Mutually Regressive Point Processes

Ifigeneia Apostolopoulou
Machine Learning Department
Carnegie Mellon University
iapostol@andrew.cmu.edu

Scott Linderman
Department of Statistics
Stanford University
scott.linderman@stanford.edu

Kyle Miller
AutonLab
Carnegie Mellon University
mille856@andrew.cmu.edu

Artur Dubrawski
AutonLab
Carnegie Mellon University
awd@cs.cmu.edu

Many real-world data represent sequences of interdependent events unfolding over time. They can be modeled naturally as realizations of a point process. Despite many potential applications, existing point process models are limited in their ability to capture complex patterns of interaction. Hawkes processes admit many efficient inference algorithms, but are limited to mutually excitatory effects. Nonlinear Hawkes processes allow for more complex influence patterns, but for their estimation it is typically necessary to resort to discrete-time approximations that may yield poor generative models. In this paper, we introduce the first general class of Bayesian point process models extended with a nonlinear component that allows both excitatory and inhibitory relationships in continuous time. We derive a fully Bayesian inference algorithm for these processes using Po´lya-Gamma augmentation and Poisson thinning. We evaluate the proposed model on single and multi-neuronal spike train recordings. Results demonstrate that the proposed model, unlike existing point process models, can generate biologically-plausible spike trains, while still achieving competitive predictive likelihoods.

1 Introduction

Many natural phenomena and practical applications involve asynchronous and irregular events such as social media dynamics, neuronal activity, or high frequency financial markets [1, 2, 3, 4, 5, 6, 7]. Modeling correlations between events of various types may reveal informative patterns, help predict next occurrences, or guide interventions to trigger or prevent future events. *Point Processes* [8] are models for the distribution of sequences of events.

Cox processes or *doubly stochastic processes* [9] are generalizations of *Poisson Processes* [10], where the intensity function is a stochastic process itself. Although there are efficient inference algorithms for some of their variants [11, 12], Cox processes do not capture explicitly temporal correlations between historical and future events. On the other hand, the *Hawkes Process (HP)* [13, 14] and its variants [15, 16, 17] constitute a class of point process models where past events linearly combine to increase the probability of future events. However, purely excitatory effects are incapable of characterizing physiological patterns such as neuronal activity where inhibitory effects are present and crucial for self-regulation [18, 19, 20]. The work in [21] can support temporal effects beyond mutual excitation that HP misses. However, capturing model uncertainty is critical in many applications [22, 23, 24, 25], especially when the size of the available data is limited compared to the model complexity. Rich literature exists on HP-based learning tasks [26, 27, 28, 29, 30, 31].

A nonlinear generalization of the HP allows for both excitatory and inhibitory interactions, but evaluating the probability density of these models requires computing integrated intensity, which is

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

generally intractable. Instead, we are forced to use discrete time approximations, which reduce to a *Poisson Generalized Linear Model (Poisson-GLM)* [32, 3], making learning of these models from data very efficient. However, the estimated regression coefficients may vary widely depending on the boundaries chosen for aggregation [33]. Empirical evidence suggests that while suitable for one-step predictions, such models may suffer stochastic instability and yield non-physical predictions [34].

There is currently limited statistical theory for point process models that support complex temporal interactions in a continuous-time regime. To this end, we develop the first class of Bayesian point process models—*Mutually Regressive Point Processes (MR-PP)*—that allow for nonlinear temporal interactions while still admitting an efficient, fully-Bayesian inference algorithm in continuous time.

2 Proposed Model

2.1 Problem statement

We are interested in learning distributions over event sequences (point processes). These distributions are mutually regressive in the sense that past event occurrences can influence the future realization of the process in an arbitrary manner. A Point Process PP($\lambda(t)$) is characterized by an intensity function $\lambda(t)$, so that in an infinitesimally wide interval $[t, t + dt]$, the probability of the arrival of a new event is $\lambda(t)dt$ [35].

2.2 Classical Hawkes Process

A Hawkes process (HP) [13, 14] of N event types HP$_N(\lambda^*_n(t))$ is characterized by the intensity functions $\lambda^*_n(t)$ for the events of type n defined as:

$$\lambda_n(t) = \lambda_n + \sum_{m=1}^{N} \sum_{i=1}^{K_m} \lambda_{m,n}(t, t_i^m) \mathrm{I}(t_i^m < t), \tag{1}$$

$$\lambda_{m,n}(t, t_i^m) = a_{m,n}\, e^{-\delta_{m,n}(t - t_i^m)}, \tag{2}$$

where $\lambda_n \geq 0$, $a_{m,n} \geq 0$, and $\delta_{m,n} > 0$. t_i^m is the arrival time of the i-th event of type m and K_m is the number of events of type m. I is the indicator function. By the superposition theorem for Poisson processes, the additive terms in Equation (1) can be viewed as the superposition of independent *non-homogeneous* Poisson processes (with intensity function that varies in time) characterized by the intensity functions $\lambda_{m,n}(t, t_i^m)$, triggered by the event i-th of type m that occurred before time t, and an exogenous, *homogeneous* Poisson process characterized by the constant intensity function λ^*_n. The HP is a *mutually exciting* point process in the sense that past events can only raise the probability of arrival of future events of the same or different type. Since $\lambda^*_n(t)$ depends on past occurrences, it is a stochastic process itself.

2.3 Mutually Regressive Point Process: a generalization of the Hawkes Process

The intensity function $\lambda_n(t)$, for events of type n occurring at times t_i^n, of a Mutually Regressive Point Process (MR-PP) is a HP intensity augmented with a probability term. It is defined as follows:

$$\lambda_n(t) = \lambda^*_n(t) p_n(t), \tag{3}$$

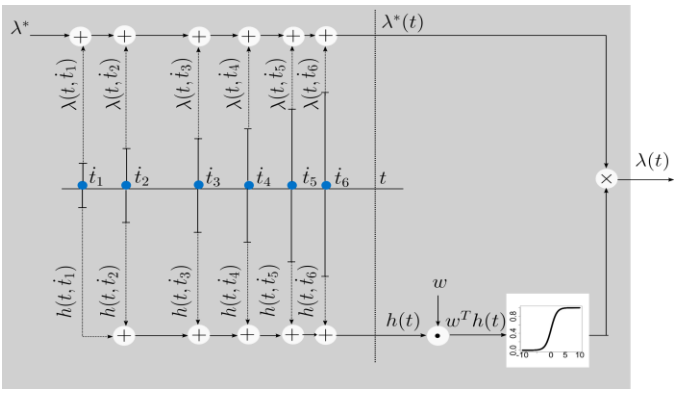
$$\lambda_n(t) = \lambda_n + \sum_{m=1}^{N} \sum_{i=1}^{K_m} \lambda_{m,n}(t, t_i^m) \mathrm{I}(t_i^m < t), \tag{4}$$

$$p_n(t) = o(\boldsymbol{w}_n^T \boldsymbol{h}(t)), \tag{5}$$

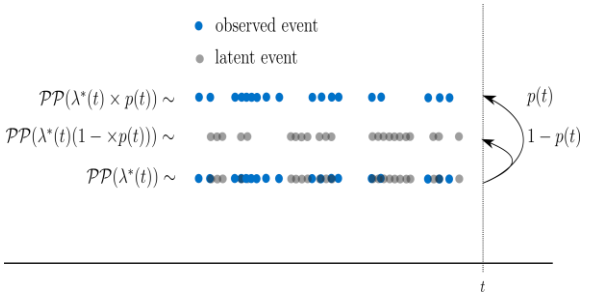
$$h_m(t) = c \sum_{i=1}^{K_m} h(t, t_i^m) \mathrm{I}(t_m^i < t), \tag{6}$$

$$h(t, t_i^n) = e^{-\gamma(t - t_i^m)}, \tag{7}$$

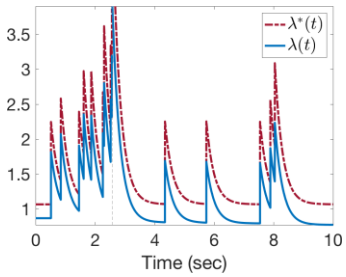
where $\lambda^*_n \geq 0$, $c > 0$, $\gamma > 0$, $\boldsymbol{w}_n = [b_n, w_{1,n}, w_{2,n}, \ldots, w_{N,n}]^T$, $\boldsymbol{h}(t) = [1, h_1(t), h_2(t), \ldots, h_N(t)]^T$ and $\lambda_{m,n}(t, t_i^m)$ defined in Equation (2). $o(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. The weight $w_{m,n}$ models the influence of type m on type n and $h_m(t)$ is the



(a) Computational flow of the MR-PP intensity function described in Equations 3-7



(b) Poisson Thinning



(c) HP vs MR-PP intensity function

Figure 1: *Explanation of the MR-PP.* The computation of the intensity function of a MR-PP at time t as a function of the past events is explained in Figure 1a. Figure 1b shows the simulation of a MR-PP which can be viewed as classification of events generated by a HP as either latent or observed. The point processes of the observed and thinned events are characterized by the $\lambda^*(t)$ intensity multiplied by the probability term $p(t)$ and $1 - p(t)$ respectively. The upper-bounding, mutually exciting, intensity and the thinned intensity $\lambda(t) = \lambda^*(t) \times p(t)$ which generates the observed events is shown in 1c.

aggregated temporal influence of type m up to time t. The computational procedure is illustrated in Figure 1a. The effect of the probability term on the upper-bounding intensity $\lambda^*_n(t)$ is demonstrated in Figure 1c. We can simulate from this model via Poisson thinning [36, 11]. First, we sample N sets of events $t_1^n, t_2^n, \ldots,$ for $n = 1, 2, \ldots, N$, from a HP $_N(\lambda_n(t))$. Afterwards, we chronologically proceed through the simulated events and accept them with probability $\lambda_n(t)/\lambda^*_n(t) = p_n(t)$, the relative intensity at that point in time (Figure 1b). In case an event at t_i^n is rejected, its offsprings (events generated by $\lambda_{n,m}(t, t^n_i)$) are pruned so that the $\lambda^*_n(t)$ defined in Equation (4) depends only on the realized events whose arrival times are notated as t_i^m. Importantly, the relative intensity $p_n(t)$ and the intensity $\lambda^*_n(t)$ only depend on the preceding events that were *accepted*; rejected events have no influence on the future intensity. Note that a negative weight $w_{m,n}$ means that events of type m inhibit future events of type n since $h_m(t)$ decreases $p_n(t)$. The correctness of this procedure is provided in the Supplementary Material.

Although $\lambda^*_n(t)$ could be replaced by a homogeneous Poisson intensity λ^*_n so that any exciting relationships are captured by a positive weight $w_{m,n}$, the upper bound λ^*_n should be given a very large value in cases where the underlying process exhibits sparse event bursts. This fact, in turn, could yield a large number of latent events and hence render the learning of the model computationally intractable (see Section 3.1 for details). Moreover, MR-PP is not hardwired to exponential kernels. Alternative kernel functions, such as the Power-Law or the Rayleigh function could be used in Equations (2) and (7).

2.4 Hierarchical MR-PP for relational constraints

A dependence between the parameters of the intensity $\lambda^*_n(t)$ and the thinning procedure $p_n(t)$ can be imposed so that an interaction between types m and n is either inhibitory or excitatory (but not both) in a probabilistic manner. To this end, we define a *Sparse Normal-Gamma* prior for the weights,

which fosters an inverse relationship between the excitatory effect $a_{m,n}$ and the repulsive effect $w_{m,n}$ of type m on type n. It is motivated by the framework of *Sparse Bayesian Learning* [37, 38], in the sense that it associates an individual precision $\tau_{m,n}$ and mean $\mu_{m,n}$ with each weight $w_{m,n}$. $\mu_{m,n}$ and $\tau_{m,n}$ follow a Normal-Gamma distribution that depends on $a_{m,n}$. It is defined as follows:

$$\tau_{m,n} \sim \text{Gamma}(\nu_\tau \varphi_\tau(a_{m,n}) + a_\tau, \beta_\tau), \tag{8}$$

$$\mu_{m,n} \sim \text{N}\left(-(\nu_\mu \varphi_\mu(a_{m,n}) + a_\mu)^{-1}, (\lambda_\mu \tau_{m,n})^{-1}\right), \tag{9}$$

$$\boldsymbol{w}_n \sim \text{N}\ \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n\ , \tag{10}$$

where $\nu_\tau > 0$, $a_\tau > 0$, $\beta_\tau > 0$, $\nu_\mu > 0$, $a_\mu \geq 0$, $\lambda_\mu > 0$, $\boldsymbol{\mu}_n = [\mu_0, \mu_{1,n}, \mu_{2,n}, \ldots, \mu_{N,n}]^T$, $\boldsymbol{\tau}_n = [1/\sigma_0^2, \tau_{1,n}, \tau_{2,n}, \ldots, \tau_{N,n}]^T$, and $\boldsymbol{\Sigma}_n = diag(\boldsymbol{\tau}_n)^{-1}$. $\varphi_\tau(x)$ and $\varphi_\mu(x)$ are monotonically increasing positive activation functions.

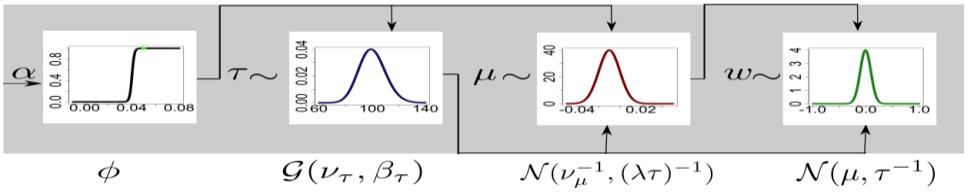
A suggested activation function for $\tau_{m,n}$ and $\mu_{m,n}$ is a shifted and scaled sigmoid function which has a soft-thresholding effect:

$$\varphi(a_{m,n}) = \frac{1}{1 + e^{-\delta_0(a_{m,n} - a_0)}}. \tag{11}$$

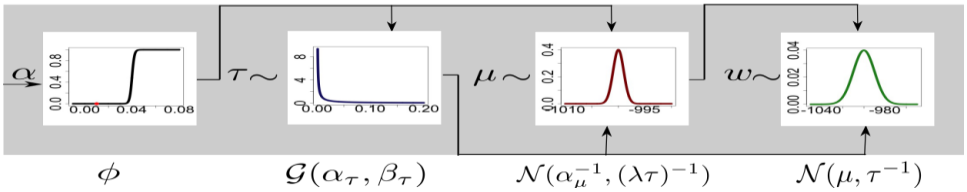
$a_0 > 0$ can be viewed as the excitation threshold (so that values of $a_{m,n}$ above a_0 indicate an excitatory relationship) and $\delta_0 > 0$ regulates the smoothness of the thresholding.

Note that when $a_{m,n}$ is large (there is excitatory relationship from type m on type n), the precision $\tau_{m,n}$ becomes large (approximately drawn from $\text{Gamma}(\nu_\tau, \beta_\tau)$) assuming $\nu_\tau >> a_\tau$ and $\nu_\tau >> \beta_\tau$. Therefore, the variance $\tau_{m,n}^{-1}$ has a value close to zero with high probability. A similar scenario holds for $\mu_{m,n}$ if $\nu_\mu >> a_\mu$. A small mean and variance for $w_{m,n}$ implies that any additional (possibly inhibitory) effect of type m on type n is suppressed. A numerical example is given in Figure 2a. On the other hand, when $a_{m,n}$ is small, the precision of $\tau_{m,n}$ will take a small value approximately drawn from $\text{Gamma}(a_\tau, \beta_\tau)$ (assuming that $\nu_\tau \varphi_\tau(a_{m,n}) << a_\tau$ and $a_\tau < \beta_\tau$). Similarly, $\mu_{m,n}$ can take large negative values coming approximately from a Normal distribution with mean $-a_\mu^{-1}$. As a consequence, inhibitory effects from type m on type n are enabled. A numerical example is given in Figure 2b.

Due to the inverse relationship between the inhibitory coefficients $w_{m,n}$ and the endogenous intensity rates $a_{m,n}$, relational constraints on pairs of types are established. Intuitively, the constants ν_τ, ν_μ control the strength of these constraints, so that $w_{m,n}$ is close to zero for a large $a_{m,n}$ with an adjustable probability. A traditional Hawkes process can be obtained by setting ν_τ, ν_μ λ_μ, a_τ, a_μ, μ_0 and τ_0 to a very large value.



(a) Hierarchical prior for an excitatory relationship



(b) Hierarchical prior for an inhibitory relationship

Figure 2: *Illustration of the behavior of the hierarchical prior for enforcing relational constraints*. In 2a the excitatory coefficient a is above the threshold value (0.05) indicating an excitatotry relationship. The prior drives the weights to a value close to zero. In 2b the coefficient is below the threshold indicating an inhibitory relationship. The prior steers the weights to a large negative value. The parameters of the hierarchical prior were set as follows: $\nu_\tau = 100$, $a_\tau = 0.01$, $\beta_\tau = 1$, $a_\mu = 0.001$, $\nu_\mu = 100$, $\lambda_\mu = 100$.

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/976105054144010041