distributed in the timeline is called as the temporal point process.

A temporal point process is a random process composed of the sequence of events and the corresponding time stamp $\{t_i\}$ these events occurred, where, $t_i \in [T_b, T_e]$, T_b is the beginning of the observation window, T_e is the end of the observation window. The temporal point process can be represented as a counting process $N = \{N(t)|t \in [T_b, T_e]\}$, where $N(t)$ is a counting process, records the number of events has happened before time t. Based on the above definition, multi-dimensional point processes with C types of the events can be represented by C counting processes $\{N_c\}_{c=1}^{C}$ on a probability space.

A point process can be described via its conditional intensity function $\{\lambda_c(t)\}_{c=1}^{C}$, the conditional intensity function is similar as the hazard function in survival analysis.

Definition 3 (*Conditional Intensity Function*)**.** Conditional intensity functions is $\{\lambda_c(t)\}_{c=1}^{C}$, where $\lambda_c(t)$ represents the expected instantaneous happening rate that the c-type event occurs instantaneously under a given history record, whose definition is shown in Eq. (1).

$$\begin{aligned}\lambda_c(t) &= \frac{\mathrm{E}(N_c(t+dt)-N_c(t)|H(t))}{dt}\\ &= \frac{P(\text{type c event occurs in}[t,t+dt)|H(t))}{dt}\\ &= \frac{P(\text{type c event occurs in}[t,t+dt)|\text{no event occurred in}[t_i,t),H(t))}{dt}\\ &= \frac{P(\text{type c event occurs in}[t,t+dt),\text{no event in}[t_i,t)|H(t))}{P(\text{no event occurred in}[t_i,t)|H(t))dt}\\ &= \frac{p(t,c)}{1-P(t,t_i)}\end{aligned} \tag{1}$$

where $p(t, c)$ is the conditional density of type c events at time t, $H(t)$ is the history affecting the type c events. t_i is the last events' timestamp before time t and $P(t, t_i)$ is the conditional cumulative function, equals the possibility if there were any events happening in $[t_i, t)$. According to Vere-Jones (2003), we can calculate $p(t, c)$ and $P(t, t_i)$ based on Eqs. (2) and (3):

$$p(t,c) = \lambda_c(t)\exp(-\sum_{c'=1}^{C}\int_{ti}^{t}\lambda_{c'}(s)ds) \tag{2}$$

$$P(t,t_i) = 1-\exp(-\textstyle\sum_{c'=1}^{C}\int_{ti}^{t}\lambda_{c'}(s)ds) \tag{3}$$

For any sequence of events $S = \{(t_i, c_i)\}_{i=1}^{I}$, we can calculate its likelihood function as Eq. (4):

$$\begin{aligned}L(S;\Theta) &= \textstyle\prod_{i=1}^{I} p(t_i,c_i)\times(1-P(T_e,t_I))\\ &= \textstyle\prod_{i=1}^{I}\lambda_{c_i}(t_i)\times\exp(-\sum_{c=1}^{C}\int_{T_b}^{T_e}\lambda_c(s)ds)\end{aligned} \tag{4}$$

Different types of point processes have different forms of conditional intensity functions, and we only discuss Hawkes process in this paper.

3.2. Weibull distribution and exponential distribution in conditional intensity function

There is a close relationship between survival analysis and point process analysis. All of these analyses are considering certain event occur at a certain time, in general the most significant difference is that in the survival analysis, the occurrence of an event will cause the observation to stop (the individual will die after the event), while the point process model generally analyzes the occurrence of the recurrent events (Zhang and Zhou, 2018).

The most widely used mathematical model is the conditional intensity function in point process which is similar as the hazard function in survival analysis, denote as $h(t)$, $h(t)$ also represents the expected instantaneous happening rate of some kind of event, shown as Eq. (5) (Duchateau and Janssen, 2007):

$$h(t) = \lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t|T\ge t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log(S(t)) \tag{5}$$

Based on Eq. (5) (Duchateau and Janssen, 2007), we can get $S(t)$, $F(t)$ and $f(t)$ (see Eqs. (6)–(8)):

$$S(t) = 1-F(t) = P(T>t) = \exp(-\int_0^t h(s)ds) \tag{6}$$

$$F(t) = \int_0^t f(s)ds = 1-\exp(-\int_0^t h(s)ds) \tag{7}$$

$$f(t) = h(t)\exp(-\int_0^t h(s)ds) \tag{8}$$

where $S(t)$ is the survival function, indicates the probability that the individual will survive to time t, which is widely applied in survival analysis and its various subdomains. $F(t)$ is the cumulative distribution function, and $f(t)$ is the probability density function.

In the point process model, if we consider an one-dimensional point process model, and no events have occurred in the past ($t_i = 0$), so based on Eqs. (2) and (3) we get that $p(t)$ and $P(t, 0)$ at this point are (see Eqs. (9) and (10)):

$$P(t,0) = \int_0^t p(s)ds = 1-\exp(-\int_0^t\lambda(s)ds) \tag{9}$$

$$p(t) = \lambda(t)\exp(-\int_0^t\lambda(s)ds) \tag{10}$$

We can see that Eqs. (7) and (9), and Eqs. (8) and (10) are identical, which proves the correlation between point process and survival analysis. In some special cases, the point process model is basically identical as the survival analysis model.

In general, the probability distribution commonly used in survival analysis is the Weibull distribution, for example, if we set hazard function $h(t) = \mu\rho t^{\rho-1}$, then the corresponding probability density function $f(t)$ is as follows:

$$f(t) = h(t)\exp(-\int_0^t h(s)ds) = \mu\rho t^{\rho-1}\exp(-\mu t^{\rho})$$

Meanwhile, in the research field of Hawkes process, most of the researchers set these base intensity as a constant, i.e., set $h_c(t) = \mu$, if we only consider influence of an arbitrary event type by its base intensity, let other type events' base intensity and all the impact function be set as zero, where we denote it as $p(t, c)'$, which is equivalent to a one-dimensional Poisson process. The conditional intensity function of the Poisson process has the following form:

$$\lambda(t) = h(t)$$

Then according to Eq. (2), when we set $h(t) = \mu$, we can calculate $p(t, c)'$ based on Eq. (11):

$$p(t,c)' = h_c(t)\exp(-\int_0^t h_c(s)ds) = \mu_c\exp(-\mu_c t) \tag{11}$$

At this point, we can see that, if the base intensity is constant, the c type event influenced by its base intensity is exponentially distributed. However, the exponential distribution has a critical shortage, i.e., memory-less property, shown in Eq. (12):

$$P(T<s+t|T>t) = \frac{\int_t^{s+t}\mu e^{-\mu t}}{\int_t^{+\infty}\mu e^{-\mu t}} = \frac{e^{-\mu t}-e^{-\mu(t+s)}}{e^{-\mu t}} = 1-e^{-\mu s} = P(T<s) \tag{12}$$

Characteristics reflected in Eq. (12) is corresponding to that the base intensity is time invariant. However, this assumption contradicts reality. The base possibility that probability of an arbitrary event

occurrence should be evolving over time, rather than being a constant. For example, high-quality TV shows will be more popular with evolving over time, and as the kids grow older, young children will become stronger and not get sick easily. If we assume the base intensity is constant, the time-invariant model cannot grasp the information mentioned above.

So, to overcome this drawback, and learn the event occurring tendency with varying time, we introduce Weibull distribution into Hawkes processes, which is widely applied in survival analysis. We set $h(t) = \mu\rho t^{\rho-1}$, and then we can get corresponding one-dimensional Poisson process $p(t,c)'$ in Eq. (13):

$$p(t,c)' = h_c(t)\exp(-\int_0^t h_c(s)ds) = \mu_c\rho_c t^{\rho_c-1}\exp(-\mu_c t^{\rho_c}) \tag{13}$$

where μ_c is the scale parameter, which defines the scale of the $h(t)$, ρ_c is the shape parameter that determines whether the function is increasing or decreasing over time. Therefore, under this assumption, the probability of event occurrence influenced by the base intensity obeys the Weibull distribution.

Let us consider a special case, when we set ρ_c=1, the base intensity is exponentially distributed. Thus, the exponential distribution is a special case of the Weibull distribution when ρ_c=1, our proposed the Hawkes processes introduced Weibull distribution is more general and take the previous ones as special cases, which is suitable for a broad range of application scenarios.

By observing the parameters obtained after learning the data of the actual event sequence, we can get an estimate of the base instantaneous happening rate at which a certain failure occurs at each time instant. Especially the estimation of shape parameters ρ_c: if $\rho_c < 1$, then we can know that the base instantaneous happening rate of c kind of failure is descending over time, on the contrary, if $\rho_c > 1$, the base instantaneous happening rate of c kind of failure is increasing over time. Then we can obtain the trend of the instantaneous occurrence rate of each type of failure, and the compressors' maintainer and manager can take targeted adjustment measures to reduce the occurrence of failures.

Competing risks and multi-dimensional point process model. In addition, we observe that multi-types events occur concurrently in competing risks analysis (Pintilie, 2006; Crowder, 2006), and there are certain competitive relationships between the events, for instance, patients may pass away from different diseases, there is competition between the diseases, a patient who died of a heart attack cannot be observed his death from cancer. In competing survival analysis, the subhazard function of each type of event is $h_c(t)$, where $c \in \{1,\dots,C\}$, there are C kinds of events. The definition of $h_c(t)$ is shown in Eq. (14) (Pintilie, 2006; Crowder, 2006):

$$h_c(t) = \lim_{\Delta t\to 0}\frac{P\{t \le T < t+\Delta t, \text{type } c \text{ events occurred}|T \ge t\}}{\Delta t} = \frac{f_c(t)}{S(t)} \tag{14}$$

where $f_c(t)$ is c-type event probability subdensity function, can be obtained by Eq. (15):

$$f_c(t) = h_c(t)\exp(-\sum_{c'=1}^{C}\int_0^t h_{c'}(s)ds) \tag{15}$$

Suppose that $S_c(t)$ is subsurvival function of type-c event, represents the probability that an individual will not be destroyed by the type-c event at the time t. $S_c(t)$ and cumulative subdistribution function $F_c(t)$ of type c event can be defined by Eq. (16):

$$S_c(t) = 1 - F_c(t) = \exp(-\int_0^t h_c(s)ds) \tag{16}$$

Overall survival function $S(t)$ is the probability that the individual successfully survives to time t, overall cumulative distribution function $F(t)$ and $S(t)$ is denoted as Eq. (17):

$$S(t) = \prod_{c'=1}^{C} S_{c'}(t) = \exp(\sum_{c'=1}^{C} -\int_0^t h_c(s)ds) = 1 - F(t) \tag{17}$$

In order to identify the similarity between multi-dimensional Hawkes process and competing risk model, we also assume that no history event happened in the past ($t_i = 0$) in a multi-dimensional Hawkes process, under this assumption, we can see that $f_c(t)$ in Eq. (15) and $p(t,c)$ in Eq. (2) are equivalence, $F(t)$ in Eq. (17) and $P(t,0)$ in Eq. (3) are equivalence. This demonstrates that to some extent the multidimensional Hawkes process is equal to the competitive risk model.

In survival analysis and competing risk model, we compare Eqs. (8) and (15), and Eqs. (6) and (17), and we can see that the probability of each event occurrence decreases because of the competition with other events, and the survival function value is also reduced due to the possibility of other possible negative events. Based on Eqs. (2) and (3), we can see a consistent pattern. The point process changes from one dimension to multiple dimensions, and the conditional probability density $p(t,c)$ and the conditional probability $1 - P(t,t_i)$ that there is no event occurred between $[t_i,t)$ also will decrease. In multi-dimensional point process, there is also competition between events.

In summary, we can see that the point process model and survival analysis are very similar in nature. Under certain conditions, multi-dimensional point process model and competitive risk model are basically equivalence in mathematical form. The multidimensional Hawkes process we applied already contains the characteristics of the risk competing model. There are competitive relationships between the different kinds of events in the multi-dimensional point process.

3.3. Hawkes processes and granger causality

A C-dimensional point processes generated a set of the event sequence, $1,\dots,C \in C$ are event types. Supposed that there is a subset of event types $\mathcal{V} \subset C$, for the type c of the event, intensity function $\lambda_c(t)$ only influenced by the history of c type of event in $\mathcal{V}$,we denote the history as $H_{\mathcal{V}}(t)$, the remaining history of event types is denoted by $H_{C\setminus\mathcal{V}}(t)$. In the view of Granger causality, $\mathcal{V}$ amount to the local independence over the dimension of the point process. The occurrence of history events in $H_{\mathcal{V}}(t)$ will affect the probability of occurrence of the future c type of events, and the $H_{C\setminus\mathcal{V}}(t)$ will not. For a subset $\mathcal{V} \subset C$, set $N_{\mathcal{V}}\{N_c(t)|c \in \mathcal{V}\}$, the filtration $H_{\mathcal{V}}(t)$, i.e., the smallest σ-algebra, which is generated by the point process, is defined as $\sigma\{N_c(s)|s \le t, c \in \mathcal{V}\}$. In particular, $H_c(t)$ is the internal filtration of the counting process $N_c(t)$, $H_{-c}(t)$ is the filtration for the subset $C\setminus\{c\}$.

Lemma 1 (*Didelez, 2008*). *The counting process $N_c(t)$ is locally independent of $N_{c'}(t)$ given $N_{C\setminus\{c,c'\}}(t)$ if the intensity function $H_c(t)$ is measurable with regard to $H_{-c}(t)$ for all $t \in [T_b^n, T_e^n]$. Otherwise, $N_c(t)$ is locally dependent.*

Based on Lemma 1, we can apply it to Hawkes processes and then establish the relationship between the impact function and Granger causality. Lemma 1 is equivalent to type c' of the event does not have Granger causality with type c of the event with regard to $H(t)$.

Generally speaking, the conditional intensity functions of multi-dimensional Hawkes processes have following form, shown in Eq. (18):

$$\begin{aligned}\lambda_c(t) &= h_c(t) + \sum_{c'=1}^{C}\int_0^t \phi_{cc'}(s)dN_{c'}(t-s)\\ &= h_c(t) + \sum_{c'=1}^{C}\int_0^t \phi_{cc'}(t-s)dN_{c'}(s)\end{aligned} \tag{18}$$

where $h_c(t)$ is the base intensity, which is independent of the history influence, and in general, $h_c(t)$ is a constant function, $\sum_{c'=1}^{C}\int_0^t \phi_{cc'}(s)dN_{c'}(t-s)$ is the endogenous intensity, indicates the history of events' effect on the $\lambda_c(t)$. $\phi_{cc'}(t)$ is impact function which measures the influence of the historical type c' events on the type c events.

From what has been discussed above, in this paper, we assume the impact function is stationary, i.e. time-invariant, and $\phi_{cc'}(t-s) \ge$

$0, (T_b \le s < t \le T_e)$. Based on this assumption, the work in Eichler et al. (2017) reveal the connection between the impact function and Granger causality.

Lemma 2 (*Eichler et al., 2017*)**.** *Assume that there exist a Hawkes process with the conditional intensity function defined in Eq. (18). If the condition* $dN_{c'}(t-s) > 0, (T_b \le s < t \le T_e)$ *holds, then if and only if* $\phi_{cc'}(t) = 0$ *for* $t \in [0, \infty)$*, type* c *of event and type* c' *of the event do not have Granger causality relationship.*

Therefore, for multi-dimensional Hawkes process, if we want to learn its Granger causality between different types of the events, we only have to confirm whether the impact function is all zero or not. It converts the Granger causality learning problem to the learning impact function problem.

After learning the sequence of failure events, we can get the impact function between various failure events, so we get the trigger pattern between all failures. In this way, we can take targeted measures to reduce or avoid secondary failure events.

4. Proposed Hawkes processes with the time varying Weibull base intensity and learning algorithm

In this section, we first introduce the Weibull base intensity to the Hawkes processes, and then, we change the form of the conditional intensity function, at last we proposed an efficient learning algorithm based on MLE method and EM algorithm. Compared with existing learning algorithms, our algorithm is more effective in identifying both the base intensity and Granger causality.

4.1. Introducing the Weibull base intensity to the Hawkes processes

Following the above analysis and discussion, if we set $h(t) = \mu\rho t^{\rho-1}$, then we can get the time varying conditional intensity function of Hawkes process, like Eq. (19):

$$\begin{aligned}\lambda_c(t) &= \mu_c \rho_c t^{\rho_c - 1} + \sum_{c'=1}^{C} \int_{T_b}^{t} \phi_{cc'}(t-s) dN_{c'}(s) \\ &= \mu_c \rho_c t^{\rho_c - 1} + \sum_{c'=1}^{C} \int_{T_b}^{t} \phi_{cc'}(s) dN_{c'}(t-s)\end{aligned} \tag{19}$$

Compared with the normal form of Hawkes process, the base intensity is replaced by $\mu_c \rho_c t^{\rho_c - 1}$ with μ_c, then in Hawkes process with time varying conditional intensity function we introduce a new parameter ρ_c, we have to propose a new effective method to learn all the parameters, including ρ_c.

Here, what we need to emphasize is that the impact function of a certain type of event on itself and its base intensity of the Weibull form are two different concepts. Impact function $\phi_{cc'}(t)$ indicates the effect of history events of type c on the occurrence probability of the current type c of event occurrence, and the base intensity of the Weibull form indicates the possibility of variation with time for occurrence of the current type c of the event. For example, if an elderly person gets sick, then in the future he will become more susceptible to illness because of decreased immunity, and although an elderly person may never get sick, the probability of his getting sick will also increase as he grows older. Note that when we set $\phi_{cc'}(t)$ in Eq. (19), this definition in Eq. (19) is equivalent to the normal Hawkes process proposed in the literature, thus the previous proposed point processes model based on Hawkes process is a special case of our proposed ones. Our Hawkes process with the time varying base intensity has a better ability to represent the time-varying characteristics of conditional intensity function.

4.2. Formulating learning task

First, we parameterize $\phi_{cc'}(t) = \sum_{m=1}^{M} a_{cc'm} g_m(t)$ as described by Lewis, one can refer (Lewis and Mohler, 2011). Where $g_m(t)$ is the mth kernel function, which is Gaussian function, and $\boldsymbol{a}_{cc'} = (a_{cc'1}, \dots, a_{cc'm})$ is the corresponding parameters of $g_m(t)$. Suppose that we have a set of the event sequence, $S = \{s_n\}_{n=1}^{N}$, $s_n = \{(t_i^n, c_i^n)\}_{i=1}^{I_n}$. $t_i^n \in \{T_b^n, T_e^n\}$ is the time instant when the ith event happened in interval $[T_b^n, T_e^n]$, and $c_i^n \in \{1, \dots, C\}$ is the corresponding type of the event. According to the definition of the conditional intensity functions Eqs. (1) and (4), the likelihood can be expressed as follows:

$$\begin{aligned}\mathcal{L}(S;\Theta) &= \prod_{n=1}^{N} \left\{ \prod_{i=1}^{I_n} p(t_i, c_i) \times (1 - P(T_e^n)) \right\} \\ &= \prod_{n=1}^{N} \left\{ \prod_{i=1}^{I_n} \lambda_{c_i^n}(t_i^n) \times \exp\left(- \sum_{c=1}^{C} \int_{T_b^n}^{T_e^n} \lambda_c(s) ds \right) \right\}\end{aligned}$$

Set $\Theta = \{\boldsymbol{A} = [a_{cc'm}] \in \mathbf{R}^{C \times C \times M}, \boldsymbol{\mu} = [\mu_c] \in \mathbf{R}^{C}, \boldsymbol{\rho} = [\rho_c] \in \mathbf{R}^{C}\}$, the log-likelihood function is shown in Eq. (20):

$$\begin{aligned}\log \mathcal{L}(S;\Theta) &= \sum_{n=1}^{N} \left\{ \sum_{i=1}^{I_n} \log \lambda_{c_i^n}(t_i^n) - \sum_{c=1}^{C} \int_{T_b^n}^{T_e^n} \lambda_c(s) ds \right\} \\ &= \sum_{n=1}^{N} \left\{ \sum_{i=1}^{I_n} \log \left(\rho_c \mu_c t_i^{n\rho_c - 1} + \sum_{j=1}^{i-1} \sum_{m=1}^{M} a_{c_i^n c_j^n m} g_m(\tau_{ij}) \right) \right. \\ &\quad \left. - \left(\sum_{c=1}^{C} \left(T_e^{n\rho_c} - T_b^{n\rho_c} \right) \mu_c + \sum_{i=1}^{I_n} \sum_{m=1}^{M} a_{cc_i^n m} G_m(T_e^n - t_i^n) \right) \right\}\end{aligned} \tag{20}$$

where $\tau_{ij} = t_i - t_j$, and $G_m(t) = \int_0^t g_m(s) ds$.

In order to enhance the robustness and accuracy of our proposed model, we consider incorporating the following two regularizers:

Temporal sparsity. The necessary condition for the stability of the Hawkes process is that $\int_0^\infty \phi_{cc'}(t) < \infty$, which mean that impact function should satisfy the asymptotic stable constraint $\phi_{cc'}(t) \to 0$ as $t \to \infty$ (Xu et al., 2016). Therefore, we add L_1-norm sparsity regularizer to the parameters of $g_m(t)$, which is denoted as $\|\boldsymbol{A}\|_1 = \sum_{c,c',m} |a_{cc'm}|$.

Local independence. Based on Lemma 2, if $\phi_{cc'}(t) = 0$ for all $t \in [0, \infty)$, then the type c' event has no effect on the type c event. Thus we incorporate L_{21}-norm regularizer (Xu et al., 2016; Song et al., 2013; Xu et al., 2010; Simon et al., 2013) to constraints the structure of coefficients of $g_m(t)$, $L_{2,1}$-norm is denoted as $\|\boldsymbol{A}\|_{1,2} = \sum_{c,c',m} \|\boldsymbol{a}_{cc'}\|_2$, where $\boldsymbol{a}_{cc'} = [a_{cc'1}, \dots, a_{cc'M}]$ is the corresponding vector of $\phi_{cc'}(t)$. The purpose of incorporating $L_{2,1}$-norm regularizer is to ensure the group sparsity of the coefficient tensor.

Thus, we can get the objective function of the Weibull–Hawkes process is Eq. (21):

$$\arg\min_{\Theta \ge 0} - \log \mathcal{L}(S;\Theta) + \alpha_S \|\boldsymbol{A}\|_1 + \alpha_G \|\boldsymbol{A}\|_{1,2} \tag{21}$$

4.3. An EM-based algorithm

Similar to Xu et al. (2016), Lewis and Mohler (2011), Zhou et al. (2013b) and Daley and Vere-Jones (2007), we propose an EM-based learning algorithm to solve the optimization problem (10) iteratively.

Update $\boldsymbol{A}$ *and* $\boldsymbol{\mu}$*.* Given the parameters of the current step, we can reconstruct the log-likelihood function by Jensen's inequality, shown in Eq. (22):

$$\begin{aligned}&\log(\mu_{c_i^n} \rho_{c_i^n} t_i^{\rho_{c_i^n} - 1} + \sum_{j=1}^{i-1} \sum_{m=1}^{M} a_{c_i^n c_j^n m} g_m(\tau_{ij}^n)) \ge \\ &p_{ii} \log\left(\frac{\mu_{c_i^n} \rho_{c_i^n} t_i^{\rho_{c_i^n} - 1}}{p_{ii}}\right) + \sum_{j=1}^{i-1} \sum_{i=1}^{M} p_{ijm} \log\left(\frac{a_{c_i^n c_j^n m} g_m(\tau_{ij}^n)}{p_{ijm}}\right)\end{aligned} \tag{22}$$

where:

$$p_{ii} = \frac{\mu_{c_i^n}^{(k)} \rho_{c_i^n}^{(k)} t_i^{\rho_{c_i^n}^{(k)}-1}}{\lambda_{c_i^n}^{(k)}(t_i^n)} \text{ and } p_{ijm} = \frac{a_{c_i^n c_j^n m}^{(k)} g_m(\tau_{ij}^n)}{\lambda_{c_i^n}^{(k)}(t_i^n)}$$

$\lambda_{c_i^n}^{(k)}(t_i^n)$ is the conditional intensity function at t_i^n with kth step parameters. Then, we can get a tight bound of log-likelihood function as Eq. (23):

$$Q(\Theta;\Theta^{(k)}) = \sum_{n=1}^{N}\left\{-\left(\sum_{c=1}^{C}\left(T_e^{n\rho_c} - T_b^{n\rho_c}\right)\mu_c + \sum_{i=1}^{I_n}\sum_{m=1}^{M} a_{cc_i^n m} G_m(T_e^n - t_i^n)\right) + \sum_{i=1}^{I_n}\left(p_{ii}\log\frac{\mu_{c_i^n}\rho_c t_i^{\rho_c-1}}{p_{ii}} + \sum_{j=1}^{i-1}\sum_{i=1}^{M} p_{ijm}\log\frac{a_{c_i^n c_j^n m} g_m(\tau_{ij}^n)}{p_{ijm}}\right)\right\} \tag{23}$$

If and only if $\Theta=\Theta^{(k)}$, we have $Q(\Theta;\Theta^{(k)}) = \log\mathcal{L}(S;\Theta)$. So, we can obtain the surrogate objective function:

$$F = -Q(\Theta;\Theta^{(k)}) + \alpha_S\|\boldsymbol{A}\|_1 + \alpha_G\|\boldsymbol{A}\|_{1,2}$$

Then, we can get the partial derivative of F about $\boldsymbol{\mu}$ and $\boldsymbol{A}$ (see Eqs. (24) and (25)):

$$\frac{\partial F}{\partial \mu_c} = \sum_{n=1}^{N}\left\{\left(T_e^{n\rho_c} - T_b^{n\rho_c}\right) - \sum_{i=1}^{I_n}\sum_{c_i=c} p_{ii}\frac{1}{\mu_c}\right\} \tag{24}$$

$$\frac{\partial F}{\partial a_{cc'm}} = \sum_{n=1}^{N}\left\{\sum_{i=1}^{I_n}\left(G_m(T_e^n - t_i^n) - \sum_{c_i^n=c}\sum_{c_j^n=c'}\frac{p_{ijm}}{a_{cc'm}}\right)\right\} + \alpha_S + \alpha_G\frac{a_{cc'm}}{\left\|a_{cc'}^{(k)}\right\|_2} \tag{25}$$

Set $\frac{\partial F}{\partial \mu_c} = 0$, and $\frac{\partial F}{\partial a_{cc'm}} = 0$, we can get closed-form solutions of μ_c and $a_{cc'm}$, shown in Eqs. (26) and (27):

$$\mu_c^{(k+1)} = \frac{\sum_{n=1}^{N}\sum_{c_i^n=c} p_{ii}}{\sum_{n=1}^{N}(T_e^{n\rho_c} - T_b^{n\rho_c})} \tag{26}$$

$$a_{cc'm}{}^{(k+1)} = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{27}$$

where:

$$A = \frac{\alpha_G}{\left\|a_{cc'}^{(k)}\right\|}$$

$$B = \sum_{n=1}^{N}\sum_{c_i^n=c} G_m(T_e^n - t_i^n) + \alpha_S$$

$$C = -\sum_{n=1}^{N}\sum_{c_i^n=c}\sum_{c_j^n=c'} p_{ijm}$$

And if $\alpha_S = 0$ and $\alpha_G = 0$, we can get $a_{cc'm}$ by Eq. (28):

$$a_{cc'm}{}^{(k+1)} = \frac{\sum_{n=1}^{N}\sum_{c_i^n=c}\sum_{c_j^n=c'} p_{ijm}}{\sum_{n=1}^{N}\sum_{c_i^n=c'} G_m(T_e^n - t_i^n)} \tag{28}$$

Update ρ. Simultaneously, we can get the partial derivative of F about $\boldsymbol{\rho}$, shown in Eq. (29):

$$\frac{\partial F}{\partial \rho_c} = -\sum_{n=1}^{N}\left\{-\mu_c(\ln T_e^n \cdot T_e^{n\rho_c} - \ln T_b^n \cdot T_b^{n\rho_c}) + \sum_{i=1}^{I_n}\sum_{c_i^n=c} p_{ii}(\ln t_i + \frac{1}{\rho_c})\right\} \tag{29}$$

If we set $\frac{\partial F}{\partial \rho_c} = 0$, it is relatively hard to get the closed form solution of $\boldsymbol{\rho}$. Thus, we tend to update the $\boldsymbol{\rho}$ by the gradient descent method like Eq. (30):

$$\rho_c^{k+1} = \rho_c^k - \alpha_\rho\frac{\partial F}{\partial \rho_c} \tag{30}$$

where α_ρ is the learning rate of $\boldsymbol{\rho}$, and in order to determine when the iterative processes terminate, we adopt a strategy called early stopping (Caruana et al., 2001), i.e., if the objective function F with current parameters will increase, then the iterative processes will be terminated. Algorithm 1 shows how to estimate the parameters of the Weibull–Hawkes process.

Algorithm 1 EM-based Algorithm for Weibull–Hawkes process (WBMLE-SGL))

Input: Event sequences $\{S_n\}_{n=1}^N$, trade-off parameters α_S,α_G,α_ρ and k.
Output: Parameters of model $\boldsymbol{\mu}$,$\boldsymbol{\rho}$ and $\boldsymbol{A}$.

```
Initialize A = [a_{cc'm}] and μ = [μ_c] randomly, set ρ = [ρ_c], ρ_c = 1 .
repeat
    repeat
        Update μ and A via Eq. (26) and Eqs. (27) or (28) respectively.
    until Convergence
    Update ρ via Eq. (30) k times.
until The likelihood function is not changing or satisfies the early stopping condition.
```

5. Experimental results

To verify the robustness and effectiveness of our proposed Hawkes processes with the time varying base intensity more accurately, we compare our model with the state-of-the-art model, MLE-SGLP (Xu et al., 2016), which can reveal the Granger causality robustly and accurately on synthetic datasets and real-world datasets with constant base intensity. First, we generate sets of event sequence both with time-varying base intensity and constant base intensity, respectively. Then we test our proposed Weibull–Hawkes process model and compare it to the MLE-SGLP and MLE model on these two kinds of synthetic datasets. In order to Figure out the influence of regularizer, we consider simultaneously two learning scenarios: the pure Weibull–Hawkes process without any regularizer (WB), the Weibull–Hawkes process with spare regularizer and group-lasso regularizer (WB-SGL). GL indicates group regularizers, S indicates spare regularizers, and P indicates pairwise similar regularizers in Xu et al. (2016). To make the evaluation of different models more intuitive, we apply the following measure criteria:

(1) The log-likelihood of testing data, denote as Loglike;

(2) The relative error of $\boldsymbol{\mu}$:

$$e_{\boldsymbol{\mu}} = \frac{\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2}{\|\boldsymbol{\mu}\|_2}$$

(3) The relative error of $\boldsymbol{\rho}$:

$$e_{\boldsymbol{\rho}} = \frac{\|\tilde{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2}{\|\boldsymbol{\rho}\|_2}$$

(4) In order to testify the error of the base intensity function, we denote a relative error of $h(t)$ as:

$$e_{h(t)} = \frac{1}{C}\sum_{c\in C}\frac{\int_{T_b}^{T_e}\left|\tilde{h}_c(t) - h_c(t)\right|dt}{\int_{T_b}^{T_e} h_c(t)dt}$$

(5) Because there are some impact functions which are all-zero, the relative error of impact functions cannot be calculated, instead, we choose the absolute error of impact functions:

$$\boldsymbol{\Phi}(t) = [\phi_{cc'}(t)]$$

$$e_{\boldsymbol{\Phi}} = \sum_{c,c'}\int_0^T\left|\tilde{\phi}_{cc'}(t) - \phi_{cc'}(t)\right|dt$$

(6) Accuracy of Granger causality analysis via distinguishing the impact function which is all-zero or not.

After testing a variety of patterns with the synthetic data experiment, we test our proposed models on real-world data: the failure event sequence of 87 compressor station during twelve years. Learning our model on these data reveals a lot of useful information in strategizing the actual operation and maintenance of the compressor station. We will introduce it in Section 5.4.

5.1. Data generating protocol

To assess the usefulness of our model, we generate the time-varying base intensity data (Ogata, 1981). Our data has the two kinds of impact functions: sine-like impact functions and square-like impact functions. Each of them is a 5-dimensional Hawkes process and contains 500 asynchronous event sequences with the time length 50. The $\boldsymbol{\mu}$ of each event type is uniformly sampled from $[0, \frac{1}{5}]$, the $\boldsymbol{\rho}$ of each event type is uniformly sampled from $[0.5, 1.5]$. The sine-like impact functions are generated as:

$$\phi cc'(t) = \begin{cases} A_{cc'}(1-\cos(\omega_{cc'}t+\varphi_{cc'})),\ t \in [0, \frac{2\pi-\varphi_{cc'}}{\omega_{cc'}}] \\ 0, \text{otherwise} \end{cases} \quad c, c' \in 1, \ldots, 5$$

where If $c' \in \{1,2,3\}$ and $c \in \{1,2,3\}$, then $A_{cc'} = 0.05, \omega_{cc'} = 0.6\pi, \varphi_{cc'} = 0$; If $c' \in \{4,5\}$ and $c \in \{4,5\}$, then $A_{cc'} = 0.05, \omega_{cc'} = 0.4\pi, \varphi_{cc'} = \pi$; If $c' \in \{5\}$ and $c \in \{1,2,3\}$ or $c \in \{5\}$ and $c' \in \{1,2,3\}$, then $A_{cc'} = 0$; If $c' \in \{4\}$ and $c \in \{1,2,3\}$ or $c \in \{4\}$ and $c' \in \{1,2,3\}$, then $A_{cc'} = 0.02, \omega_{cc'} = 0.2\pi, \varphi_{cc'} = \pi$. The square-like impact functions are the truncated results of above sine-like impact functions. In order to testify the MLE-SGLP model, we set three pairs of different events type: [1, 2; 2, 3; 4, 5], which pairs of events type have the similar Granger causality.

Algorithm 2 Generating Sequence Data of Hawkes Processes.

Input: scale parameter $\boldsymbol{\mu}$, shape parameter $\boldsymbol{\rho}$, impact functions $\boldsymbol{\Phi}(t) = [\phi_{cc'}(t)]$ and length of sequence T.
Output: Event sequence S_n.
Initialize t = $\varepsilon(\varepsilon > 0)$, ε is a small positive number, $S_n = \emptyset$.
repeat
Calculate $M = \max \sum_1^C \lambda_c(t')$,
where $t' \in [t, t+\max(T_{cc'})]$, $T_{cc'}$ is the max period of all the $\phi_{cc'}(t)$,
and $\lambda_c(t') = \mu_c \rho_c t'^{\rho_c - 1} + \sum_{j=1}^{i-1} \sum_{m=1}^{M} a_{c_i^n c_j^n m} g_m(t'-t)$
Generate an exponential r.v. E with mean $\frac{1}{M}$ and an r.v. U uniformly distributed on (0, 1).
repeat
Calculate $o = \sum_{c=1}^{c'} \lambda_c(t+E), c' \in C$.
until $o \geq M \cdot U$
$t \leftarrow t + E, S_n = S_n \cap \{(t, c')\}$
until $t > T$

Based on the Algorithm 2, we generate a set of event sequences we have mentioned. Fig. 2 depicts the conditional intensity function and the event-occurrence time of 5 types' event sequences.

Fig. 2 describes the condition intensity functions' curves and when and which type the event occurred. We can find out that the higher conditional intensity function, the higher possibility of the event occurred synchronously. The following test of our model will be executed on these event sequences.

5.2. Experimental results of time-varying base intensity data

5.2.1. The choice of hyper-parameters

First, we are supposed to test our model with various hyper-parameters in a wide range, $\alpha_S, \alpha_G \in [10^{-2}, 10^3]$, the curves of Loglike with regard to the two hyper-parameters are shown in following four sub-figure in Fig. 3, the upper part is the sine-impact function and the lower part is square-impact function.

When α_G is changing within the range of $[10^{-2}, 10^3]$, α_S is fixed at 10, when α_S is changing within the range of $[10^{-2}, 10^3]$, α_G is fixed at 100. We can find out that our model is relatively stable when the hyper-parameters is changing in a wide range. According to the experimental result, we can set $\alpha_S = 10$, and $\alpha_G = 100$.

5.2.2. Relative error of parameters

In the following subsections, we will compare the performance of different models on all kinds of measure criteria, such as $e_{\boldsymbol{\mu}}$, $e_{\boldsymbol{\rho}}$, $e_{h(t)}$, $e_{\boldsymbol{\Phi}}$ and Loglike.

Based on discussion in Section 4.1, we know that MLE models fix the value of ρ_c to 1, so these models cannot fit the $h_c(t)$ well, intuitively, we anticipate that the relative error of scale parameter $\boldsymbol{\mu}$ for MLE will bigger than our WB model. The estimate of $\boldsymbol{\mu}$ for MLE model is misled by setting the ρ_c as the fixed value, In Fig. 4, it is shown that the estimated relative error $e_{\boldsymbol{\mu}}$ of the parameter $\boldsymbol{\mu}$ with our model is much smaller than the MLE model. With increasing the number of training sequences, the performance of our model has improved greatly, however the pure MLE model have an increasing relative error, because the time-varying base intensity interferes the estimation of the impact functions and $\boldsymbol{\mu}$ without the regularizers.

As mentioned above, all the MLE-based algorithms assume the scale parameter is fixed to $\rho_c = 1$, so the relative error of $\boldsymbol{\rho}$ is fixed to a constant, while our WB-based models can accurately learn the parameter$\boldsymbol{\rho}$ from the asynchronous event sequences (see Fig. 5). The fluctuation of $\boldsymbol{\rho}$ is due to the randomness of the event sequences. In addition, since our method does not guarantee that each iteration will make $\boldsymbol{\rho}$ to reach a global minimum. It is still difficult to get an unbiased estimate of $\boldsymbol{\rho}$.

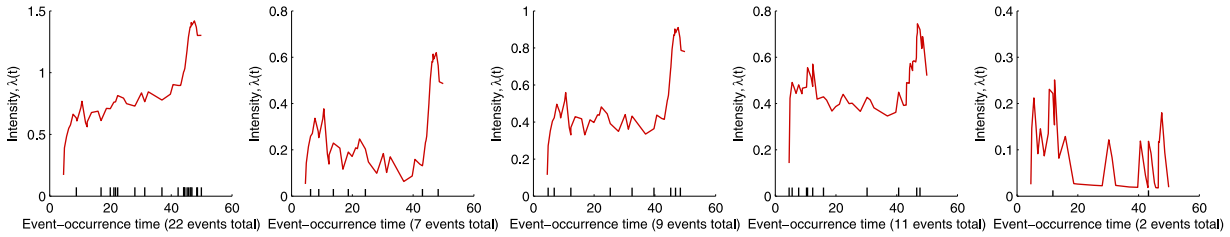
Comparing the scale parameter and the shape parameter of different algorithms separately might be relatively unfair for the MLE-based model (because in fact there are only base intensity parameters which are needed to be estimated), thus, we first define a new measure criterion, the relative error of the base intensity $e_{h(t)}$, by comparing the estimate accuracy of the base intensity, we can figure out which approach's estimate for $h(t)$ is better. From Fig. 6 we can find that, with increasing the number of training sequences, the relative error of all the algorithms is descending, but limited by model assumptions, MLE-Based model cannot fit the base intensity $h(t)$ well. The estimates of $h(t)$ with our models are superior to the MLE models.

In this subsection, we will compare the estimation error of impact functions with the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP, which is the most important part to reflect Granger causality relationship between events. From Fig. 7, we can find that, in sine-like case, when the number of training sequences is 250, the estimation error values of $e_{\boldsymbol{\Phi}}$ with WB-SGL, WB and MLE-SGLP algorithm are 0.763295, 0.781096 and 0.7729, respectively. Meanwhile, in the square-like case, we can figure out that the estimate of the impact function with WB-SGL algorithm is the most accurate, however, due to the influence of high frequency components in the square wave, the performance of the method based on Gauss basis function is affected. The $e_{\boldsymbol{\Phi}}$ values of all MLE-based algorithms are influenced by the larger estimating errors of the base intensity $h(t)$.

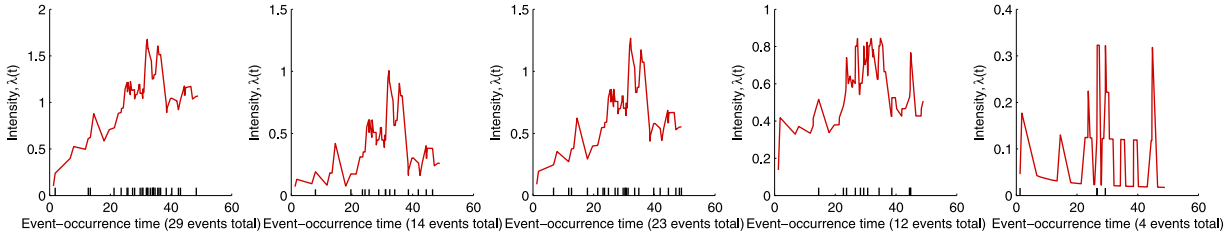
Based on the above analysis, we can validate that when base intensity $h(t)$ is time-varying, WB-based models are superior to the MLE-based ones. The curves of Loglike also verify that. From Fig. 8, we can see that the Loglike of WB-SGL and WB algorithms are much larger than MLE-SGLP and MLE algorithms. Specifically, WB-SGL is better than WB algorithm, both in the sine-like and square-like cases.

5.2.3. Curves of the impact functions

Fig. 9 illustrates the estimates of the impact functions obtained by these four algorithms. The true values of the function $\phi_{1,5}$, $\phi_{2,5}$, $\phi_{3,5}$,

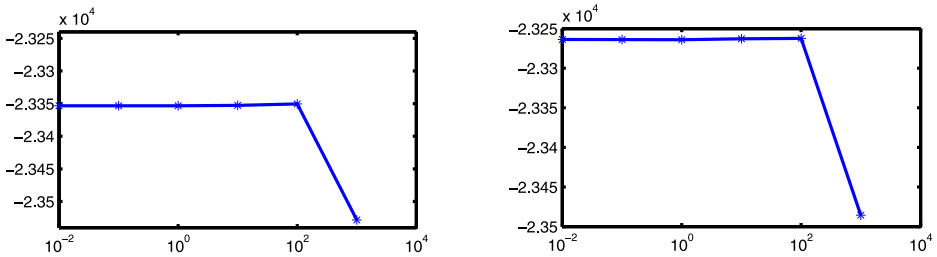


(a) sine-impact function

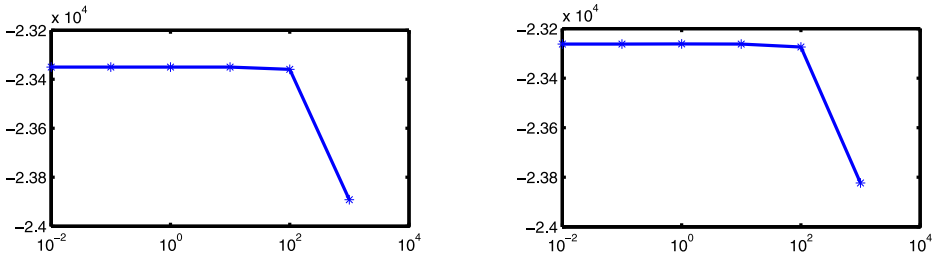


(b) square-impact funciton

Fig. 2. The conditional intensity function and event-occurrence time. The red curve above represents the intensity function, black points represent the time stamp of the corresponding event.

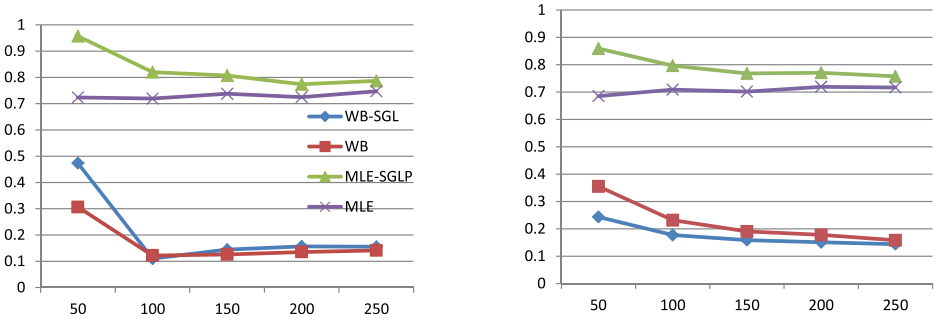


(a) Sine-impact function and α_S (b) Sine-impact function and α_G



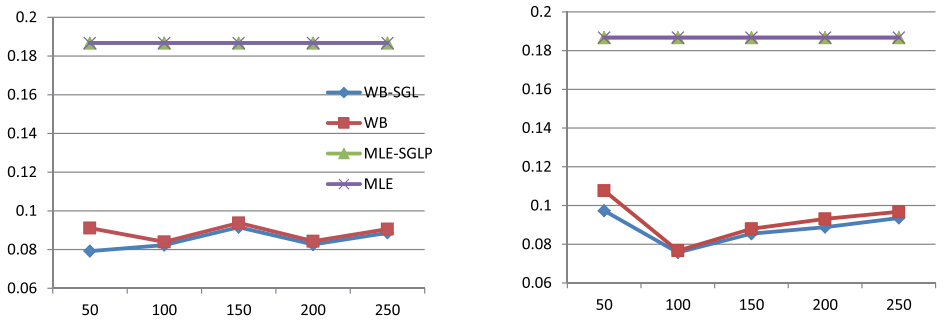
(c) Square-impact function and α_S (d) Square-impact function and α_G

Fig. 3. The curves of Loglike with the change of α_S and α_G.



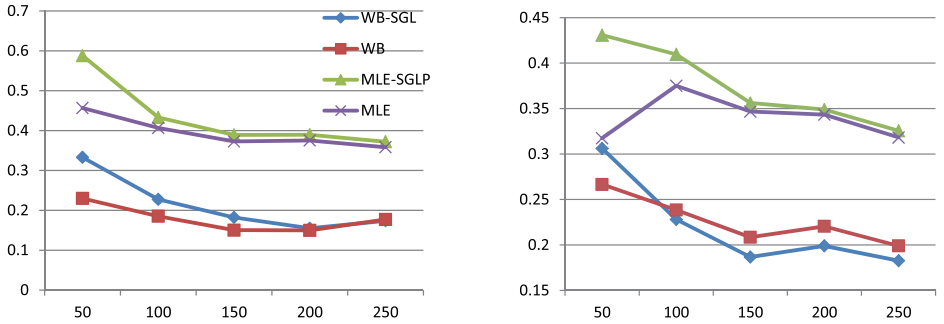
(a) Sine-impact function (b) Square-impact function

Fig. 4. The curves of e_μ of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



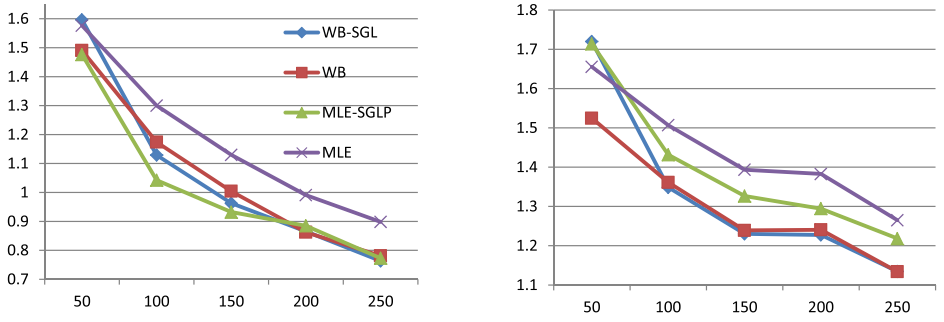
(a) Sine-impact function (b) Square-impact function

Fig. 5. The curves of e_{ρ} of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



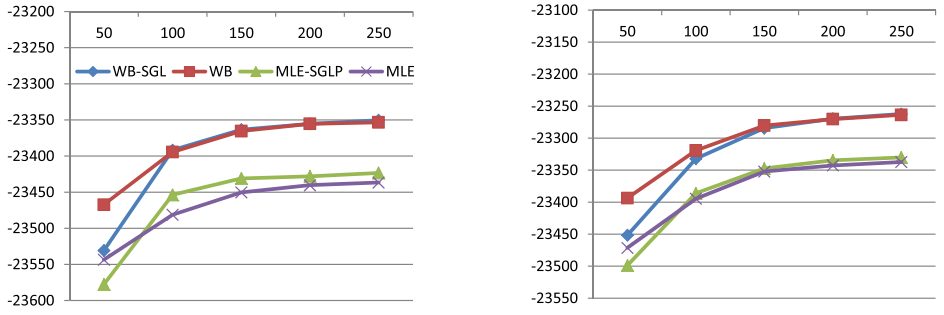
(a) Sine-impact function (b) Square-impact function

Fig. 6. The curves of $e_{h(t)}$ of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



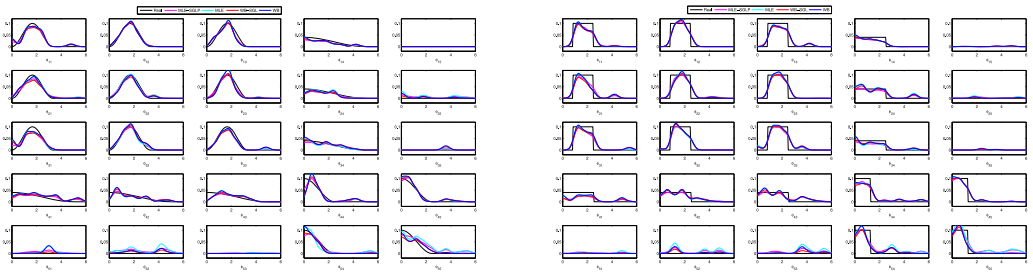
(a) Sine-impact function (b) Square-impact function

Fig. 7. The curves of e_{Φ} of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



(a) Sine-impact function (b) Square-impact function

Fig. 8. The curves of Loglike of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



(a) Sine-impact function (b) Square-impact function

Fig. 9. The curves of $\phi_{cc'}$ of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.

$\phi_{5,1}$, $\phi_{5,2}$ and $\phi_{5,3}$ are all zero. From Fig. 9, we can see that the estimate of the WB-SGL model is the most accurate, which is consistent with the value of $e_{\boldsymbol{\Phi}}(t)$. MLE-based model is misled by the time varying base intensity, due to $\rho_5 \approx 1.24$, thus time varying base intensity $h_5(t)$ is increasing. Meanwhile, MLE-based model assume that $h_5(t)$ is a constant, thus the change of $h_5(t)$ is ignored and deteriorate the estimate accuracy of $\phi_{5c'}$, reflected in the experimental results, the $\phi_{5c'}$ estimate with MLE-based algorithm has a large margin of error. Even the Gaussian basis functions do not fit the square-like impact functions well, our model still can estimate them robustly. However, the algorithm without regularizer has a worse performance, when the impact functions are all zero, they cannot estimate the impact functions accurately. Regularizers restrict the value of $a_{cc'm}$, make the estimate more accurate, but also bring a disadvantage, the restricting the value of $a_{cc'm}$ will obtain a lower parameter estimate of $a_{cc'm}$ and $\phi_{cc'}(t) = \sum_1^M a_{cc'm} g(t)$, the non-zero impact function estimate with regularizers will lower than the ones without regularizers, this point of view can be verified by the experimental results in Fig. 9.

5.3. Experimental results of constant base intensity data

In this subsection, to verify the robustness of our algorithms, we test our model on the constant base intensity data (Xu et al., 2016). In this case, assuming $h(t) = \mu$ will reduce the computational complexity and reduce the interference caused by variable parameters, thus the MLE-based algorithm will have a big advantage. We will still compare all these models on all kinds of measure criteria, such as, $e_{\boldsymbol{\mu}}$, $e_{\boldsymbol{\rho}}$, $e_{h(t)}$, $e_{\boldsymbol{\Phi}}$ and Loglike.

5.3.1. Relative error of parameters

In Fig. 10, we can find that although in WB-based algorithm, we already set the initial value of ρ_c to 1, the interference caused by the randomness of the event sequences still affects the estimate of the parameter $\boldsymbol{\mu}$. However, we can note that as the number of event sequences increases, our algorithm's estimates of $\boldsymbol{\mu}$ are more and more accurately. When the number of event sequences is 250, estimate accuracy of our proposed algorithm is basically close to the MLE-based algorithm.

In Fig. 11, we see that MLE-based algorithms assume ρ_c is fixed to 1, thus there is no the estimate error for ρ_c with MLE-based algorithms. Although WB-based algorithms set the initial value of ρ_c to 1, but because of the randomness of the event sequences, the estimate of ρ_c will fluctuate around the true value. From the experimental results we can get that with increasing the number of event sequences, the estimate of ρ_c is more accurate, the relative error is already less than 5%.

As showed in Fig. 12, because of the estimate error of $\boldsymbol{\rho}$ and $\boldsymbol{\mu}$, the relative error of $h(t)$ with WB-based algorithm is larger than MLE-based algorithm, the relative error with WB-SGL algorithm is closed to the MLE algorithm, which will affect the estimate accuracy of the impact functions.

Table 1
The $e_{\boldsymbol{\Phi}(t)}$ effected by the time varying $h(t)$ (N=250)

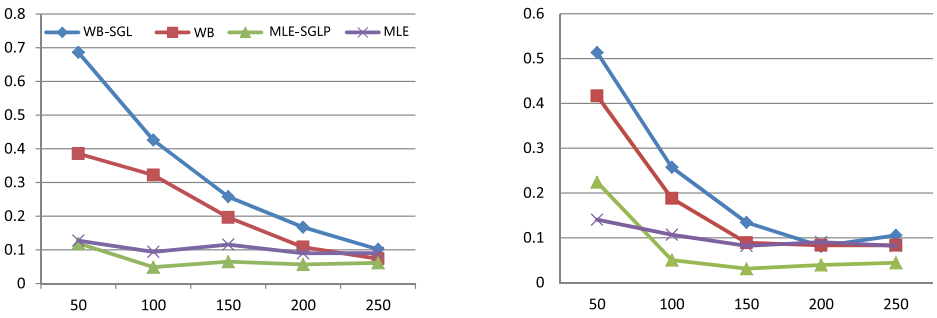
(a) WB-based method				
	Sine-impact function		Square-impact function	
	WB-SGL	WB	WB-SGL	WB
Time varying $h(t)$	0.7633	0.7811	1.1344	1.1337
Constant $h(t)$	0.7712	0.8216	1.1197	1.1157
(b) MLE-based method				
	Sine-impact function		Square-impact function	
	MLE-SGLP	MLE	MLE-SGLP	MLE
Time varying $h(t)$	0.7729	0.8983	1.2184	1.265
Constant $h(t)$	0.6613	0.795	1.0846	1.103

From Fig. 13 and Table 1, we can find out that our proposed algorithm is more robust than MLE-based algorithm, the varying range of the estimate error of MLE-based algorithm's impact function $e_{\boldsymbol{\Phi}(t)}$, is quite larger than WB-based algorithm, although in constant intensity event sequence, MLE-SGLP model has the most accurate estimate of $\boldsymbol{\Phi}(t)$. There is another very interesting findings. In square-like case, the performance of WB-based algorithm in constant $h(t)$ event sequence is a little better than in time varying $h(t)$ event sequence. In our view, the reason for this phenomenon is that, with constant base intensity, the assumption that the base intensity is time-varying compensates for incompatibility between Gaussian basis functions $g_m(t)$ and square-like impact functions $\phi_{cc'}(t)$.

Based on the above discussions, we will analyze Loglike of different models. Undoubtedly, from Fig. 14, we can find that the Loglike of MLE-SGLP model is the highest than other models, which verifies that MLE-SGLP model is superior to other models, and WB-SGL model is superior to the pure MLE model. However, the most important point is that the difference between the Loglike of those algorithms is smaller than in time-varying $h(t)$ event sequences. At the same time, the analysis results of the relative errors of other parameters can also verify this, such as $e_{\boldsymbol{\mu}}$, $e_{\boldsymbol{\rho}}$, $e_{h(t)}$, and $e_{\boldsymbol{\Phi}}$. This validates that the WB-based algorithm we proposed have a strong stability and robustness, both on the time-varying $h(t)$ event sequences and constant $h(t)$ event sequences, WB-based algorithm can obtain more accurate parameter estimation and have a stronger generalization ability.

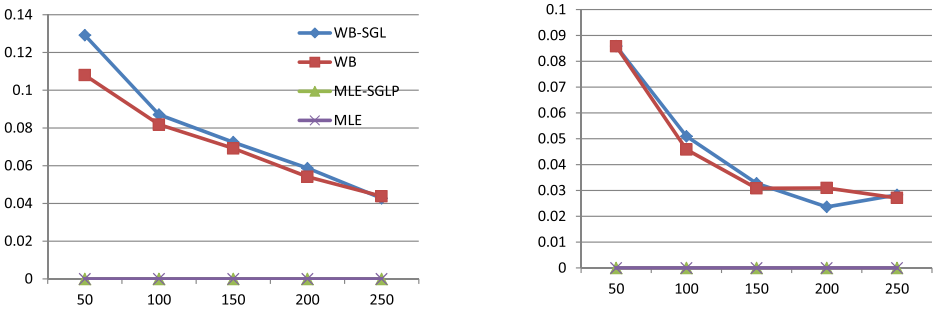
5.3.2. Curves of the impact functions

Fig. 15 illustrates the estimates of the impact functions obtained by these four algorithms on constant $h(t)$ event sequences. The algorithms with regularizers, such as MLE-SGLP and WB-SGL algorithm, are more accuracy in finding the Granger causality. Pure WB algorithm's performance is worst, because of the estimation error of $h(t)$ and without using the regularizers for $a_{cc'm}$.



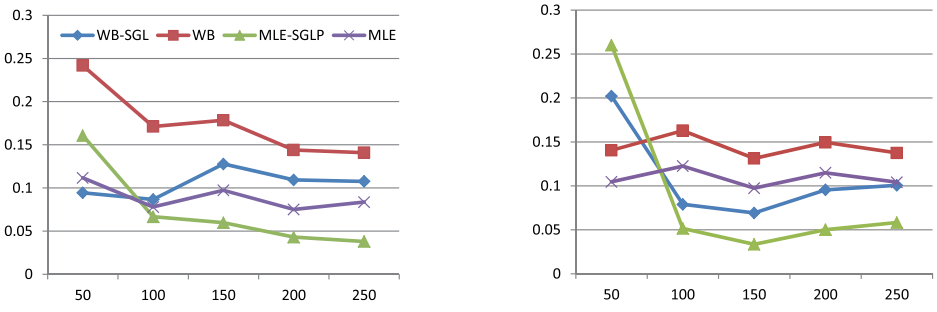
(a) Sine-impact function (b) Square-impact function

Fig. 10. The curves of e_{μ} of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



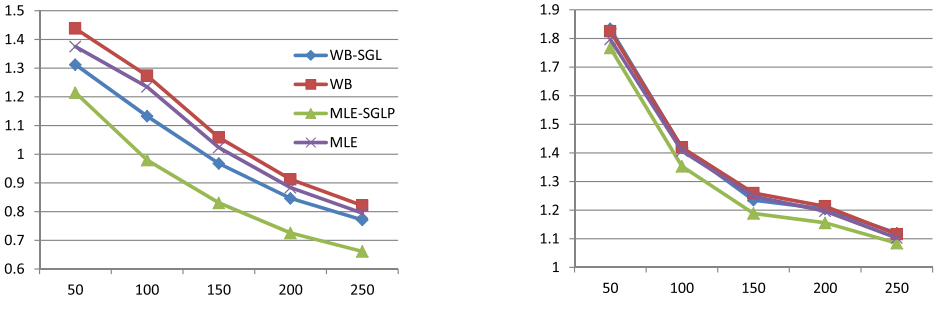
(a) Sine-impact function (b) Square-impact function

Fig. 11. The curves of e_{ρ} of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



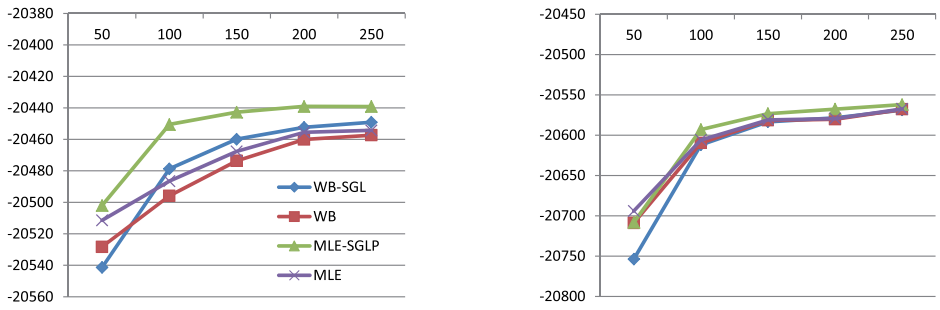
(a) Sine-impact function (b) Square-impact function

Fig. 12. The curves of $e_{h(t)}$ of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



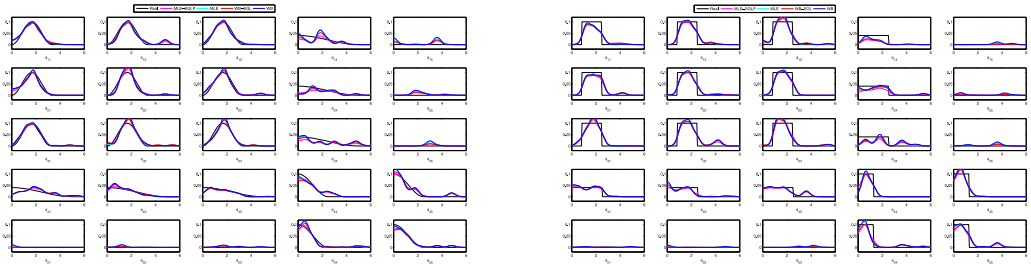
(a) Sine-impact function (b) Square-impact function

Fig. 13. The curves of e_{Φ} of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



(a) Sine-impact function (b) Square-impact function

Fig. 14. The curves of Loglike of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.



(a) sine-impact function (b) square-impact function

Fig. 15. The curves of $\phi_{cc'}$ of the four sorts of models: WB,WB-SGL, MLE, and MLE-SGLP.

Table 2
The 14 types of failures.

Index	Failure types
1	Gas Generator system failure
2	Motor system failure
3	Engine system failure
4	Power supply system failure
5	Fuel gas system failure
6	Lubricating oil system failure
7	Intake system failure
8	Control system failure
9	Cooling system failure
10	Firefighting system failure
11	Compressor system failure
12	Instrument air system failure
13	Process system failure
14	Other failure

5.4. Real-world data

In this section, we will verify the validity of our proposed model on real-world data. This data is labeled by us and is named as compressor station failure dataset. And on this data set we set $\alpha_S = 1$ and $\alpha_G = 1$. The data set contains 87 compressor stations' failure event sequences from March 2006 to October 2018. There are 14 subsystems in each compressor station, so all the failures are categorized into 14 categories as Table 2 according to the subsystem in which the failure occurred.

All of these types of failures have both self-triggering and mutually-triggering patterns. For instance, if the engine system fails, the possibility of future engine system failure will increase (self-triggering), if the gas generator system fails, the possibility of future control system failure will increase (mutually-triggering). Not only that, the trend of the base intensity of each system failure with time can also be obtained through learning which is depending on the parameters ρ and μ. Thus, we model the failure event sequence via a Hawkes process with time-varying base intensity, which is from Weibull distribution, while Weibull distribution is already widely used in reliability analysis, then we can learn the Granger causality among the failure categories and the varying trend of base intensity.

Table 3
The Loglike of various models on real-world data.

WB series model				MLE series model			
WB-SGL	WB-S	WB-GL	WB	MLE-SGL	MLE-S	MLE-GL	MLE
−5054.8	−5055.8	−5056.21	−5061.3	−5755.3	−5610.3	−5760.7	−5764.5

5.4.1. The Loglike of different models

The sequence of failure events of eighty compressor stations is used as training data, and the sequence of failure events of the remaining compressor stations is used as test data. Consider the fact that the compressor station is inspected for each production quarter and the shortest stable running time is three days, we set the length of time of the impact function to be 90 days (the influence of a failure will not exist in a production quarter), and the number of basis function $M = 31$ (sampling every 72 h). The parameter of the kernel Gaussian function is defined as:

$$\sigma = \frac{90}{30 \times 2} = 1.5$$

Table 3 lists the Loglike obtained via 8 sorts of models (WB-SGL, WB-S, WB-GL, WB, MLE-SGL, MLE-S, MLE-GL, and MLE) with compressor station failure data. We can figure out that the WB models are far better than the MLE models, and WB-SGL model obtains the best experimental result. This is in line with the assumptions we made at the beginning of this paper, the base intensity of the Hawkes process is not a constant in practical applications, but a time-varying function.

And we assume that the base intensity is consistent with the Weibull distribution and it has achieved better results in practical applications.

5.4.2. The estimation of Weibull base intensity parameters

In order to get the most accuracy estimation of the parameters to analyze the failures' occurrence, we learn our model on the all failure event sequence. Table 4 shows the estimates of μ and ρ via four different WB-based models. We can see that the parameter estimates of different models are basically the same, which can verify that our algorithm is stable. Below we analyze the reliability of the compressor station with the parameters of the best model on test results, i.e., the WB-SGL model.

Based on the estimates of parameters μ and ρ with WB-SGL model, we can figure out that these compressors stations are in normal working condition and have superior reliability. However, there are still two issues that need our attention. For instance, Control system, Cooling system, Compressor system and Instrument system failure have a higher value of scale parameter, and a lower value of shape parameter ρ, according to the definition of hazard function of Weibull base intensity $h(t) = \mu\rho t^{\rho-1}$, these failures are more likely to occur when the compressor station is just starting production, and the likelihood of these occurred failures will gradually decrease over time.

On the other hand, for Fuel gas system failures, Intake system failures, Firefighting system failures and other failures, due to the high value of ρ ($\rho > 0.9$), although it is still less than 1, the parameter estimation error, the environmental negative affect, misoperation, and equipment aging, etc., may gradually shift the hazard function from decreasing over time to increasing over time, thus we need to pay attention to the risks of these failures after a long run. The exciting thing is that the estimated trend of system reliability based on parameters is consistent with the experiences of the operation and maintenance experts of the compressor station. This proves the validity and accuracy of our proposed algorithm.

5.4.3. Granger causality of different models

In this subsection, we will analyze the trigger patterns of all kinds of failures. We test the four models both on our model and the MLE model on the entire data set. In order to more intuitively represent the causal relationship between each type of fault, we construct the infectivity matrices of different types of failures in Fig. 16. The element in the cth row and the c'th column of the infectivity matrices is $\int_0^\infty \phi_{cc'}(s)ds$. The infectivity matrices of the same series of models are basically the same, so here we only show the infectivity matrices of the WB-SGL and the MLE-SGL model.

The closer the color in the matrix is to blue, which indicates the weaker the causal relationship between the two types of faults. We can see that most of the elements in most infectivity matrices are dark blue, which reflects the sparseness of Granger causality. From Fig. 16, It can be seen that the sparsity of our proposed model is better than that of the MLE-SGL model.

In general, after a certain type of fault occurs, the probability of occurrence of this fault will increase. This means that after a certain type of fault occurs, there is a possibility of a secondary failure in a short period of time. The infectivity matrices reveal this phenomenon: the main diagonal elements of the infectivity matrices are larger than most other elements in the matrices. Especially in engine system, motor system, power supply system, and lubricant system's diagonal elements are larger than other main diagonal elements. This result implies that Engine system failure has the strongest self-trigger correlation, which implies that if an engine system failure occurred, then the probability of a secondary failure occurring within 90 days is quite large. Motor system, power supply system, and lubricant system are also prone to secondary failures.

In addition, after other systems fail, the control system is most susceptible to failures caused by remain system failures, such as the gas generator system failure, power supply system failure, instrument air system failure and other failure. This reflects the susceptibility of the control system, because there are intimate relations between control system and other system.

For a more detailed analysis of Granger causality between failures and impact function (which indicates failure probability under the influence of historical fault events) over time, we rank the infective element from high to low, and show the top-28 impact functions in Fig. 17. Observing top-28 impact functions, we can roughly divide the impact function into three categories.

First, depending on whether the impact functions have a delay part, we divide the trigger pattern of failures into two categories. If the impact functions are close to 0 at the beginning, then we call these trigger patterns as delay trigger patterns. Such as $\phi_{8,13}$ and $\phi_{7,10}$ in Fig. 17 at the beginning, the value of this type of impact function is not very high, even close to 0, but one or more spikes suddenly will appear over time, operation and maintenance personnel need to pay more attention to possible secondary faults at the corresponding moments where the spikes appear. From the perspective of the control field, the reason of this occurring phenomenon may be that there is a delay between these systems, i.e. firefighting system to intake system and process system to control system.

In Table 5, we will list the characteristics of delay trigger patterns, including the length of the delay and the moment when the peak risk occurs.

Secondly, according to the attenuation of the impact function, we divide the rest impact function into two categories. If the impact function can decay to 10% of its maximum value within 50 days, then we consider the trigger pattern between the two failures is stable, otherwise, it is unstable.we use $\phi_{3,3}$ and $\phi_{2,2}$ as examples, which is the illustration in Fig. 17. When a fault just occurs, the probability of a secondary fault is very high, but it gradually decreases with time. The impact functions of this category are mainly the self-trigger impact function of the failures. When these faults occur, the operation and maintenance personnel should remain vigilant within a short period of time after handling the fault to prevent occurring secondary faults, such as engine system failure, motor system failure, power supply system failure, control system failure, cooling system failure, gas generator system failure, Fuel gas system failure, compressor system failure, instrument air system failure, and Intake system failure. In Table 6, we will list the higher risks duration of stable failure trigger.

More specifically, for a stable trigger pattern, after a failure occurs, the probability of occurrence of a derivative fault is generally reduced with time, while for the unstable trigger pattern, the probability of occurrence of a derivative failure is only slightly reduced, or instead will increase. Such as $\phi_{8,12}$ and $\phi_{8,5}$ depicting in Fig. 17, this kind of impact function has a large value in a short time after the source failure occurs, and there are still several peaks appearing with the passage of time. After handling the source fault, the operation and maintenance personnel should not only be vigilant in a short period of time, but also should be vigilant at the moment of the corresponding peak value to prevent the occurrence of secondary faults, Such as instrument air system to control system and fuel gas system to control system. We will list the peak appearance time of unstable failure trigger patterns in Table 7.

In summary, after we know the sequence of historical failure events, we can use the learned Weibull–Hawkes model to predict future failures, so as to respond to measure and reduce potential economic and security losses.

6. Conclusion

In this paper, we modify the structure of the Hawkes process, introduce a time-varying base intensity obeying Weibull distribution, the Weibull base intensity, and propose a series of Hawkes process model. Based on our proposed model and the EM-based algorithm, we provide an algorithm that effectively learns the parameters in the

Table 4
The parameter estimation of various models on real-world data.

Index	WB-SGL		WB-S		WB-GL		WB	
Parameter	μ	ρ	μ	ρ	μ	ρ	μ	ρ
1	0.004168	0.676046	0.004158	0.676254	0.004166	0.676083	0.004164	0.676103
2	0.002214	0.69372	0.002216	0.693572	0.002219	0.69345	0.002219	0.69341
3	0.000413	0.894198	0.000412	0.894285	0.000413	0.894166	0.000412	0.894279
4	0.003276	0.886825	0.003273	0.88689	0.003276	0.886823	0.003273	0.886886
5	0.000433	0.907548	0.000432	0.907589	0.000433	0.907548	0.000432	0.907592
6	0.001012	0.810709	0.001011	0.810765	0.001013	0.810642	0.001011	0.810756
7	0.000263	0.940528	0.000262	0.940535	0.000263	0.940528	0.000262	0.940557
8	0.01711	0.763604	0.017102	0.763622	0.017122	0.763525	0.017098	0.763646
9	0.001053	0.779977	0.001052	0.780047	0.001054	0.779939	0.001053	0.779989
10	9.08×10^{-5}	0.980752	9.07×10^{-5}	0.980747	9.08×10^{-5}	0.980753	9.07×10^{-5}	0.980745
11	0.0129	0.595825	0.012873	0.596012	0.012892	0.595895	0.012874	0.596006
12	0.008312	0.433294	0.00829	0.43353	0.008317	0.433229	0.008294	0.433486
13	0.000224	0.895607	0.000223	0.895686	0.000224	0.895594	0.000224	0.895627
14	1.85×10^{-5}	0.985088	1.85×10^{-5}	0.9851	1.85×10^{-5}	0.985088	1.85×10^{-5}	0.985104

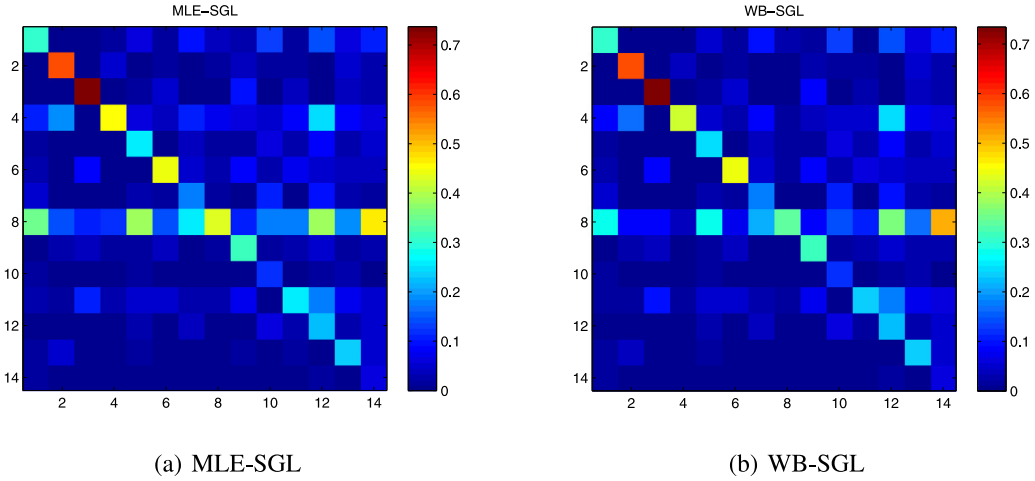


Fig. 16. The infectivity matrices of WB-SGL and MLE-SGL. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

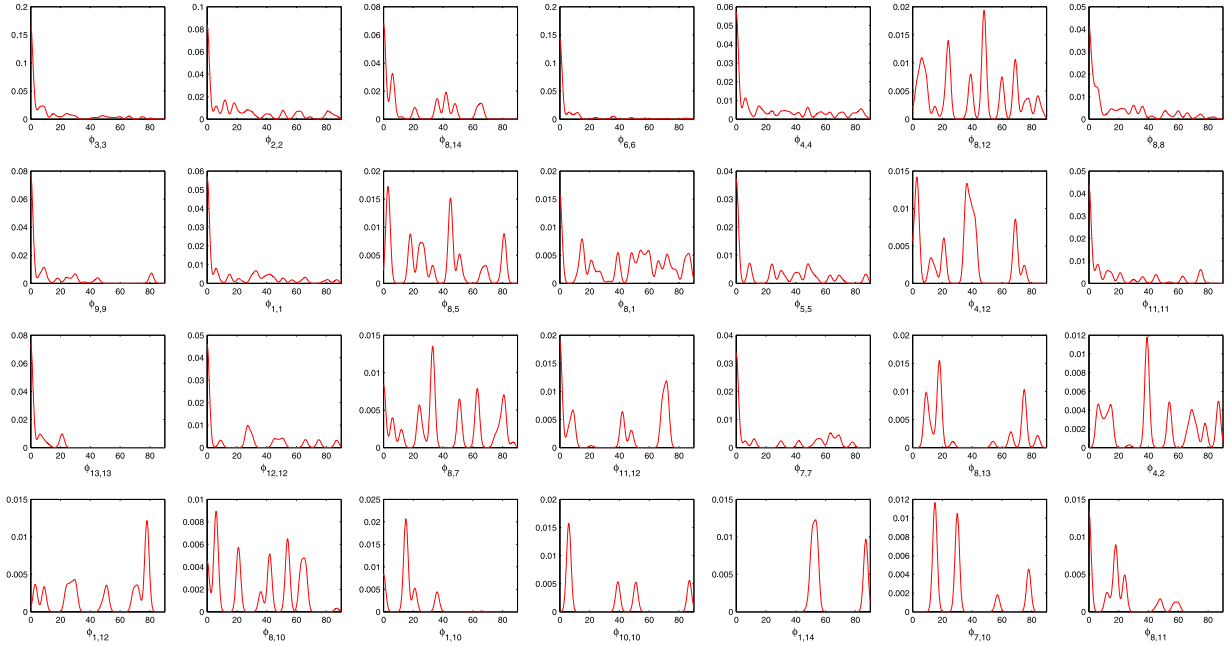


Fig. 17. The curves of the top-28 $\phi_{cc'}(t)$ between different types of failures.

Table 5
The characteristics of delay trigger patterns.

Failure trigger	Length of the delay (d)	Peak occurrence time (d)
Process system failure to Control system failure	10	18.00
Firefighting system failure to Intake system failure	4	14.65
Other failure to Gas Generator system failure	45	53.20

Table 6
Higher risks duration of stable failure trigger.

Failure trigger	Higher Risks duration (d)
Engine system failure to Engine system failure	0–10.0
Motor system failure to Motor system failure	0–20.5
Power supply system failure to Power supply system failure	0–48.7
Lubricating oil system failure to Lubricating oil system failure	0–3.4
Control system failure to Control system failure	0–37.2
Gas Generator system failure to Gas Generator system failure	0–33.9
Process system failure to Process system failure	0–22.1
Instrument air system failure to Instrument air system failure	0–30.9

Table 7
Peak appearance time of unstable failure trigger.

Failure trigger	Peak occurrence time (d)
Other failure to Control system failure	0.0
Instrument air system failure to Control system failure	48.0
Fuel gas system failure to Control system failure	3.0
Cooling system failure to Cooling system failure	0.0
Gas Generator system failure to Control system failure	0.0
Instrument air system failure to Power supply system failure	2.9
Fuel gas system failure to Fuel gas system failure	0.0
Intake system failure to Control system failure	33.0
Instrument air system failure to Compressor system failure	0.0
Process system failure to Control system failure	18.0
Intake system failure to Intake system failure	0.0
Motor system failure to Power supply system failure	39.0
Instrument air system failure to Gas Generator system failure	78.0
Firefighting system failure to Control system failure	6.0
Firefighting system failure to Gas Generator system failure	15.0
Firefighting system failure to Firefighting system failure	6.0
Compressor system failure to Control system failure	0.0

model. To demonstrate the effectiveness, stability, and robustness of our model, we tested our proposed approaches both on time-varying base intensity data and constant base intensity data.

The most important point, which deserves our attention, is that in the practical application of the Hawkes process, assuming the base intensity is constant is extremely unrealistic. This assumption constricts the development of the Hawkes process in practical applications. Setting $h(t)$ is time-varying is more flexible and more generalization. This assumption is more in line with the needs of real-world applications. We train our model on the historical data of the compressor station failure, obtain a lot of valuable experimental results worth analyzing. After that, we make some suggestions for the actual production of the compressor station.

In the future, we are interested in the following work: create a better parameter learning algorithm and devise a new point process model that better fits the needs of actual production applications.

CRediT authorship contribution statement

Lu-ning Zhang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Visualization, Data curation. **Jian-wei Liu:** Conceptualization, Methodology, Validation, Resources, Formal analysis, Writing - review & editing, Supervision, Project administration. **Xin Zuo:** Project administration, Supervision, Funding acquisition, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

I would like to thank Ms. Zhang Yu-ying, Ms. Wang Yue-qiao and Prof. Luo Xiong-lin for their selfless help when I was facing difficulties. Thanks Dr. Xu Hong-teng for his generous help in the early days of my research. Thank all my friends for their company and help.

References

Adams, Ryan Prescott, Murray, Iain, MacKay, David J.C., 2009. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 9–16.

Ahmed, Amr, Xing, Eric P., 2009. Recovering time-varying networks of dependencies in social and biological studies. Proc. Natl. Acad. Sci. 106 (29), 11878–11883.

Ali, Jaouher Ben, Chebel-Morello, Brigitte, Saidi, Lotfi, Malinowski, Simon, Fnaiech, Farhat, 2015. Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network. Mech. Syst. Signal Process. 56, 150–172.

Arnold, Andrew, Liu, Yan, Abe, Naoki, 2007. Temporal causal modeling with graphical granger methods. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 66–75.

Bacry, Emmanuel, Dayri, Khalil, Muzy, Jean-François, 2012. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. Eur. Phys. J. B 85 (5), 157.

Bacry, Emmanuel, Mastromatteo, Iacopo, Muzy, Jean-François, 2015. Hawkes processes in finance. Mark. Microstruct. Liq. 1 (01), 1550005.

Bain, Lee, 2017. Statistical Analysis of Reliability and Life-Testing Models: Theory and Methods. Routledge.

Basu, Sumanta, Shojaie, Ali, Michailidis, George, 2015. Network granger causality with inherent grouping structure. J. Mach. Learn. Res. 16 (1), 417–453.

Caruana, Rich, Lawrence, Steve, Giles, C. Lee, 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In: Advances in Neural Information Processing Systems. pp. 402–408.

Christodoulou, Symeon E., 2011. Water network assessment and reliability analysis by use of survival analysis. Water Resour. Manage. 25 (4), 1229–1238.

Chwialkowski, Kacper, Gretton, Arthur, 2014. A kernel independence test for random processes. In: International Conference on Machine Learning. pp. 1422–1430.

Cox, David Roxbee, 2018. Analysis of Survival Data. Routledge.

Crowder, Martin, 2006. Competing risks. Encyclopedia Actuar. Sci. 1.

Daley, Daryl J., Vere-Jones, David, 2007. An Introduction to the Theory of Point Processes: Vol. II: General Theory and Structure. Springer Science & Business Media.

Daneshmand, Hadi, Gomez-Rodriguez, Manuel, Song, Le, Schölkopf, Bernhard, 2014. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. In: International Conference on Machine Learning. pp. 793–801.

Didelez, Vanessa, 2008. Graphical models for marked point processes based on local independence. J. R. Stat. Soc. Ser. B Stat. Methodol. 70 (1), 245–264.

Du, Nan, Dai, Hanjun, Trivedi, Rakshit, Upadhyay, Utkarsh, Gomez-Rodriguez, Manuel, Song, Le, 2016. Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1555–1564.

Du, Nan, Song, Le, Yuan, Ming, Smola, Alex J., 2012. Learning networks of heterogeneous influence. In: Advances in Neural Information Processing Systems. pp. 2780–2788.

Duchateau, Luc, Janssen, Paul, 2007. The Frailty Model. Springer Science & Business Media.

Eichler, Michael, 2012. Graphical modelling of multivariate time series. Probab. Theory Related Fields 153 (1–2), 233–268.

Eichler, Michael, Dahlhaus, Rainer, Dueck, Johannes, 2017. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. J. Time Series Anal. 38 (2), 225–242.

Embrechts, Paul, Liniger, Thomas, Lin, Lu, 2011. Multivariate Hawkes processes: an application to financial data. J. Appl. Probab. 48 (A), 367–378.

Farajtabar, Mehrdad, Du, Nan, Rodriguez, Manuel Gomez, Valera, Isabel, Zha, Hongyuan, Song, Le, 2014. Shaping social activity by incentivizing users. In: Advances in Neural Information Processing Systems. pp. 2474–2482.

Gunawardana, Asela, Meek, Christopher, Xu, Puyang, 2011. A model for temporal dependencies in event streams. In: Advances in Neural Information Processing Systems. pp. 1962–1970.

Hall, Eric C., Willett, Rebecca M., 2016. Tracking dynamic point processes on networks. IEEE Trans. Inform. Theory 62 (7), 4327–4346.

Han, Fang, Liu, Han, 2013. Transition matrix estimation in high dimensional time series. In: International Conference on Machine Learning. pp. 172–180.

Hansen, Niels Richard, Reynaud-Bouret, Patricia, Rivoirard, Vincent, et al., 2015. Lasso and probabilistic inequalities for multivariate point processes. Bernoulli 21 (1), 83–143.

Hawkes, Alan G., 1971. Spectra of some self-exciting and mutually exciting point processes. Biometrika 58 (1), 83–90.

Laub, Patrick J., Taimre, Thomas, Pollett, Philip K., 2015. Hawkes processes. arXiv preprint arXiv:1507.02822.

Leger, Stefan, Zwanenburg, Alex, Pilz, Karoline, Lohaus, Fabian, Linge, Annett, Zöphel, Klaus, Kotzerke, Jörg, Schreiber, Andreas, Tinhofer, Inge, Budach, Volker, et al., 2017. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci. Rep. 7 (1), 1–11.

Lemonnier, Remi, Vayatis, Nicolas, 2014. Nonparametric markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 161–176.

Lewis, Erik, Mohler, George, 2011. A nonparametric EM algorithm for multiscale Hawkes processes. J. Nonparametr. Stat. 1 (1), 1–20.

Li, Kan, Chan, Wenyaw, Doody, Rachelle S, Quinn, Joseph, Luo, Sheng, 2017. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. J. Alzheimer's Dis. 58 (2), 361–371.

Lian, Wenzhao, Henao, Ricardo, Rao, Vinayak, Lucas, Joseph, Carin, Lawrence, 2015. A multitask point process predictive model. In: International Conference on Machine Learning. pp. 2030–2038.

Lloyd, Chris, Gunter, Tom, Osborne, Michael, Roberts, Stephen, 2015. Variational inference for Gaussian process modulated Poisson processes. In: International Conference on Machine Learning. pp. 1814–1822.

Luo, Dixin, Xu, Hongteng, Zhen, Yi, Ning, Xia, Zha, Hongyuan, Yang, Xiaokang, Zhang, Wenjun, 2015. Multi-task multi-dimensional Hawkes processes for modeling event sequences. In: Twenty-Fourth International Joint Conference on Artificial Intelligence.

Meek, Christopher, 2014. Toward learning graphical and causal process models. In: Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction, Vol. 1274. CEUR-WS. org, pp. 43–48.

Mei, Hongyuan, Eisner, Jason M., 2017. The neural Hawkes process: A neurally self-modulating multivariate point process. In: Advances in Neural Information Processing Systems. pp. 6754–6764.

Mohammadi, Kasra, Alavi, Omid, Mostafaeipour, Ali, Goudarzi, Navid, Jalilvand, Mahdi, 2016. Assessing different parameters estimation methods of Weibull distribution to compute wind power density. Energy Convers. Manage. 108, 322–335.

Ogata, Yosihiko, 1981. On Lewis' simulation method for point processes. IEEE Trans. Inform. Theory 27 (1), 23–31.

Peña, Edsel A., Hollander, Myles, 2004. Models for recurrent events in reliability and survival analysis. In: Mathematical Reliability: An Expository Perspective. Springer, pp. 105–123.

Pereira, Gabriel KR, Guilardi, Luís F, Dapieve, Kiara S, Kleverlaan, Cornelis J, Rippe, Marilia P, Valandro, Luiz Felipe, 2018. Mechanical reliability, fatigue strength and survival analysis of new polycrystalline translucent zirconia ceramics for monolithic restorations. J. Mech. Behav. Biomed. Mater. 85, 57–65.

Pintilie, Melania, 2006. Competing Risks: A Practical Perspective, Vol. 58. John Wiley & Sons.

Reynaud-Bouret, Patricia, Schbath, Sophie, et al., 2010. Adaptive estimation for Hawkes processes; application to genome analysis. Ann. Statist. 38 (5), 2781–2822.

Samo, Yves-Laurent Kom, Roberts, Stephen, 2015. 2015. scalable nonparametric Bayesian inference on point processes with Gaussian processes. in: international conference on machine learning. pp. 2227–2236.

Simon, Noah, Friedman, Jerome, Hastie, Trevor, Tibshirani, Robert, 2013. A sparse-group lasso. J. Comput. Graph. Statist. 22 (2), 231–245.

Song, Dong, Wang, Haonan, Tu, Catherine Y, Marmarelis, Vasilis Z, Hampson, Robert E, Deadwyler, Sam A, Berger, Theodore W, 2013. Identification of sparse neural functional connectivity using penalized likelihood estimation and basis functions. J. Comput. Neurosci. 35 (3), 335–357.

Tierney, Jayne F, Stewart, Lesley A, Ghersi, Davina, Burdett, Sarah, Sydes, Matthew R, 2007. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials 8 (1), 16.

Vere-Jones, David, 2003. An Introduction to the Theory of Point Processes: Vol. I: Elementary Theory and Methods. Springer.

Xiao, Shuai, Yan, Junchi, Yang, Xiaokang, Zha, Hongyuan, Chu, Stephen M, 2017. Modeling the intensity function of point process via recurrent neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence.

Xu, Hongteng, Farajtabar, Mehrdad, Zha, Hongyuan, 2016. Learning granger causality for Hawkes processes. In: International Conference on Machine Learning. pp. 1717–1726.

Xu, Zenglin, Jin, Rong, Yang, Haiqin, King, Irwin, Lyu, Michael R, 2010. Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th International Conference on Machine Learning. ICML-10, Citeseer, pp. 1175–1182.

Yan, Junchi, Zhang, Chao, Zha, Hongyuan, Gong, Min, Sun, Changhua, Huang, Jin, Chu, Stephen, Yang, Xiaokang, 2015. On machine learning towards predictive sales pipeline analytics. In: Twenty-Ninth AAAI Conference on Artificial Intelligence.

Zhang, Quan, Zhou, Mingyuan, 2018. Nonparametric Bayesian lomax delegate racing for survival analysis with competing risks. In: Advances in Neural Information Processing Systems. pp. 5002–5013.

Zhao, Qingyuan, Erdogdu, Murat A, He, Hera Y, Rajaraman, Anand, Leskovec, Jure, 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 1513–1522.

Zhou, Ke, Zha, Hongyuan, Song, Le, 2013a. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In: Artificial Intelligence and Statistics. pp. 641–649.

Zhou, Ke, Zha, Hongyuan, Song, Le, 2013. Learning triggering kernels for multi-dimensional Hawkes processes. In: International Conference on Machine Learning. pp. 1301–1309.