University of Belgrade
Faculty of Organizational Sciences

XVII INTERNATIONAL SYMPOSIUM

2020 symorg

BUSINESS
AND
ARTIFICIAL INTELLIGENCE

September 07 - 09, 2020

online

SERBIA

UNIVERSITY OF BELGRADE
FACULTY OF ORGANIZATIONAL SCIENCES

symorg 2020

BUSINESS
AND
ARTIFICIAL INTELLIGENCE

SYMORG 2020

Belgrade, September 7-9, 2020
(Online)

XVII INTERNATIONAL SYMPOSIUM

BUSINESS AND ARTIFICIAL INTELLIGENCE

SYMPOSIUM PROCEEDINGS

Editors:
Dušan Starčević, Ph.D.
Sanja Marinković, Ph.D.

Belgrade, 2020

BUSINESS
AND
ARTIFICIAL INTELLIGENCE

Publisher
University of Belgrade – Faculty of Organizational Sciences
Jove Ilića 154, Belgrade, Serbia
www.fon.bg.ac.rs

Dean of Faculty of Organizational Sciences
Milija Suknović, Ph.D.

Designed by
Marina Dobrota, Ph.D., Minja Marinović

Printing
SAJNOS DOO NOVI SAD.
Momčila Tapavice 2, Novi Sad, SERBIA

Year
2020

Conference Organizer
University of Belgrade - Faculty of Organizational Sciences, Serbia

---

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

005:004(082)(0.034.2)
005.6(082)(0.034.2)
005.8(082)(0.034.2)
658:004(082)(0.034.2)
004.738.5:658(082)(0.034.2)

INTERNATIONAL symposium business and artificial intelligence (17 ; 2020 ; Beograd)
Symposium proceedings [Elektronski izvor] / XVII International Symposium Business and Artificial Intelligence , SYMORG Belgrade, September 7-9, 2020 ; editors Dušan Starčević, Sanja Marinković ; [conference organizer Faculty of organizational sciences]. - Belgrade : Faculty of organizational sciences, 2020 (Novi Sad : Sajnos). - 1 elektronski optički disk (CD-ROM) : tekst ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovnog ekrana. - Tiraž 100. - Bibliografija uz svaki rad.

ISBN 978-86-7680-385-9

а) Бизнис -- Информациона технологија -- Зборници
б) Електронско пословање -- Зборници
в) Пројектни менаџмент -- Зборници
г) Менаџмент људских ресурса -- Зборници
д) Предузећа -- Пословање -- Информациона технологија -- Зборници

COBISS.SR-ID 22734345

A NOVEL APPROACH FOR LEARNING TEMPORAL POINT PROCESS

Dimitrije Milenković[1], Andrija Petrović*[2], Uglješa Bugarić[2]
[1]University of Belgrade Faculty of Organizational Sciences
[2] University of Belgrade Faculty of Mechanical Engineering
*Corresponding author, e-mail: aapetrovic@mas.bg.ac.rs

Abstract: *In this paper, we presented a novel methodology for learning temporal point process based on the implementation of one-dimensional numerical integration techniques. The implementation of numerical methodology is used for linearizing negative maximum likelihood (neML) function to enable backpropagation of neML derivative. The presented approach is tested on highway toll dataset. Moreover, four different well-known point process baseline models were compared: first-order and second-order polynomial Poisson inhomogeneous process and Hawkes with exponential and Gaussian kernel. The results showed that different numerical integration techniques influence the quality of the obtained models.*

Keywords*: traffic prediction, temporal point process, Hawkes process, Poisson process, numerical integration*

1. INTRODUCTION

Nowadays, one of the most popular research areas is focused on modelling event sequence. Event sequence has become extremely popular in a variety applications such as traffic (Ryu & Steven, 1998), epidemiology (Zahrieh, 2017), network activities (Liu, 2018), bioinformatics (Farajtabar et al., 2017), e-commerce etc. The event data carry important information about timestamps when an event occurred. Additionally, event data can also provide information concerning event attribute such as class of event, type, participator, etc. This type of point process is known as a marked point process. Compared to the time-series event occurrences are treated as random variables generated in an asynchronous manner, which makes them fundamentally different from time series where equal and fixed time intervals are considered. This property makes them useful in a wide variety of applications where discretizing events to fixed interval results in bad prediction performances and high computational cost.

Generally, there are two types of point process models: temporal (univariate) point process and spatial-temporal (multivariate) point process. In the case of the univariate point process, the objective is to model temporally correlated event occurrence, whereas in spatial-temporal point process the event occurrences are correlated in space and time. Multivariate point process is generally mostly used in the analysis of protein patterns (Jacobsen et al., 2007) and financial market predictions (Bowsher, 2007). The general formulation of the point process makes them available to model event occurrences continuous or discontinuous (with jumps). Additionally, the point process can be further generalized by stochastic differential equations to stochastic point process.

The main idea behind different types of point process models is hidden in modelling conditional intensity function. Conditional intensity function can be interpreted heuristically as expected number of events that are going to occur in infinitesimally small timestamp (*dt*). This intensity function can be modelled as constant (homogeneous process) or function of time (inhomogeneous process). Learning intensity function (Mei & Eisner, 2017) from given dataset present one of the most popular subjects of research. A point process is extremely useful in modelling traffic congestion and traffic event occurrences (Jia, Jiang, Liu, Cui, & Shi, 2018) (arrival of vehicles, pedestrian movement, etc.). Simulating highway traffic and predicting highway congestion (Nguyen, Krishnakumari, Calvert, Vu, & Van Lint, 2019) is one of the main problems connected with point process modelling. In the case of highway congestion, the event occurrences can be described as the number of vehicles that pass highway toll.

In this paper, we presented the data-driven approach for learning different types of conditional intensity functions used in temporal point process models. Our approach is based on implementing numerical integration methods to linearize negative maximum likelihood (neML) and backpropagate derivative of neML. Based on the dataset that consists of exact timestamps when vehicle passes the highway toll we implemented provided methodology and showed that it can be successfully used for any type of conditional intensity function. Furthermore, we compared four different baseline models based on neML scores: first-order and second-order polynomial Poisson inhomogeneous process, Hawkes with exponential and Gaussian kernel.

2. RELATED WORK

We structure the discussion of related work onto the two broad previously mentioned categories: intensity approaches and intensity-free approaches. The intensity approaches present the methods where the point process is modelled by different functional forms of conditional intensity functions (Rasmussen, 2011). In the case of an intensity-free methods point process is modelled with some type of unsupervised machine learning algorithms.

Intensity approaches present the oldest approaches in modelling point process. They rely on functional form that completely depends on the form of conditional intensity function. The poisson process presents the simplest point process where conditional intensity function has constant value (Last & Penrose, 2017). The more complicated variant of this process is observed when conditional intensity function is modelled as a product of kernels (Kirchner, 2017). Recent research proposed different variants of modelling intensity function by deep neural networks (Mei & Eisner, 2017; Xiao, Yan, Yang, Zha, & Chu, 2017). Xiao et al. presented an interesting approach of modelling intensity function by a recurrent neural network. However, in this paper authors assumed that integral in negative maximum likelihood is correlated only with the current timestamp. Even though, this strong assumption cannot be justified by theoretical properties of point process models the obtained results were significantly better compared to well-known baseline models. In the paper neural ordinary differential equations (Chen, Rubanova, Bettencourt, & Duvenaud, 2018; Zhang et al., 2019) authors presented an interesting approach for modelling models dynamic by deep neural networks. Moreover, the authors presented an interesting example where point process is modelled by differential equation and solved using Euler method. Besides, the authors implemented the backpropagation technique for reducing memory complexity during the training phase.

Intensity-free approaches are based on modelling point process by unsupervised learning techniques (Ghahramani, 2003). Compared to intensity approaches this methods can obtain better results, however, they are more prone to overfitting due to small datasets or large expressive powers of the model. Variational autoencoders (VAE) present unsupervised machine learning algorithms that are mostly used for point process modelling. The Action Point Process variational autoencoder (APP-VAE) presents a variational auto-encoder that can capture the distribution over the times and categories of action sequences (Mehrasa et al., 2019). The APP-VAE obtained state-of-the-art results on the MultiTHUMOS and Breakfast datasets. A declustering based hidden variable model that leads to an efficient inference procedure via a variational autoencoder for solving multivariate highly correlated point process is presented in (Yuan et al., 2020). Besides of VAE, generative adversarial networks (GANs) have recently been proposed as a method for describing event occurrences (Xiao et al., 2017). The authors proposed an intensity-free approach for point processes modelling that transforms nuisance processes to a target one by using Wasserstein GANs. Experiments on various synthetic and real-world data substantiate the superiority of the proposed point process model over conventional ones.

The model presented in this paper belongs to the class of intensity approaches. Compared to the standard intensity approaches our model has more expressive power and is less prone to overfitting compared to intensity-free approaches.

3. POINT PROCESS

In the term point process, the word point is used as a representation of the event on the timeline. Furthermore, it is accepted by the authors to think about the temporal point pattern as an ordered array of times when events occurred. Mutual to such events is that there is no information about how many there will be and when they will happen. Usually, there is some more complex mechanism behind them that explains their nature. To explain this nature and predict future events, it's convenient to use a tool for stochastic process modeling point patterns - a temporal point process.

To describe phenomena over time, the evolutionary character is essential. *Evolutionarity* means that what is happening now only depends on what happened in the past, so the future events don't have any impact on the current state. Accepting evolutionarity, the challenge to describe and predict the temporal point pattern comes down to finding a stochastic model for the time of the next event given the times of previous events. This knowledge of the times of previous events up to but not including current time t is given by history H_{t_n}:

$$H_{t_n} = (t_0, t_1, t_2, ..., t_{n-1}, t_n)$$

One of the possible approaches to define a process is by finding the distribution of time length between subsequent events. This time is also known as *interevent time*. Let $f(t_{n+1} \mid H_{t_n})$ be the *conditional density function* of the time of the next event t_{n+1} given the history of previous events H_{t_n}. Since the density functions $f(t_n \mid H_{t_{n-1}})$ specify the distributions of all interevent times, one by one, and according to the evolutionary character of the process, distribution of all events is given by the joint density (respecting the rule that the joint density for a bivariate random variable can be represented as $p(x, y) = p(x) \cdot p(y \mid x)$):

$$f(t_0, t_1, t_2, ..., t_{n-1}, t_n) = \prod_n f(t_n \mid H_{t_{n-1}})$$

Another popular approach involves a conditional intensity function as a function that defines expectation that an event will occur in the infinitesimal interval around *t* given the history *H* at times before time *t*.

The conditional intensity $\lambda(t)$ is defined to be the expected rate at which events will tend to occur around time *t* given the H_{t_n}:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{E\left(N(t, t+\Delta t) \mid H_{t_n}\right)}{\Delta t}$$

where N denotes the number of events occurred in interval (t+Δt). Previously mentioned papers mostly agreed that the conditional intensity function is a more convenient and intuitive way of specifying how the present depends on the past in an evolutionary point process.

Considering the conditional density $f(t_{n+1} \mid H_{t_n})$ and its corresponding cumulative distribution function $F(t_{n+1} \mid H_{t_n})$ for any $t > t_n$, the conditional intensity function is defined by Liu (2018):

$$\lambda(t_{n+1}) = \frac{f(t_{n+1} \mid H_{t_n})}{1 - F(t_{n+1} \mid H_{t_n})}$$

In other words, the conditional intensity function specifies the mean number of events in a region conditional on the past. It's assumed that there are no points coincide so that there is either zero or one point in an infinitesimal interval.

Proposition 1. A conditional intensity function $\lambda^*(t)$ [1] uniquely defines a point process if it satisfies the following conditions for any point pattern $(t_0, t_1, t_2, ..., t_{n-1}, t_n)$ and any $t < t_n$:

1. $\lambda^*(t)$ is non-negative and integrable on any interval starting at t_n, and
2. $\int_{t_n}^{t} \lambda(s) ds \to \inf \text{ for } t \to \inf$ (Rasmussen, 2011)

Different functional forms have been found helpful in concrete examples of point process modelling. In this paper, we aim to present a novel approach for the incorporation of well-known forms into learning models that describe the temporal point patterns. Four different functional forms are considered: first and second-order polynomial Poisson inhomogeneous processes, Hawkes with exponential and Gaussian kernels.

The form of the homogeneous *Poisson process* describes a completely random process independent of history. Such a conditional intensity function is equal to conditional density function:

$$\lambda(t) = b = const$$

The common way to introduce a dependency to a model is to turn into a polynomial form. First and second-order forms are presented as:

$$\lambda(t) = a + b \cdot t$$

[1] Notation * is used for stating that the function λ depends on the history H_{t_n}

and equivalently

$$\lambda(t) = a + b \cdot t + c \cdot t^2$$

where $a \geq 0,\ b \geq 0 \text{ and } c \geq 0$ are learnable coefficients.

On the other hand, Hawkes suggests the one possible solution for the clustered point process (Hawkes, 1971). A significant number of real phenomena have clustered nature. The process of a car arriving on the highway and the process of a customer arriving at a store, both have such a nature. The clustered nature explains that observation of a few points in the recent past increases the chance that there will be new points soon. In other words, the probability of seeing a new event increases due to previous events.

$$\lambda(t) = \mu + \alpha \cdot \sum_{t<t_i} \exp\left(-\left(t - t_i\right)\right)$$

where $\mu, \alpha \geq 0$.

The Hawkes form emphasizes that each time a new point arrives in this process, the conditional intensity grows by α and then decreases exponentially back towards μ. Hawkes process gives a felicitous form for conditional intensity function that can be used as a foundation for adapting to a specific problem. The standard form is given by:

$$\lambda(t) = \mu + \alpha \cdot \sum_{t_i<t} \varphi\left(t - t_i\right)$$

where φ represents the kernel function that can be replaced to better suit specific problems. The base form of Hawkes uses an exponential kernel. We are considering another form where the kernel is represented as a sum of Gaussian basis functions (Zha & EDU):

$$\lambda(t) = \mu + \alpha \cdot \sum_{t<t_i} \left(2 \cdot \pi \cdot \sigma^2\right)^{-1} \cdot \exp\left(\frac{-t^2}{2 \cdot \sigma^2}\right)$$

3.1. Likelihood function – learning point process

For the observed point pattern $\left(t_0, t_1, t_2, \ldots, t_{n-1}, t_n\right)$ on $\left[0, T\right]$ for some given $T > 0$, a likelihood function is given by:

$$L = \prod_{i=1}^{n} \lambda^*\left(t_i\right) \cdot exp\left(-\Lambda^*\left(T\right)\right)$$

where $\Lambda^*\left(T\right)$ stands for integrated conditional intensity function, given by:

$$\Lambda^*\left(T\right) = \int_0^T \lambda^*\left(s\right) ds$$

The λ^* function should be selected concerning Proposition 1 and the integral is going to be solvable.

In order to fit the parameters of the point process to the observed event data, it is necessary to define a loss function. The loss function can be defined as a negative log-likelihood:

$$\log L = \sum_{i=1}^{N} \log \lambda^*\left(t_i\right) - \int_0^t \lambda^*\left(t\right) dt$$

$$loss = -\log L = -\sum_{i=1}^{N} \log \lambda^*\left(t_i\right) + \int_0^t \lambda^*\left(t\right) dt$$

This provides the right and powerful instrument to be able to optimize the parameters of the defined function λ^* intending to maximize negative loss.

Since the function λ^* can take any functional form, it's expected that many of them will not have an analytic solution of the integral. Although integration would be possible, it is often not feasible due to a computational limitation. For the Hawkes form of the conditional intensity function, given the previous equation, loss have the following form:

$$loss = -\log L = -\sum_{i=1}^{N} \log \lambda^*(t_i) + \int_0^t \lambda^*(t)\,dt = -\sum_{i=1}^{N} \log \lambda^*(t_i) - \left(\mu \cdot t + \alpha \cdot \sum_{t_i < t} \int_0^t \exp(t - t_i)\,dt \right)$$

In the example, the $\Lambda^*(t)$ has an analytical solution, but it quickly tends to fall into the trap of overflowing the memory space of float numbers. Given the behavior of the exponential function and the constraints of float numbers in Python, for each $t > 92$, $\Lambda^*(t)$ will tend to infinity.

In this paper, we offer an alternative solution to avoid the problem explained, whether integration is not possible due to a non-integrable function or due to an overflow problem. We introduce the utilization of a numerical approximation of a one-dimensional integration in a previously defined loss function. There are several widely accepted approximations of solutions of a one-dimensional integral.

In this paper, we are going to present the result of using a Trapezoid and Simpson's rule, Euler method, and Gaussian quadrature rules based on interpolating functions. The trapezoidal rule is a technique for approximating the definite one-dimensional integral. It works by approximating the region under the graph of the function as a trapezoid and calculating its area. Applied to conditional intensity function, it takes the following form:

$$\int_{t_1}^{t_2} \lambda(t)\,dt \approx (t_2 - t_1) \cdot \frac{f(t_1) + f(t_2)}{2}$$

Euler's Method is another technique, which uses the idea of local linearity or linear approximation, where the small tangent lines over a short distance are used to approximate the solution to an initial-value problem. Simpson's rule is the approximation for n+1 values bounding n equally spaced subdivisions. Gaussian quadrature rule is an approximation of the definite integral of a function, usually stated as a weighted sum of function values at specified points within the domain of integration, named after a mathematician Carl Friedrich Gauss.

4. EXPERIMENTAL SETUP

The presented approach was successfully tested on highway toll dataset and four different well known point process models were compared. The sequence of cars arriving at the ramp toll on the E 75 highway was taken as a concrete example of interest. Highway European route E 75 is part of the International E-road network, which is a series of main roads in Europe. The E 75 starts at the town of Vardø in Norway, goes through seven countries and ends at town Sitia in Greece, after about 4,380 kilometres. The observed part connects two large Serbian cities, Belgrade and Niš. More precisely, we decided that the goal of the modelling was the process of arrivals in Niš from the direction of Belgrade on the busiest ramp toll, marked number 3. The average time between two passes in one day is about 20 seconds. Figures 1 and 2 show that there is a correlation between the number of cars and hours of the day, but no significant correlation with a minute of hours.

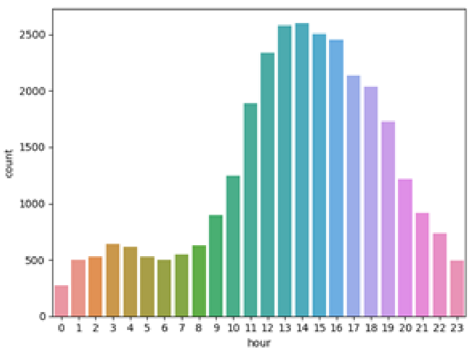

Figure 1: Number of cars that crossed the toll by hour

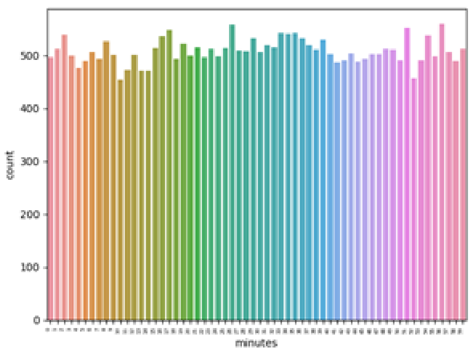

Figure 2: Number of cars that crossed the toll by minute

The implementation of numerical methodology is used for linearizing the neML function to enable backpropagation of neML derivatives. We have trained four different forms of the conditional intensity function with five different numerical approximations of the integral part and an analytical solution where possible. In total, we present the results of 23 models. All models were trained in 50 epochs and with a constant learning rate of 0.001. As an optimization method, Adam stochastic optimization was used. Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. All defined models are implemented in Pytorch (Paszke et al., 2019), an optimized tensor library for deep learning using GPUs and CPUs implemented in Python.

The obtained results are presented in Table 1. The best results are obtained using the Hawkes baseline with the sum of the Gaussians kernel. We hereby confirm that the process of the car arriving at the ramp toll is most likely a clustered point process, so Hawkes' models proved to be significantly better. Using the sum of Gaussians kernels, the dependence on time is emphasized in a sophisticated way, so this leads to the best results. On the other hand, polynomial forms let down, since they don't hold the information about any clustered pattern. In terms of execution time, polynomial models take less time than Hawkes models due to their derivation complexity and the number of parameters being learned.

However, with all models used, we have shown that the functions can be optimized following the given goal.

Table 1. Results

Model name	Integration scheme	Loss
Poisson Polynomial Second Order	Euler	1.03E+13
Poisson Polynomial Second Order	Implicit Euler	1.03E+13
Poisson Polynomial Second Order	Trapezoid	9.55E+12
Poisson Polynomial Second Order	Simpsons	8.77E+12
Poisson Polynomial Second Order	Gaussian_Q	7.99E+12
Poisson Polynomial Second Order	Analytical	8.77E+12
Poisson Polynomial First Order	Euler	2.18E+08
Poisson Polynomial First Order	Implicit Euler	2.18E+08
Poisson Polynomial First Order	Trapezoid	1.87E+08
Poisson Polynomial First Order	Simpsons	1.55E+08
Poisson Polynomial First Order	Gaussian_Q	1.24E+08
Poisson Polynomial First Order	Analytical	1.55E+08
Hawkes	Euler	57141.27
Hawkes	Implicit Euler	58971.43
Hawkes	Trapezoid	55488.9
Hawkes	Simpsons	52010.85
Hawkes	Gaussian_Q	48537.86
Hawkes	Analytical	50361.51
Hawkes Sum Gaussians	Euler	27313.95
HawkesSumGaussians	Implicit Euler	29099.78
HawkesSumGaussians	Trapezoid	26773.55
HawkesSumGaussians	Simpsons	24525.94
HawkesSumGaussians	Gaussian_Q	22345.78

5. CONCLUSION

In this paper, we presented a novel methodology for learning temporal point process based on the implementation of one-dimensional numerical integration techniques. The likelihood function of intensity point process has integral of conditional intensity function given in the limits of data observation. Bearing in mind that conditional intensity function can take any kind of mathematical form, in many cases this integral is analytically intractable. Due to this, in this paper, we presented a possibility to linearize integral with standard numerical techniques and to backpropagate derivative through it. The presented approach was successfully tested on highway toll dataset and four different well known point process models were compared. In addition, we presented that different numerical techniques for integration can be successfully implemented in automatic differentiation package such as Pytorch. Further studies should address using deep neural networks (feed-forward and recurrent networks) as a conditional intensity function to better capture dependencies between event occurrences.

ACKNOWLEDGEMENT

This work was supported in part by the ONR/ONR Global under Grant N62909-19-1-2008. In addition, this research is partially supported by the Ministry of Science, Education and Technological Development of the Republic of Serbia grants TR35004 and TR41008. The authors would like to express gratitude to company

Saga d.o.o Belgrade, for supporting this research. The authors would also like to thank public enterprises Roads of Serbia for providing data for this research.

REFERENCES

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

[2] Chen, T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). *Neural ordinary differential equations.* Paper presented at the Advances in neural information processing systems.

[3] Farajtabar, M., Wang, Y., Gomez-Rodriguez, M., Li, S., Zha, H., & Song, L. (2017). Coevolve: A joint point process model for information diffusion and network evolution. *The Journal of Machine Learning Research, 18*(1), 1305-1353.

[4] Ghahramani, Z. (2003). *Unsupervised learning.* Paper presented at the Summer School on Machine Learning.

[5] Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological), 33*(3), 438-443.

[6] Jacobsen, S., Grove, H., Nedenskov Jensen, K., Sørensen, H. A., Jessen, F., Hollung, K., . . . Søndergaard, I. (2007). Multivariate analysis of 2-DE protein patterns–Practical approaches. *Electrophoresis, 28*(8), 1289-1299.

[7] Jia, R., Jiang, P., Liu, L., Cui, L., & Shi, Y. (2018). Data driven congestion trends prediction of urban transportation. *IEEE Internet of Things Journal, 5*(2), 581-591.

[8] Kirchner, M. (2017). An estimation procedure for the Hawkes process. *Quantitative Finance, 17*(4), 571-595.

[9] Last, G., & Penrose, M. (2017). *Lectures on the Poisson process* (Vol. 7): Cambridge University Press.

[10] Liu, S. (2018). Generalized Mahalanobis Depth in Point Process and Its Application in Neural Coding and Semi-Supervised Learning in Bioinformatics.

[11] Mehrasa, N., Jyothi, A. A., Durand, T., He, J., Sigal, L., & Mori, G. (2019). *A variational auto-encoder model for stochastic point processes.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[12] Mei, H., & Eisner, J. M. (2017). *The neural hawkes process: A neurally self-modulating multivariate point process.* Paper presented at the Advances in Neural Information Processing Systems.

[13] Nguyen, T. T., Krishnakumari, P., Calvert, S. C., Vu, H. L., & Van Lint, H. (2019). Feature extraction and clustering analysis of highway congestion. *Transportation Research Part C: Emerging Technologies, 100*, 238-258.

[14] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Antiga, L. (2019). *PyTorch: An imperative style, high-performance deep learning library.* Paper presented at the Advances in Neural Information Processing Systems.

[15] Rasmussen, J. G. (2011). Temporal point processes: the conditional intensity function. *Lecture Notes, Jan*.

[16] Ryu, B., & Steven, B. L. (1998). Point process models for self-similar network traffic, with applications. *Communications in statistics. Stochastic models, 14*(3), 735-761.

[17] Xiao, S., Farajtabar, M., Ye, X., Yan, J., Song, L., & Zha, H. (2017). *Wasserstein learning of deep generative point process models.* Paper presented at the Advances in Neural Information Processing Systems.

[18] Xiao, S., Yan, J., Yang, X., Zha, H., & Chu, S. M. (2017). *Modeling the intensity function of point process via recurrent neural networks.* Paper presented at the Thirty-First AAAI Conference on Artificial Intelligence.

[19] Yuan, B., Wang, X., Ma, J., Zhou, C., Bertozzi, A. L., & Yang, H. (2020). *Variational Autoencoders for Highly Multivariate Spatial Point Processes Intensities.* Paper presented at the International Conference on Learning Representations.

[20] Zahrieh, D. (2017). Bayesian point process modeling to quantify excess risk in spatial epidemiology: an analysis of stillbirths with a maternal contextual effect.

[21] Zha, H., & EDU, G. Learning Granger Causality for Hawkes Processes Supplementary File.

[22] Zhang, T., Yao, Z., Gholami, A., Gonzalez, J. E., Keutzer, K., Mahoney, M. W., & Biros, G. (2019). *ANODEV2: A Coupled Neural ODE Framework.* Paper presented at the Advances in Neural Information Processing Systems.