Temporal Point Processes: A Survey

Johanna Sommer

Abstract

Numerous applications in healthcare, social media and other areas produce vast amounts of temporal event data. Capturing the behaviour of this sequential data is central to the better understanding of patterns and predicting the future. Temporal point processes offer a mathematical framework to model discrete events in continuous time. In this paper we survey classical as well as recently proposed models, challenges and ongoing research.

1. Introduction

In fields such as healthcare, consumer behaviour and stock market transactions, a great deal of information is hidden in the system's history and the intricate mechanisms that trigger events. In modern medical care, electronic health records contain a sequence of visits to the hospitals, tests, diagnoses and medications. If we could understand and model dependencies between such events and predict future developments, it could enable prevention and patient-specific care. Similarly, understanding the evolution of stocks and purchasing patterns could help businesses make informed decisions in business-critical situations. Other fields of application include seismology, social media interactions and speech recognition.

In such situations, we do not know when in the future or how many events will occur. Hence, the goal is to use information from past events to understand the system and model the causal structure of consecutive events. Temporal point processes offer an elegant way to describe such mechanisms. From that we can retrieve information about the future and use predictions to recommend, intervene or act.

1.1. Theory of Temporal Point Processes

Notation. A temporal point process τ is a stochastic process with discrete points/events on a continuous timeline.[1] We denote the time of an event i as t_i with $t_i \in \mathbb{R}^+$ and $i \in \mathbb{Z}^+$. The knowledge of the times of all events $H_t = \{t_i | t_i \leq t\}$ up to and including time t is called the history. Temporal point processes assume that points do not coincide and are strictly ordered in time. This means the time interval between subsequent events, called interevent time, is strictly positive. We are interested in the conditional probability density function $f(t_{i+1}|H_{t_i})$ which denotes the probability of the next event t_{i+1} occuring in the time interval $[t, t+dt)$ given the history up to time t [14]. This can then be used to sample events from the distribution and perform predictive inference. The inverse method or Ogata's thinning algorithm are commonly used to draw samples [14][10]. The difference between temporal point processes and time series is that we handle time in a continuous way; it is a random variable rather than an index. Temporal point processes act in a different scenario compared to e.g. times series or autoregressive models, as time is handled in a continuous way. It is a random variable rather than an index.

Conditional Intensity Function. When trying to model a temporal point process with a conditional density function, one quickly runs into several problems: there is no intuition when designing the underlying structure with $f(t_i|H_{t_{i-1}})$, it needs to be a valid probability distribution and combining multiple process models is non-trivial [14]. Thus, the conditional density function may not be the best choice. Instead, we use a combination of the conditional density function $f(t_i|H_{t_i})$ and the cumulative distribution function $F(t_i|H_{t_i})$ to characterize the event times of a process by the conditional intensity function (CIF)

$$\lambda^*(t) = \frac{f(t_i|H_{t_i})}{1 - F(t_i|H_{t_i})} \tag{1}$$

for any $t > t_i$.[2] A more intuitive explanation of the conditional intensity function would be that it specifies the rate at which events come in – dependent on the past [1].

Marked TPPs. In the case of marked temporal point processes, we have discrete *marked* events $\{(t_i, \kappa_i)\}$ with $\kappa_i \in \mathbb{M}$ where the domain of the marks $\mathbb{M}$ is specific to the application. A concrete example of a mark would be the places a person visits during the day. The history for a marked TPP includes the time as well as the marks of all previous events. The conditional intensity function is

[1] The terms point and event refer to the same entity and will be used interchangeably.

[2] We denote the conditional intensity with $*$ to make clear that it depends on the history.

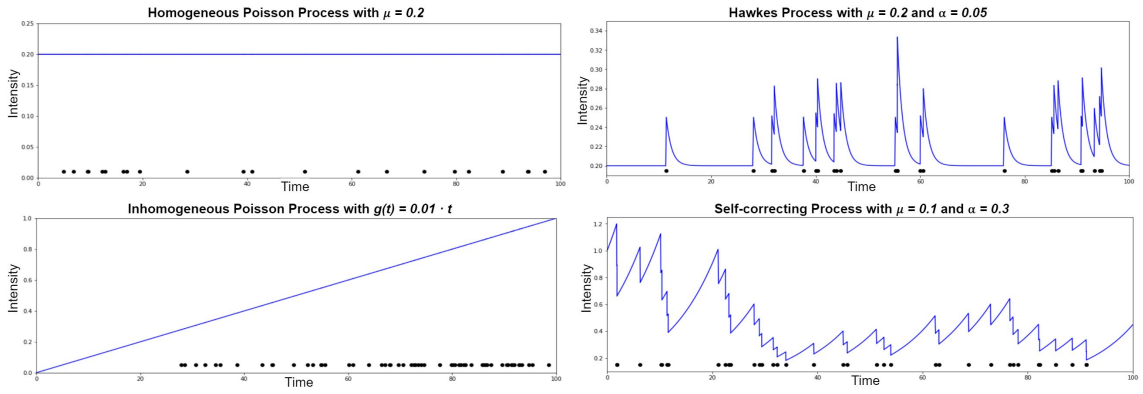

Figure 1. Example CIF and process timings

defined as

$$\lambda^*(t,\kappa) = \lambda^*(t) * f(\kappa|t, H_{t-1}) = \frac{f(t,\kappa|H_{t_i})}{1 - F(t|H_{t_i})} \quad (2)$$

where we call $\lambda^*(t)$ the ground intensity. To evaluate the goodness of fit of the marker distribution, metrics like cross-entropy are used.

Learning. To train the model from past events, we need to learn the parameters of the conditional intensity function. The likelihood function of a temporal point process is defined as

$$L = (\prod_{i=1}^{N} \lambda^*(t_i)) * \exp(-\Lambda^*(T)) \quad (3)$$

where we have a point pattern $t_1, \ldots, t_N$ on an observation interval $[0, T)$ and the integrated conditional intensity function

$$\Lambda^*(T) = \int_0^T \lambda^*(t)dt. \quad (4)$$

In the marked case, we use the same formula but have accompanying marks for each event. With the likelihood function, we can now use, for example, maximum likelihood estimation to estimate the parameters of the CIF. To evaluate the event timing predictions, metrics like the negative log-likelihood or the mean squared error are used.

1.2. Challenges

There are several challenges to modelling a temporal point process. To define the conditional intensity function we need to make assumptions about its form. The different parametrizations encode prior beliefs about the process, but in practice the true model is not known [2]. Those assumptions can restrict its expressive power and the ability to generalize across application fields [12]. Having to pick between those restricted models even for a specific application increases the risk of model misspecification. Additionally, while we have an explicit expression for the likelihood function, it is hardly ever trivial to analytically find the optimal parameters.

In the next section, we will first discuss some basic TPPs and then see how recent work addresses these challenges. In section 4 we will evaluate and discuss the proposed methods and look at open research questions.

2. Types of Temporal Point Processes

While there exists work on application specific models [5][18][13], the focus of this survey lies on temporal point processes that try to model general scenarios.

2.1. Classical Models

We now discuss classical models and their conditional intensity functions. To give intuition about their behaviour, Figure 1 shows examples of the CIF as well as the timings of events.

Poisson process [16][2]. In a Poisson process, the interevent times are i.i.d. random variables and exponentially distributed, as they do not depend on the past. In a homogeneous Poisson process, the conditional intensity is constant $\lambda^*_\mu(t) = \mu \geq 0$ and we learn the parameter μ. In an inhomogeneous Poisson process the conditional intensity $\lambda^*_\theta(t) = g_\theta(t) \geq 0$ varies over time and we learn the parameters θ of the function $g(t)$. In both cases, $\lambda^*(t)$ is independent of the history.

Hawkes process [15][4]. The assumption that the interevent times are independent of the history does not always hold. In reality, we have e.g. contagious processes like the spread of measles or earthquakes, where the arrival of an event causes the conditional intensity to increase. Such processes are called self-exciting and the Hawkes process is a popular example. Its conditional intensity is defined by

$$\lambda^*(t) = \lambda_0(t) + \sum_{t_i \in H_t} \Phi(t - t_i) \quad (5)$$

where $\lambda_0(t)$ is some deterministic baseline intensity with $\lambda_0(t) : \mathbb{R} \to \mathbb{R}_+$ and $\Phi(t)$ is referred to as the memory kernel. Popular choices for this kernel are the exponential kernel $\Phi(t) = \alpha * exp(-\delta t)$ with $\alpha \geq 0$, $\delta > 0$ and $\alpha < \delta$

and the power-law kernel $\Phi(t) = \frac{\alpha}{(t+\delta)^{\eta+1}}$ with $\alpha \geq 0$, $\delta, \eta > 0$ and $\alpha < \eta\delta^{\eta}$. One can observe that the Hawkes process is a specific case of the inhomogeneous Poisson process where the intensity explicitly depends on past events through the memory kernel [15].

Self-correcting process [6][14]. In some cases we want the opposite of a self-exciting process; a process wherein the arrival of a new event causes the probability for another event to decrease. We can not simply change the plus sign before the memory kernel in eq. 5 to a minus as the conditional intensity function has to be non-negative. Instead, we can use functions like

$$\lambda^*(t) = exp(\mu t - \sum_{t_i \in H_t} \alpha). \tag{6}$$

2.2. Advanced Models

Recent work will be compared under the following aspects:
- *how they model the conditional intensity*
- *learning the parameters of the proposed TPP*

2.2.1. Neural Temporal Point Processes

Du et al. [2] treat the conditional intensity as a non-linear function of the history, which is approximated by means of a recurrent cell. They propose using a Recurrent Neural Network **RMTPP** to jointly model the time and marker information. Recall that in RNNs the output of the hidden units at the current time step will be fed in as the input at the next time step. This makes the connection between a recurrent layer and the history H_t clear which allows us to rewrite the conditional density for the next event as $f(t_{i+1}|H_t) = f(t_{i+1}|\mathbf{h}_i)$. The hidden state $\mathbf{h}_i$ is updated after recieving the current input. We can depend on it to model the generation of the markers $f(\kappa|t, H_{t-1})$ with a multinomial distribution and the conditional intensity with

$$\lambda^*(t) = exp(v^{t\top} * \mathbf{h}_i + w^t(t - t_i) + b^t) \tag{7}$$

where v^t is a column vector and w^t, b^t are scalars. Thus, the intensity is a combination of the accumulated influence of past events $v^{t\top} * \mathbf{h}_i$, the influence of the current event $w^t(t-t_i)$ and a base intensity b^t. The loss is then calculated and backpropagated to learn the parameters v^t, w^t and b^t of the process.

In Mei & Eisner's work on the **Neural Hawkes Process**[8], they modify the Hakwes process such that its intensity evolves according to a continuous-time LSTM. Similar to the recurrent cell in the RMTPP, the hidden state $\mathbf{h}(t)$ is a sufficient statistic for the history. The difference to a regular LSTM is that the update does not depend on the previous hidden state but its value $\mathbf{h}(t_i)$ at time t_i after it has *continuously* decayed towards a baseline intensity. This is done to better match the continuity of the random variable t. The interval $(t_i, t_{i+1}]$ ends when the next event occurs and the cell is updated. Fitting the Neural Hawkes Process means learning the parameters of the LSTM by maximizing the likelihood; the intractable integral is handled with Monte Carlo sampling.

Because RMTPP and the Neural Hawkes Process still make assumptions about the functional form of the intensity, Omi et al. propose **FullyNN** [12], an RNN that represents the conditional intensity in a general functional form. They aim to generalize RNN-based approaches by modelling the cumulative intensity function $\Lambda^*(t)$ (see eq. 5) instead of the CIF. The cumulative intensity is then simply the output of the network and the conditional intensity can be derived by differentiation.[3]

2.2.2. Intensity-Free Approaches

Instead of modelling the intensity function, there are approaches that try to avoid the CIF and, thereby, also the major drawbacks that come with it. Shchur et al. [17] propose to model the conditional density instead and do so by (a.) modelling the probability distribution via normalizing flows (**DSFlow**) and (b.) with a mixture model of log-normal distributions (**LogNormMix**). While there has been work on using mixture models for temporal point processes [19][11], the focus there lies on modelling the CIF.

2.2.3. Multivariate Point Processes

A new line of research that has gained traction over the last few years is that on multivariate point processes. In such processes one has to account not only for interactions between events of the same marks but also the interactions between events of different types. An example would be a social network setting, where events are members of a type of process, for example tweets from a certain user. Since in such scenarios one often encounters a large number of users and thereby marks, work in this area focuses on modelling such processes at scale.

Türkmen et al. [20] introduce an RNN to model the mutual excitation between marks, which is combined with a Hawkes process to capture temporal relationships. They scale their approach with a proposal for efficient sampling in multivariate point processes. Similarly, Nickel and Le [9] tackle the problem by paying attention to sparsity, which is often inherent in such processes, as only a small fraction of all possible entries participate in any given sequence. They exploit this sparsity in their likelihood and gradient computation, making for a scalable approach for modeling multivariate point processes.

[3] The constraints imposed on the CIF do not hold for the FullyNN model. A flaw has been pointed out by Shchur et al. [17] where the PDF does not integrate to 1 and the interevent times are not always strictly positive, but this can easily be fixed.

Model	What do they model? How?
RMTPP	$\lambda^*(t)$ with RNN hidden states
Neural H.	$\lambda^*(t)$ with continuous LSTM
Fully NN	$\Lambda^*(t)$ with RNN
DSFlow	$f*(t)$ with Normalizing Flows
LogNormMix	$f*(t)$ with Mixture Model

Table 1. Overview Neural Temporal Point Processes

3. Learning of Temporal Point Processes

Some temporal point process models suffer from an intractable likelihood function where one has to resort to Monte Carlo sampling or limiting the CIF to integrable functions. This can lead to inexact specification of the conditional intensity function.

Guo et al. [3] propose INITIATOR, a framework to tackle the intractable integral. A novel loss function based on noise-contrastive estimation is proposed, where, similar to the RMTPP approach, the likelihood can then be learned by a set of parameters. A similar approach is taken by Xiao et al. [23]. While they also treat this topic as a conditional probability distribution modelling problem, they use a Wasserstein GAN to do so. Here, a network-based generator is used for event prediction, whereas in the work by Guo et al. [3] the generator refers to an explicit parametric point process. They extend upon their WGAN work in [24] by conditioning the generator on the event history. Yan et al. [25] further improve MLE learning by discriminative and adversarial learning with the Wasserstein distance introduced in the WGAN work. The initial model is learned by MLE, then improved by reversely approximating the integral of the conditional intensity in closed form.

Upadhyay et al. [21] as well Li et al. [7] look at the problem of learning the parameters of a point process from a reinforcement learning perspective. In this case, the generation of samples from a TPP is treated as an action taken by a stochastic policy. The goal is to find the optimal policy and mark distribution for the agents actions that maximizes any arbitrary reward function. This is done by iterating between learning the reward function and the stochastic policy.

4. Discussion

Both Poisson processes are memoryless and the events arrive independently. In the homogeneous case, the events arrive at a constant rate, whereas in the inhomogeneous case the rate is goverened by the intensity function. Thus, Poisson is a good choice when the events are independent of each other. However, this is not always the case as there are applications where the arrival of an event increases the likelihood of another event. Such self-exciting processes are often used in social media modelling or seismology [15]. Figure 1 shows that this leads to a more clustered pattern of points. The counterpart is a self-correcting process, which seeks to produce regular events.

Experiments for TPPs usually consist of synthetic and real data. The synthetically constructed data is used to evaluate how good of a fit the intensity or density function is, as the ground truth is known. It is also important to have data from different domains in the experiments, as different fields have different characteristics in their event distributions. Empirical results show that while FullyNN yields consistently good results, its performance is inferior to that of LogNormMix and DSFlow across all domains [17]. The experiments also show that the assumptions that RMTPP and Neural Hawkes make about the functional form of the CIF limits their flexibility and they are outperformed by other approaches. A short summary of all models can be found in Table 1.

Current literature fails to show experimental results for the complexity of the model and the overhead of training. It would be valuable to have standardized results of training and inference runtime - not only for multivariate point processes but also for the models introduced in section 2.2.

The question of which function to model in order to specify a TPP remains - the conditional density or the conditional intensity. Beside accuracy, the CIF is commonly used because of its interpretability and because it is easier to specify, as it has less constraints. While learning with the CIF can pose challenges because of intractable integrals, section 3 has shown approaches to overcome such challenges. However Shchur et al. [17] rightfully point out that with recent recurrent formulations of the CIF, it is not interpretable anymore and the conditional density function is easier to learn.

Still, it would be valuable to restore the interpretability of the conditional intensity function to gain an understanding of the latent mechanisms of a process. While the domains and applications of processes are vastly different, there could still be similarities between them. Does, for example, the spread of information behave similar to the transmission of a disease? Could we measure such similarities? The field of meta learning investigates such questions with the goal of *learning to learn*. It covers any type of learning based on experience with other tasks [22]. With these techniques, we could extend upon existing equality measures for stochastic processes, like the cross-correlation function, by using the experience we have from modelling previous TPPs. This could give suggestions about the functional form (constant, self-exciting, etc.) of an unseen process – ultimately not only boosting performance and restoring interpretability, but enabling knowledge discovery of the similarity of different domains.

5. Conclusion

In this survey we presented state-of-the-art techniques to express temporal point processes: classical models as well as neural TPPs and intensity-free approaches. We extended upon this by discussing possible improvements to the learning process and concluded with an outlook on future research on temporal point processes.

References

[1] De, A., Upadhyay, U., and Gomez-Rodriguez, M. Temporal point processes. Technical report, Saarland University, 2019.

[2] Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. 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*, 2016.

[3] Guo, R., Li, J., and Liu, H. Initiator: Noise-contrastive estimation for marked temporal point process. In *IJCAI*, 2018.

[4] Hawkes, A. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B*, 33, 07 1971.

[5] Huang, H., Wang, H., and Mak, B. Recurrent poisson process unit for speech recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2019.

[6] Isham, V. and Westcott, M. A self-correcting point process. *Stochastic Processes and their Applications*, 8(3):335 – 347, 1979. ISSN 0304-4149.

[7] Li, S., Xiao, S., Zhu, S., Du, N., Xie, Y., and Song, L. Learning temporal point processes via reinforcement learning. In *Advances in neural information processing systems*, 2018.

[8] Mei, H. and Eisner, J. M. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, 2017.

[9] Nickel, M. and Le, M. Learning multivariate hawkes processes at scale, 2020.

[10] Ogata, Y. On lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27 (1):23–31, 1981.

[11] Okawa, M., Iwata, T., Kurashima, T., Tanaka, Y., Toda, H., and Ueda, N. Deep mixture point processes: Spatio-temporal event prediction with rich contextual information. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[12] Omi, T., Ueda, N., and Aihara, K. Fully neural network based model for general temporal point processes. In *Advances in Neural Information Processing Systems 32*. 2019.

[13] Qian, Z., Alaa, A. M., Bellot, A., Rashbass, J., and van der Schaar, M. Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. *arXiv preprint arXiv:2001.02585*, 2020.

[14] Rasmussen, J. G. Lecture notes: Temporal point processes and the conditional intensity function, 2018.

[15] Rizoiu, M.-A., Lee, Y., Mishra, S., and Xie, L. A tutorial on hawkes processes for events in social media. *arXiv preprint arXiv:1708.06401*, 2017.

[16] Ross, S. M., Kelly, J. J., Sullivan, R. J., Perry, W. J., Mercer, D., Davis, R. M., Washburn, T. D., Sager, E. V., Boyce, J. B., and Bristow, V. L. *Stochastic processes*, volume 2. Wiley New York, 1996.

[17] Shchur, O., Biloš, M., and Günnemann, S. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020.

[18] Tabibian, B., Valera, I., Farajtabar, M., Song, L., Schölkopf, B., and Gomez-Rodriguez, M. Distilling information reliability and source trustworthiness from digital traces. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[19] Taddy, M. A., Kottas, A., et al. Mixture modeling for marked poisson processes. *Bayesian Analysis*, 7, 2012.

[20] Türkmen, A. C., Wang, Y., and Smola, A. J. Fastpoint: Scalable deep point processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2019.

[21] Upadhyay, U., De, A., and Rodriguez, M. G. Deep reinforcement learning of marked temporal point processes. In *Advances in Neural Information Processing Systems*, 2018.

[22] Vanschoren, J. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[23] Xiao, S., Farajtabar, M., Ye, X., Yan, J., Song, L., and Zha, H. Wasserstein learning of deep generative point process models. In *Advances in Neural Information Processing Systems*, 2017.

[24] Xiao, S., Xu, H., Yan, J., Farajtabar, M., Yang, X., Song, L., and Zha, H. Learning conditional generative models for temporal point processes. In *AAAI*, 2018.

[25] Yan, J., Liu, X., Shi, L., Li, C., and Zha, H. Improving maximum likelihood estimation of temporal point process via discriminative and adversarial learning. In *IJCAI*, 2018.