CHAPTER NINE

Selective review of penalized learning methods for event processes

Myrto Limnios[a] **and Niels R. Hansen**
University of Copenhagen, Department of Mathematical Sciences, Copenhagen, Denmark

9.1. Introduction

Temporal event processes are ubiquitous in many applications, with their instantaneous evolution being dependent on past events. Such processes are traditionally modeled as point processes. One early important model was introduced in seismology for analyzing the temporal propagation of earthquakes, see, e.g., (Ogata, 1988), and more recently, a series of works consider applications in mathematical finance (Bacry et al., 2015), neuroscience (Paninski, 2004), social science (Park et al., 2021; Mohler et al., 2011), etc.

One can consider temporal event processes through their time-stamps only, which results in *temporal point processes* or, equivalently, counting processes. It is also possible to consider temporal event processes, where each event is associated with a mark characterizing the event. Marks can be categorical, which results in *multivariate point processes*, but they can also be continuous, encoding, e.g., earthquake magnitudes. Marks can, furthermore, be multivariate (spatial) giving *spatio-temporal point processes*.

Seminal works developed probabilistic frameworks, allowing for, e.g., self-exciting models, which can provide a mechanistic description of the temporal dynamics of the underlying system. We refer for instance to Hawkes (1971a) for Hawkes processes, and more generally to the classic books (Daley and Vere-Jones, 1998; Brémaud, 2020; Jacobsen, 2005). However, classic statistical procedures for estimating parameters in a model suffer from the intrinsic memory of the system, which results in complex forms of empirical loss functions, which are difficult to compute and optimize. For instance, the likelihood function will generally not have an

[a] Present affiliation: Institute of Mathematics, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Stochastic Modeling and Statistical Methods
https://doi.org/10.1016/B978-0-44-331694-4.00014-1

analytically tractable expression, and it may require numerical integration over multivariate sets. Moreover, to express complicated dependencies, the models need to be flexible, which often leads to a high-dimensional parameter, which we need to estimate from data.

In the present chapter, we focus on penalized statistical methods tailored for dealing with both high-dimensional parameters and for being computationally attractive, even for large datasets and complex models. Some of these solutions involve replacing the log-likelihood with another loss function, the quadratic contrast, which can sometimes be computationally and theoretically attractive.

We highlight that our review is not exhaustive, and rather intends to provide a general perspective of the existing literature. For additional details, we thus refer to some of the excellent existing reviews, specific to particular models, e.g., to González et al. (2016) for environmental and epidemiological models; Renner et al. (2015) for models describing species distribution with point processes dynamics; Rommel et al. (2022) for data-augmentation approaches specific to electroencephalogram signals (EEG), or statistical methods; Banerjee et al. (2014) for a summary of statistics applied to purely spatio-(temporal) point processes; Reinhart (2018) for a general review, and Rommel et al. (2022) reviewed data augmentation approaches based on deep learning algorithms and testing of their efficiency with a systematic approach.

The chapter is organized as follows: Section 9.2 states the general framework and exposes prototypical examples of temporal point processes. Section 9.3 describes the general framework for statistical methods based on penalized loss functions first, then reviews the proposed methods in the literature for two specific cases: likelihood-based and least-squares-based losses.

9.2. Event processes: definitions, fundamental properties, and prototypical examples

This section introduces the notations used throughout the chapter, exposes general types of evolutionary point processes considered as *event processes*, of which prototypical examples are then described.

Notation

For any nonempty set A, we denote the corresponding indicator function by 1_A and its closure by $\overline{A}$. For the half-line $\mathbb{R}_+ = [0, \infty)$, the corresponding

Borel σ-algebra is denoted by $\mathcal{B}_+$. We denote by $L_2(I)$, with I a compact set, the Hilbert space of functions $f : I \to \mathbb{R}$ equipped with the norm $\|f\|_I^2 = \int_I f(t)^2 \mathrm{d}t$. The positive part of $x \in \mathbb{R}$ is denoted by $(x)_+ = \max\{x, 0\}$.

9.2.1 General framework

We introduce in this section the general framework of marked temporal point processes, which is used throughout the chapter, and that forms the foundation for the statistical methods reviewed in Section 9.3.

Let $(\Omega, \mathcal{F}, \boldsymbol{\mathcal{F}}, \mathbb{P})$[1] denote an abstract probability space equipped with a right-continuous filtration $\boldsymbol{\mathcal{F}} = (\mathcal{F}_t)_{t \geq 0}$. Let $(K, \mathcal{K}, \eta)$ be a Borel measure space equipped with a probability measure η. We refer to Protter (1992), Section I.1 therein, for basic definitions and properties. We will consider stochastic processes defined on Ω, which represent event times, and we will use elements in K to denote the type of an event. The space K is called the mark space to signify that elements in K are marks of the event types.

A *marked temporal point process* (MTPP) X is defined on Ω and takes values in the set of counting measures on the space $(\mathbb{R}_+ \times K, \mathcal{B}_+ \otimes \mathcal{K})$. That is, for any set $B \in \mathcal{B}_+ \otimes \mathcal{K}$, the random variable $X(B)$ takes values in $\{0, 1, 2, \ldots, \infty\}$, and it counts the number of points in B. For any set $A \in \mathcal{K}$ and all $t \in \mathbb{R}_+$, we introduce the counting process $N_t(A) = X([0, t] \times A)$, which counts the number of events up to time t and with marks in A. We let $N_t = N_t(K) = X([0, t] \times K)$ denote the counting process of all events up to time t, and we say that N is nonexploding if $N_t < \infty$ for all $t \in \mathbb{R}_+$. Note that if N is nonexploding, then $N_t(A) < \infty$ for all A. We will throughout assume that N is nonexploding and that the counting processes $N_t(A)$ for $A \in \mathcal{K}$ are $\boldsymbol{\mathcal{F}}$-adapted.

The homogeneous Poisson random measure is a special MTPP, which will serve as a baseline for the construction of other MTPPs. If X is a homogeneous Poisson random measure, there is a $\lambda^0 > 0$ such that the counting processes $N_t(A)$ are homogeneous Poisson counting processes with intensity $\lambda^0 \eta(A)$ with regard to the abstract filtration $\boldsymbol{\mathcal{F}}$ for all $A \in \mathcal{K}$. Since η is a probability measure, N_t is then a homogeneous Poisson process with intensity λ^0, and it is, in particular, nonexploding. For any $t > s \geq 0$, the increment $N_t(A) - N_s(A)$ is Poisson distributed with mean $\lambda^0(t - s)\eta(A)$ and independent of $\mathcal{F}_s$. Moreover, for any disjoint sets $A_1, \ldots, A_r$ of $\mathcal{K}$, the counting processes $N_t(A_1), \ldots, N_t(A_r)$ are independent.

[1] We suppose Ω to be nonempty and $\mathcal{F}$ to be a σ-algebra composed of the subsets of Ω.

We will throughout assume that there is a probability measure $\mathbb{Q}$ on Ω and a random counting measure X such that X is a homogeneous Poisson random measure under $\mathbb{Q}$. By the following definition and theorem, we will be able to define other distributions of X by a change-of-measure.

Definition 9.1. Let X be a homogeneous Poisson random measure on $(\mathbb{R}_+ \times K, \mathcal{B}_+ \otimes \mathcal{K})$ defined on the filtered probability space $(\Omega, \mathcal{F}, \boldsymbol{\mathcal{F}}, \mathbb{Q})$. Let $t \mapsto \lambda_t(x)$ denote a nonnegative, locally finite, and $\mathcal{F}_t$-predictable process for all $x \in K$ and define the likelihood process:

$$L_t = \exp\left\{ \int_K \int_0^t \log \frac{\lambda_u(x)}{\lambda^0} X(\mathrm{d}u, \mathrm{d}x) - \int_K \int_0^t (\lambda_u(x) - \lambda^0) \mathrm{d}u \, \eta(\mathrm{d}x) \right\} . \quad (9.1)$$

The condition that $\lambda_t(x)$ is locally finite, meaning that $\sup_{s \leq t, x \in K} \lambda_s(x) < \infty$, implies that the second integral in (9.1) is finite and that the likelihood process L_t is well defined. We see from the definition that L_t is a nonnegative and $\mathcal{F}_t$-adapted process, and that $L_t > 0$, unless X has a point mass in a point (u, x) with $\lambda_u(x) = 0$. Generally, the process L_t is a local martingale and a supermartingale with regard to $\mathcal{F}_t$ so that

$$\mathbb{E}_{\mathbb{Q}}[L_t \mid \mathcal{F}_s] \leq L_s .$$

Since $L_0 = 1$, this implies that $\mathbb{E}_{\mathbb{Q}}[L_t] \leq 1$. If the process is a martingale, the theorem that follows allows us to construct another distribution of the random counting measure X by a change-of-measure.

Theorem 9.1. *With the same setup as in Definition 9.1, and if the likelihood process L_t is a martingale, we can define a measure $\mathbb{P}_t$ on $(\Omega, \mathcal{F})$ such that the likelihood process is the Radon–Nikodym derivative:*

$$\frac{\mathrm{d}\mathbb{P}_t}{\mathrm{d}\mathbb{Q}} = L_t . \quad (9.2)$$

We refer to Jacobsen (2005) for the sequential construction of *canonical* MTPPs, which lead to Corollary 5.1.2 in Jacobsen (2005), which gives a constructive proof of the likelihood process as a Radon–Nikodym derivative showing that it is a martingale. For MTPPs defined on an abstract filtered probability space, and with a finite K, see Theorem 2.4 and Corollary 2.3 in Sokol and Hansen (2015) for general conditions ensuring that the likelihood process is a martingale. Example 4.3 in Sokol and Hansen (2015) gives that

$$\lambda_t(x) \leq \alpha + \beta N_{t-} , \quad (9.3)$$

for some constants $\alpha, \beta \geq 0$, and all $x \in K$ is a sufficient condition when K is finite. Since (9.3) also implies that $\lambda_t(x)$ is locally finite, condition (9.3) guarantees that for any $t \geq 0$ there exists a probability measure $\mathbb{P}_t$ with the likelihood process L_t as Radon–Nikodym derivative with respect to $\mathbb{Q}$.

When L_t is a martingale, the probability measures $\mathbb{P}_t$ are consistent in the sense that $(X_s)_{s\in[0,t]}$ has the same distribution under $\mathbb{P}_t$ as under $\mathbb{P}_T$ for any $T \geq t$. We will later fix a maximal time horizon $T > 0$ and consider the observation of X only within the time window $[0, T]$ and under $\mathbb{P}_T$. It is nontrivial to extend the family of measures to a T-independent measure on an abstract space, but to ease notation, we will usually drop the subscript T and just denote the measure by $\mathbb{P}$. The distribution of X under $\mathbb{P}$ is then determined by the predictable process λ. Thus to define a statistical model with particular properties, we choose λ, possibly parameterized by $\theta \in \Theta$ for a parameter space Θ, to yield these desired properties.

To construct models in terms of λ, we need to be able to interpret and understand λ. We will usually take $\lambda^0 = 1$, in which case it holds that

$$\mathbb{P}[X(\mathrm{d}t \times \mathrm{d}x) = 1 \mid \mathcal{F}_{t-}] \approx \lambda_t(x)\, \mathrm{d}t\, \eta(\mathrm{d}x)\ . \tag{9.4}$$

That is, conditionally on everything up to just before time t, $\lambda_t(x)$ is the conditional rate of a new event. We refer to $\lambda_t(x)$ as the *intensity*. We can further decompose $\lambda_t(x)$ by introducing the ground intensity

$$\lambda_{g,t} = \int \lambda_t(x)\eta(\mathrm{d}x)\ , \tag{9.5}$$

and then in terms of the ground intensity, we define

$$f(x \mid t) = \frac{\lambda_t(x)}{\lambda_{g,t}}\ , \tag{9.6}$$

for $\lambda_{g,t} > 0$. The ground intensity has an interpretation similar to (9.4), that is, as the conditional rate of an event with any mark type: $\mathbb{P}[N_{t+\mathrm{d}t} - N_{t-} = 1 \mid \mathcal{F}_{t-}] \approx \lambda_{g,t}\, \mathrm{d}t$. The function $f(x \mid t)$ is then the conditional density with respect to η of the mark given an event at time t and the history $\mathcal{F}_{t-}$ up to just before time t.

Models defined in terms of an $\mathcal{F}_t$-predictable intensity λ are sometimes referred to as a *causal* description of the system, in the sense that the model description at the present time t only depends on the past. We note that this notion of being causal differs from other notions of *causality*, see, e.g.,

(Peters et al., 2017), which should also reflect what happens when we intervene in the system. See also Mogensen et al. (2018); Mogensen and Hansen (2020).

The intensity process has, in addition to (9.4), another interpretation. Define the process $\Lambda_t(A) := \int_A \int_0^t \lambda_s(x)\mathrm{d}s\eta(\mathrm{d}x)$ for $A \in \mathcal{K}$. Under minimal assumptions on λ, namely that it is a nonnegative and locally bounded[2] $\mathcal{F}_t$-predictable process, the process $M_t(A) = N_t(A) - \Lambda_t(A)$ is a local $\mathcal{F}_t$-martingale. The process $\Lambda_t(A)$ is known as the *compensator* of the process $N_t(A)$, and $N_t(A) = M_t(A) + \Lambda_t(A)$ is the *Doob–Meyer decomposition* of the submartingale $N_t(A)$ as a sum of a (local) martingale and an increasing predictable process.

With a given intensity model determining $\mathbb{P}$ and an observation of the random counting measure X on $[0, T]$ under $\mathbb{P}$, we can compute the log-likelihood as $\log(L_T)$. It is worth writing out the explicit formula for the log-likelihood in the corollary that follows; see also Proposition 7.3.III in Daley and Vere-Jones (1998).

Corollary 9.1. *Consider the MTPP X with intensity process λ and observed on the interval $[0, T]$ for some fixed and finite $T > 0$. Let $(t_1, x_1), \ldots, (t_n, x_n)$ denote the finite observed points for X with event times $0 < t_1 < t_2 < \ldots < t_n < T$, and the associated marks $x_1, \ldots, x_n \in K$. Then the log-likelihood process at time T equals*

$$\log L_T = \sum_{i=1}^{n} \log \frac{\lambda_{g,t_i}}{\lambda^0} - \int_0^T (\lambda_{g,u} - \lambda^0)\mathrm{d}u + \sum_{i=1}^{n} \log f(x_i \mid t_i) \ . \tag{9.7}$$

Note how the log-likelihood in (9.7) is decomposed into the first two terms pertaining only to the ground intensity and the event times and the last term being effectively a conditional log-likelihood of the marks given the event times.

Canonical probability space

MTPPs are often studied on a so-called canonical probability space (the space of counting measures, say) with the internal filtration generated by the process X itself. See Jacobsen (2005) for a detailed treatment of this perspective. One main benefit of this approach is that the way that λ can depend on the history becomes very explicit.

[2] An $\mathcal{F}_t$-adapted is said to be locally bounded if there exists a sequence of $\mathcal{F}_t$-adapted increasing stopping times, such that the stopped process is bounded.

Suppose that we have observed the points $(t_1, x_1), \ldots, (t_{n-1}, x_{n-1})$ in the time interval $[0, t_{n-1}]$, we can then write the conditional survival function of the waiting time $\tau_n = t_n - t_{n-1}$ from t_{n-1} to the next event as

$$S_n(s) = \exp\left(-\int_0^s h_n(u)\mathrm{d}u\right) ,$$

where h_n denotes the conditional hazard function. From this, we identify the ground intensity as $\lambda_{g,t} = h_n(t) = h_n(t \mid (t_1, x_1), \ldots, (t_{n-1}, x_{n-1}))$. Combined with (9.6), the intensity can be written as $\lambda_t(x) = h_1(t)f_1(x \mid t)$ for $0 < t \leq t_1$, and for $i \geq 2$ and $t_{i-1} < t \leq t_i$ as

$$\lambda_t(x) = h_i(t \mid (t_1, x_1), \ldots, (t_{i-1}, x_{i-1}))f_i(x \mid (t_1, x_1), \ldots, (t_{i-1}, x_{i-1}), t) \ . \quad (9.8)$$

Though the above representation of the intensity is fairly explicit in terms of how the conditional hazard functions and conditional mark distributions should depend on the observed history, the mere computation of the log-likelihood can be a practical challenge. The sequential conditionings, when i ranges in $\{1, \ldots, n\}$, results in an *intractable* form of L_T, limiting its direct maximization. Strategic choices of parameterization for the joint distribution of (t, x), resulting in elegant models; see, e.g., (Schoenberg, 2013), can circumvent some limitations related to the high dimensionality of the process, possibly with long-range memory. We will, however, not focus on those works, and rather discuss methods alleviating those intrinsic characteristics in the dedicated Section 9.3. In the paragraph that follows, we briefly discuss the case of abstract filtration, and how it can encode dependencies on external covariates in particular.

Dependence on exogeneous covariates

In many applications, the observed X over $[0, T]$ depends on auxiliary covariates, which we would like to include in a statistical model. We can model such a dependence by assuming that such covariate processes are adapted to the abstract filtration so that the intensity process can depend on them.

Example 9.1. (*Explicit form of dependence.*) Let Y be a *covariate* process defined on $(\Omega, \mathcal{F}, \boldsymbol{\mathcal{F}}, \mathbb{Q})$, with values in $\mathbb{R}^d$, and predictable with regard to the filtration $\boldsymbol{\mathcal{F}}$. We suppose that X is a homogeneous Poisson random measure under $\mathbb{Q}$, and that Y is independent of X under $\mathbb{Q}$. If we define the intensity process $\lambda_t(x)(\omega) = \mu(t, x, Y_t(\omega))$ for a continuous mapping $\mu : \mathbb{R}_+ \times K \times \mathbb{R}^d \to \mathbb{R}_+$, Theorem 9.1 applies and X depends, under $\mathbb{P}$, on

Y through its intensity. In this particular example, X will be an inhomogeneous Poisson counting measure conditionally on Y.

In general, there can be some technicalities arising from considering external dependencies, related to the $\mathcal{F}$-measurability/predictability of the process Y in particular. We will not go into the details, and refer to Daley and Vere-Jones (1998) page 236, Brémaud (2020) Chapter 5.1, and to Christgau et al. (2023) for resulting limitations in the context of statistical modeling. In the following section, we expose prototypical examples of (marked) point processes, which will be extensively used in Section 9.3.

Now that the general framework is stated, the following section presents a series of prototypical examples of temporal point processes defined through their intensity process only.

9.2.2 Examples of event processes

This section presents fundamental examples of point processes from the simplest form, i.e., homogeneous Poisson processes, to specific classes of MTPP. We implicitly consider processes indexed by time $t \in [0, T]$, with $T > 0$ finite, to avoid additional technicalities. We define the processes via their associated model for the intensity that is generally formulated as a transformation of a predictable process, and parameterized by a deterministic set of square-integrable functions. Precisely, without any loss of generality, consider a counting process N_t with associated intensity process λ_t with regard to an abstract filtration $(\mathcal{F}_t)_{t\geq 0}$ such that the framework in Section 9.2.1 holds true. Based on a Hilbert class of real-valued basis functions $\mathcal{H}$, we present explicit classical models by the relation

$$\lambda_t = \varphi_t(h) \ , \tag{9.9}$$

where $h \in \mathcal{H}$, and φ_t is the predictable transformation characterizing the time dependent stochastic structure of the process. As we shall see in Section 9.3, the goal is to estimate the function h based on an observed process on $t \in [0, T]$.

Poisson processes

Poisson processes are the simplest models for random counting processes, where the distribution of the jumps are specified by a rate being constant if homogeneous, or a deterministic function of time otherwise. The intensity process is deterministic and equals

$$\lambda_t = h(t) \ , \tag{9.10}$$

with $\mathcal{H} \subseteq L_2([0, T], \mathbb{R}_+)$. Their fundamental property of having independent jumps when conditioned on the past makes this class a building block for any other MTPP. We refer to dedicated analysis in Brémaud (2020), Chapters 2 and 3 therein, among others.

Hawkes processes

Introduced in Hawkes (1971a,b), Hawkes processes (HPs) are fundamental models of temporal point processes, which have been used extensively to model evolutionary phenomena with memory. By allowing past occurrences to have inhibitory or excitatory effects on the future events, this class achieves great modeling flexibility. The first works modeled epidemic-type aftershock sequence (ETAS) of earthquakes, see, e.g., (Ogata, 1988), more recently being used to model portfolio dynamics in finance markets (Bacry et al., 2015), or neuronal networks (Reynaud-Bouret et al., 2013; Lambert et al., 2018) for instance. Consider $\psi : \mathbb{R} \mapsto \mathbb{R}_+$ and a kernel $h : \mathbb{R}_+ \mapsto \mathbb{R}$, then the general formulation of the intensity process for Hawkes processes is explicitly given by

$$\lambda_t = \psi \left(\mu_0 + \int_{t-T}^{t-} h(t-u) \, \mathrm{d}N_u \right) , \tag{9.11}$$

where $\mu_0 \in \mathbb{R}$ defines the *spontaneous rate*, representing the value of the process at the origin, and often chosen to be equal to zero. If the *interaction function* h is nonnegative, then the resulting process is self-exciting, and it is self-inhibiting otherwise. The simple choice of $\psi(u) = u$ generates linear Hawkes process, whereas typical examples are $\psi(u) = (u)_+$, accounting for possibly inhibitory effects, and $\psi(u) = \exp(u)$ ensures positivity. We refer to Brémaud and Massoulié (1996) for stability results for nonlinear Hawkes processes under Lipschitz-type conditions for ψ mainly, generalizing the results of Hawkes and Oakes (1974) established in the linear case. If now one considers multivariate HPs with $d \in \mathbb{N}^*$ coordinates, then for any $j \leq d$,

$$\lambda_t^{(j)} = \psi^{(j)} \left(\mu_0^{(j)} + \sum_{i=1}^{d} \int_{t-T}^{t-} h_i^{(j)}(t-u) \, \mathrm{d}N_u^{(i)} \right) , \tag{9.12}$$

where the kernels $h_i^{(j)}$ model the transfer from the ith coordinate $N^{(i)}$ to $N^{(j)}$. Notice that the upperbound of the integral $(t-)$ ensures predictability of the intensity, that could alternatively be obtained by fixing all kernels $h_i^{(j)}(0) = 0$. The associated class of kernel functions can be chosen to be $\mathcal{H} = (\mathbb{R} \times L_2([0, T], \mathbb{R})^d)^d$, with associated squared norm

$\|h\|_{\mathcal{H}}^2 = \sum_j (\mu_0^{(j)})^2 + \sum_j \sum_i \int_0^T (h_i^{(j)})^2$, with $i, j \leq d$, see, e.g., (Hansen et al., 2015). HPs can alternatively be defined via a system of stochastic equations. For instance, Bacry and Muzy (2014) proposed an estimation method for multivariate linear HPs associated with marks, as being the solution of a linear discretized scheme of the Wiener–Hopf system of equations, characterizing up to their second-order. Corollary 1 therein proves that if the considered process has stationary increments and some conditions, then it is uniquely defined by its first- and second-order statistics, ensuring thus that the proposed system has a unique solution.

Temporal point processes

We now define the general setting of temporal point processes (TPP). Following the pattern set by Section 9.2.1, consider X to be a homogeneous Poisson random measure on $(\mathbb{R}_+, \mathcal{B}_+)$, defined on the filtered abstract probability space $(\Omega, \mathcal{F}, \boldsymbol{\mathcal{F}}, \mathbb{P})$. Then by defining the likelihood process L_t of any process λ_t fulfills the conditions of Definition 9.1 to be

$$L_t = \exp\left\{ \int_0^t \log \frac{\lambda_u(x)}{\lambda^0} X(\mathrm{d}u) - \int_0^t (\lambda_u(x) - \lambda^0) \mathrm{d}u \right\} . \tag{9.13}$$

If it is a martingale, we can construct $\mathbb{P}_t$ such that L_t is the change-of-measure $L_t = \mathrm{d}\mathbb{P}_t / \mathrm{d}\mathbb{Q}$ by Theorem 9.1. Let $n \in \mathbb{N}^*$ be finite, and consider the strictly increasing time-stamps over $(0, T)$ under $\mathbb{P}$: $0 < t_1 < t_2 < \ldots < t_n < T$. The likelihood process then is

$$L_T = \prod_{i=1}^n \left(\frac{\lambda_{t_i}}{\lambda^0} \right) \exp\left\{ - \int_0^T (\lambda_u - \lambda^0) \mathrm{d}u \right\} . \tag{9.14}$$

We refer to Proposition 7.2.III in Daley and Vere-Jones (1998), for the proof using the explicit Janossy density functions.

Multivariate and spatial TPP

The framework introduced in Section 9.2.1 encompasses many important models for MTPP. If the set of marks is of the form $K = \{1, \ldots, d\}$, with $d \in \mathbb{N}^*$ finite, $d \geq 2$, and variation independent, then X can model a multivariate TPP, for which the marks represent the coordinates of X. Such processes can be represented by (finite) graphs, with the set of nodes being K, and the edges possibly oriented, which importantly encode the interactions between coordinates. The related adjacency matrix is defined by $A = (A_{i,j})_{1 \leq i,j \leq d}$, with $A_{i,j} = 1$ if there is an edge between node i and j,

and $A_{i,j} = 0$. It is a key tool for the methods in Section 9.3 for summarizing the structure of the graph. If K is a subset of a Euclidean space, the marks can indicate the spatial location visited at the event time, defining spatio-temporal point processes, whereas many models restrict $K \subseteq \mathbb{R}^2$ for geospatial models. Lastly, the marks can indicate a weight coefficient $K \subseteq \mathbb{R}_+$, attributed to the associated event time, or can be multivariate for factorial MTPP.

In the next section, we expose the main statistical methods proposed in the literature to approximate and estimate (M)TPPs. We focus on two different methods formulated as a penalized loss functional of the intensity process mainly. Section 9.3.2 is devoted to losses related to the likelihood process as derived in Theorem 9.1, while Section 9.3.3 focuses on quadratic contrast functions.

9.3. Approximation models and estimation methods

This section presents the most common approximation models and estimation methods for event processes, specifically formulated as solution of a penalized loss functional. We distinguish two main risk functions, being either the likelihood functional or the least-squares functional. We consider the general framework and related assumptions exposed in Section 9.2.1 to hold true.

9.3.1 Generic formulation and empirical optimal estimators

Let the filtered probability space be $(\Omega, \mathcal{F}, \mathbb{P})$, and suppose the time-interval $[0, T]$ to be finite, during which we observe the process X_t. As motivated in Section 9.2.2, we focus on presenting approximation models for the true intensity process λ_t of X_t, adapted to the filtration $\mathcal{F}_t$, with associated estimation procedures. Following Section 9.2.1, let Θ be a parameter space, and consider a statistical model $(\lambda(\theta))_{\theta\in\Theta}$ approximating the true intensity λ, defined on $(\Omega, \mathcal{F}, \mathbb{P})$, with associated probability measure $(\mathbb{P}_\theta)_{\theta\in\Theta}$. Then X has intensity $\lambda(\theta)$ under $\mathbb{P}_\theta$ on $[0, T]$; see Corollary 2.3 in Sokol and Hansen (2015). Choose $\ell_t : \Theta \to \mathbb{R}_+$ to denote a smooth loss functional, which is supposed to admit at least one local minimizer. The general form of the penalized loss function to minimize is formulated by the penalized risk

$$\theta \in \Theta \mapsto \ell_T(\theta) + \pi(\rho, \theta) \ , \tag{9.15}$$

where $\pi : (0,\infty)^q \times \Theta \to \mathbb{R}_+$, $q \in \mathbb{N}^*$, is the penalization function, being usually a linear combination of ℓ_p-norms, with $p \in \mathbb{N}$, of positive weights $\rho \in (0,\infty)^q$ being fixed or data-driven. Typical choices are $p \in \{0,1,2\}$, recovering namely lasso ($p=1$), Tikhonov ($p=2$), and elastic net (linear combination of the two) penalizations, also considered as regularizations for some; we refer to Hastie et al. (2001) for a dedicated analysis. Suppose that there exists at least an optimal minimizer θ^* that we want to estimate, and possibly control the expected risk defined for any θ by

$$\mathcal{E}_T(\theta) = \mathbb{E}_\theta[\ell_T(\theta)] - \ell_T(\theta^*) \ , \tag{9.16}$$

which we suppose to be nonnegative, where $\mathbb{E}_\theta$ denotes the expectation with regard to $\mathbb{P}_\theta$. In practice, however, the approximation models under $\mathbb{P}_\theta$ are unknown, and the goal here is to minimize the penalized risk (9.15) based on an observed random sample $(X_t)_{t\in[0,T]}$. Consider the sequence of strictly increasing event times $\{t_1,\dots,t_n\}$, possibly with associated marks $\{x_1,\dots,x_n\}$. Notice that, because the process is nonexploding, there is at most a finite number of events occurring within a finite time-interval, with probability one. Letting $\Theta_0 \subseteq \Theta$ be fixed, we define an optimal empirical minimizer of the empirical counterpart of the penalized risk, when it exists, by

$$\theta_n \in \operatorname*{arg\,min}_{\theta\in\Theta_0} \ \ell_T(\theta) + \pi(\rho,\theta) \ . \tag{9.17}$$

Some reviewed methods consider having observed an i.i.d. sample drawn from X denoted by $\{X_{t,i},\ i \le m\}$ with $m \in \mathbb{N}^*$. Notice that, depending on the application, the parameters T or m are supposed to be *large*, or the aggregated process to contain enough information, insofar as enough occurrences should be observed to be able to model the memory effect of the process. In addition, when the penalty function is well calibrated on the data sample and the model class, it can be related to adaptive models and model selection procedures; see Massart and Picard (2007), Chapter 7 therein, in the context of density estimation. In the present framework, those guarantees are established for the least-squares estimator (LSE) mainly, and by means of concentration inequalities resulting in (minimax and) oracle generalization bounds of the empirical solution (9.17). As we will see, those recent results provide explicit choices of empirical criteria for penalty functions, while enjoying minimal assumptions on the regularity of the true model λ_t, however they come at a highly technical cost.

We specifically explain the risk function for each of the methods in the sections that follow, and particularly emphasize the choices for the approximation models for the intensity process. In the paragraph that follows, we present some typical choices of nonparametric function classes used in the literature.

Examples of estimators

We briefly outline some prototypical examples of Hilbert spaces $\mathcal{H}$, which are commonly used to model the intensity process for instance, and for which θ represents the weights and possibly additional parameters appearing in those estimators. The simplest class, particularly used for modeling Poisson and counting processes, is that of histograms:

$$\mathcal{H} = \left\{ h, \quad h : t \mapsto \sum_{k \in I} \theta_k 1_{t \in k}, \quad (\theta_k)_{k \in I} \in \Theta \right\} , \tag{9.18}$$

with I being a set of disjoint intervals partitioning $[0, T]$, right-closed (i.e., of the form $(a, b]$), and $\Theta \subset \mathbb{R}$. By counting the number of events occurring in those bins, it drastically simplifies the derivations of the losses. An important related class is the Haar family of intervals, composed of basis functions of the form $x \mapsto 2^{j/2}(1_{0 \leq 2^j x - l < 1/2} - 1_{1/2 \leq 2^j x - l < 1})$ for all $l \in \mathbb{Z}$, $j \in \mathbb{N}$, and $x \mapsto 1_{l \leq x < l+1}$ for $j = -1$, $k = (l, j)$. For multivariate TPP, the θ's appearing in the linear decomposition depend on all couples (i, j)'s of coordinates. For example, the interaction kernels in the linear MHP of Eq. (9.12), are usually written as $h_i^{(j)} = \sum_{k \in I} \theta_{i,k}^{(j)} g_{i,k}^{(j)}$, with the g's being typically exponential or power-law functions, and for modeling the fixed jth coordinate. Such strategic choices can reduce the computational complexity of minimizing Eq. (9.17), in particular when chosen to be common to all pairs of coordinates (i, j), cf. Section 9.3.3. It is thus interesting to choose a group-norm penalty function inducing the same characteristic across the I's basis functions, as will be seen in the sequel.

In the next sections, we will present the two main methods studied in the literature for learning the optimal model, either by minimizing the penalized likelihood-based risk function, by sequentially estimating the conditional p.d.f. or intensity processes, or by minimizing a penalized quadratic loss function when specifying a nonparametric approximation model for the intensity process. To avoid additional technicalities, we formulate the loss functions for temporal point processes only, and refer to the general structure of MTPPs in Section 9.2.1.

9.3.2 Likelihood-based loss functions

We first recall the general form of the likelihood process, which incidentally highlights inherent computational limitations, resulting from its intrinsic evolutionary structure.

Formulation of likelihood loss processes

Let X be a TPP observed on the finite time-interval $[0, T]$, and with intensity process λ_t defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Without loss of generality, consider the probability measure under which X_t is a counting process, say N_t for clarity, compatible with the abstract probability space, and the generic filtration $(\mathcal{F}_t)_{t\geq 0}$. Then the maximum likelihood estimator (MLE) of (9.7) is equivalently the minimizer of

$$\ell_t(\theta) := -\frac{1}{t}\int_0^t \log \lambda_u(\theta) \mathrm{d}N_u + \frac{1}{t}\int_0^t \lambda_u(\theta)\mathrm{d}u \,, \tag{9.19}$$

where we set $\lambda^0 = 1$ for simplicity, and the normalization factor $1/t$ is to comply with the asymptotic regime. In the case of MTPP, we refer to Section 9.2.1, and the extension follows from Eq. (9.6). Suppose the number of events $n \in \mathbb{N}^*$ of random nondecreasing times $(t_i)_{i\leq n}$ in $[0, T)$. Following the sequential decomposition of the intensity process in Eq. (9.8), a natural approach for fitting a statistical model relies on proposing an adequate model for the conditional kernels of, equivalently, the intensity processes, the p.d.f.'s, or the hazard functions for increasing t_i's (equivalently the τ_i's). Table 9.1 gathers typical transmission functions used in the literature. This direct approach, however, results in an intractable minimization scheme converging at least with rate $\mathcal{O}(n^2)$, as the log-likelihood (Eq. (9.7)) has a finite sum of the logarithms of conditional intensities, themselves depending on the past occurrences. For small n, it is possible to numerically evaluate and estimate θ_n. We refer to classic references directly treating Eq. (9.19), such as Brillinger (2004); Brillinger and Segundo (1979); Ozaki (1979). For linear approximation models of the intensity process, classic results have proved the existence and unicity of the MLE; see Kutoyants (1984); Ogata (1978) for theoretical analysis, and, e.g., Ogata et al. (1993), for parametric modeling resulting in a closed-form with faster computational rate $\mathcal{O}(n)$, which is applied to earthquakes activity. We now present some recent contributions proposed in the literature, and refer to additional methodological reviews, such as Reinhart (2018) for classical approaches, and to Yan (2019) for machine learning based methods. We will emphasize the choice of penalization function π, and the algorithmic complexity of the related practical implementations.

Table 9.1 Typical examples of basis functions to model the log-transformation of the survival conditional function S and the corresponding conditional hazard function h. The two event times are defined by t_i and t_j such that $t_j > t_i$, and the rate of transmission being $\alpha_{i,j} > 0$.

Model	Log survival function	Hazard function
Exponential	$-\alpha_{i,j}(t_j - t_i)$	$\alpha_{i,j}$
Power Law	$-\alpha_{i,j}\log(t_j - t_i)$	$\alpha_{i,j}/(t_j - t_i)$
Rayleigh	$-\alpha_{i,j}(t_j - t_i)^2/2$	$\alpha_{i,j}(t_j - t_i)$

Approximation models and estimation of the intensity process

We start by presenting recent works estimating a nonparametric model for the intensity process, particularly motivated for providing faster algorithmic procedures, to allow for efficient model fitting when using large datasets (i.e., large n, m, d). For instance, multivariate Hawkes processes typically require estimating d^2 interaction kernels (see Eq. (9.12)), whereas the likelihood minimization is quadratic in n in addition. Lemonnier and Vayatis (2014) achieves faster rates of order $\mathcal{O}(nd^2)$ by linearly decomposing the real-valued kernels onto a linear exponential basis $h_{i,I}^{(j)}(t) = \sum_{k\in I} \beta_{i,k}^{(j)} \exp(-k\alpha t)$, for $k \in I$ being the degree of decomposition, the coefficients $\beta_{i,k}^{(j)} \in \mathbb{R}$ encoding the intensity of the jumps, while considering a constant rate α for all kernels, and $\varphi(u) = (u)_+$. They proposed a concave approximation of the log-likelihood (Proposition 3), for which the resulting approximation functions valued at the optimal parameters converge in sup-norm towards the true functions μ_0, h's (Proposition 1) with explicit polynomial rate $\mathcal{O}(|I|^{-r})$ if the $r \in \mathbb{N}^*$th derivative functions are continuous, with $|I|$ being the number of basis functions, and with geometric rate $\mathcal{O}(\exp(-|I|))$ if the functions are analytic. The gradients of the Hessian of the loss can be analytically computed, yielding a Newton-based optimization sequential approach, leveraging the Markovian property of the kernels. The implemented unweigthed L_2-roughness penalization of the form $\alpha \int_0^T h'^2$ empirically shows some limitations for estimating power-laws with fast decay around 0. When considering structural assumptions of the underlying data, a series of methods achieve faster rates. Lemonnier et al. (2017) extended Lemonnier and Vayatis (2014) by assuming a low-rank structure induced by clusters in the associated graph. They used the concept of self-concordant barriers circumventing the constrained loss functions (see Nesterov et al. (1994)), and by projecting the d-dimensional HP into a smaller r-dimensional space characterizing the number of *event*

groups. The obtained method achieves an algorithmic complexity of order $\mathcal{O}(n|I|r(\Delta + r))$, with Δ the maximum node degree of the underlying graph. We also refer to Zhou et al. (2013) for sparse low-rank network applied to learning social infectivity networks, and to Du et al. (2015) for weighted nuclear norm penalization for instance. Lastly, Liu et al. (2018) proposed an extension for multivariate HP, which integrates prior spatial structural knowledge encoded into a connection matrix, of coordinates measuring the similarity between two spatial locations. The functions h's are linearly decomposed over the exponential basis with fixed transmission rates. The elastic net regularization with an additional weighted ℓ_2-norm forces the alignment between the interaction and the connection matrices, and a third weighted ℓ_2-norm ensuring both to be of low rank. The Lagrangian formulation of $R_{t,n}$ is then optimized by the alternating direction method of multipliers (Boyd et al., 2011). For theoretical results on nonparametric linear models of λ_t, cf. Eq. (9.19), Hansen (2013) studied general classes of basis filters valued in a generic Banach space. The author considered a finite number of events n, and quadratic π. For Sobolev spaces in particular, e.g., reproducing kernel Hilbert spaces, if φ is continuously differentiable, then the gradients of ℓ_t exist and can be numerically computed; see Proposition 3.6 therein. Later, Hansen (2015) proved the almost identical estimated intensity model when based on either Sobolev kernels or B-spline basis classes, with quadratic penalization, and ρ being chosen to minimize the Takeuchi's information criterion (Claeskens and Hjort, 2008).

Bayesian inference. When considering a reference probability measure on the parameter space Θ, a series of works maximize the posterior intensity model, implemented with an expected maximization (EM) algorithm usually. Briefly, let Π be a chosen prior distribution on Θ, then the posterior distribution for any subset $\Theta_0 \subseteq \Theta$ is given by $\Pi(\Theta_0) = \int_{\Theta_0} L_T(\theta)\mathrm{d}\Pi(\theta) / \int_{\Theta} L_T(\theta)\mathrm{d}\Pi(\theta)$ conditioned on the observed process X. However, these procedures inevitably result in higher computational complexity due to the additional integrals with regard to Π, known as *doubly intractable*. They are derived for specific parametric models of processes, for which we expose some recent advances here, and further refer to Reinhart (2018). For instance, Veen and Schoenberg (2008) introduced a latent variable indicating offspring events for modeling earthquakes (epidemic) propagation in seismology. Nonlinear HPs with real-valued h's were studied by Sulem et al. (2024, 2023) in particular. They proved concentration rates for the convergence of the posterior distribution to the true parameter

measured with regard to the L_1-distance, and with explicit rate of the radius $\varepsilon_T = o(1)$ of order $\log^3 T = \mathcal{O}(T\varepsilon_T^2)$. The necessary assumptions to obtain a consistent recovering of A, rely on typical entropic on Θ_0, and depend on the estimation scenario. Xu et al. (2016) proposed a linear decomposition of the interaction functions h over the Gaussian kernel density estimators, with data-driven bandwidth Silverman (1986). Then, the Fourier transformation of λ_t yields the optimal decomposition by estimating the cut-off frequency, such that the residual error is controlled at a fixed level. The penalization π function of the adjacency matrix A ensures the following: local independence with ℓ_2-norm, temporal sparsity with its ℓ_1-norm, and pairwise similarity measure for events belonging in the same cluster given by $A \mapsto \sum_{j \leq d} \sum_{j' \in \mathcal{C}_j} \|A_{j\cdot} - A_{j'\cdot}\|_F^2 + \|A_{\cdot j} - A_{\cdot j'}\|_F^2$, where $\mathcal{C}_j$ is the cluster of the jth coordinate, numerically optimized by an iterative EM algorithm. We further refer to Lewis and Mohler (2011); Zhou et al. (2020); Yuan et al. (2019); Salehi et al. (2019) for instance, and references therein, for non-parametric inference for Poisson and Hawkes processes with an EM-type procedure.

Sequential minimization of the loss for multivariate TPP

In the context of large data analysis for network learning, a substantial series of works deployed algorithms tailoring specific applications, wherein the idea lies in sequentially learning the MLE by recurrent or feed-forward algorithms, depending on the choice of kernel functions, the embedding maps between two events, and the optimization algorithm.

Cascade modeling for network learning. Myers and Leskovec (2010) introduced a recurrent network procedure for multivariate TPP, defined by constant parameters for all density functions between two coordinates (i, j), to be exponential, power-law, or Weibull distributions. To ensure a convex objective function, the penalization π is the ℓ_1-norm of $(\rho, A_{i,j}) \mapsto \rho \exp\{-(1 - A_{i,j})\}$, with A being the adjacency matrix. The authors proposed a convex relaxation achieving polynomial time algorithm, which recovers the global optimal estimator. They did not allow for self-contamination of the coordinates, and considered having access to the sequence of event-times, while not of their origin that encode hidden events. Gomez-Rodriguez et al. (2012) proposed `NetInf` algorithm, finding the optimal weighted directed spanning tree as a representation of a multivariate TPP, wherein cascades define the hidden layers composed of all the nodes of the tree. The contagion spread over the tree is parameterized by power or exponential laws, and only depends on the time-interval since

the previous event, with constant rate $\alpha > 0$ for all nodes. The procedure aims to learn the graph maximizing the log-likelihood loss using submodularity, instead of considering all possible propagation trees. It is penalized to ensure sparsity of the network, with a fixed number of edges in the tree, which is used as a stopping rule for the greedy approach when starting from the empty graph. Gomez-Rodriguez et al. (2011) extended those models by learning transmission rates between the layers as well, and for all transmission models gathered in Table 9.1 with weighted ℓ_1-norm penalization. The model only allows pairwise time-dependent transmissions between two nodes; it ignores dependence between unknown covariates, while assuming the event times to be independent conditionally on the past events. Theorem 3 proves the consistency of the MLE for the unconstrained loss for the three different transmission models. When considering the loss as function of the conditional survival process S_n valued at the intertimes τ_i's, Du et al. (2012) proposed a linear model for the hazard rate of the form $h_{i,j} = \sum_{k\in I} \theta_{i,j}^k g(\delta_k, t_j - t_i)$, where $(\delta_k)_{k\in I}$ is a uniform grid of the observational interval $[0, T]$; g is a kernel function such that the associated matrix of embeddings $(g(\delta_l, \delta_s))_{l,s\in I}$ is positive definite. The weighted group-lasso penalization function selects a few number of groups having interactions: $\pi(\rho, \theta) = \rho(\sum_i \|\theta_{i,j}\|)^2$, such that $\theta_{i,j} \geq 0$, for all $j \leq n$. The authors used Gaussian kernels to obtain a closed-form of the survival process, although not restricted to it, as one can use numerical integration schemes. By noticing that $\pi(\rho, \theta) \leq \rho \sum_i \|\theta_{i,j}\|^2/\gamma_j$, with $\gamma_j \geq 0$ and $\sum_j \gamma_j = 1$, the resulting loss is convex in θ. A learning algorithm based on block coordinate descent method alternating between the parameters involved is implemented.

Neural point processes. Neural networks (NN), termed as *diffusion networks* here, are characterized by their deep learning architecture. They are reference algorithms for learning an approximation model for massive datasets, possibly valued in high-dimensional spaces, with competitive generalization performances. Recurrent and feedforward NN are mainly used as a natural algorithmic architecture to estimate the MLE, when the risk Eq. (9.7) is formulated in terms of the conditional p.d.f.'s. Starting from t_0, the node/neuron of the network updates at each next time t_i the value of the loss by computing the probability of an event conditionally on the past. The neurons characterize the *hidden layers* of the network, defined by the high-dimensional feature real-valued vector y_i, $i \leq n+2$, where the first and last layers, respectively, characterize the input and output. At the current time t_i, the hidden layer is a (nonlinear) function of the past y_{i-1} and of the *input layer* encoding the information observed at time t_i, and possibly the mark x_i. Once the current value of the hidden layer is computed,

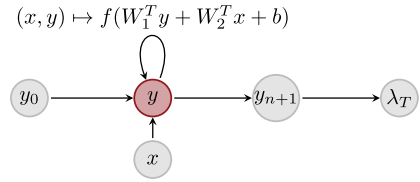

Figure 9.1 Recurrent neural network architecture, with the input observation y_0, the hidden layers represented by the y's, computed via the embedding $f(x, y)$, including possible marks x, the output layer being y_{n+1}, and final mapping to obtain the predicted intensity process valued at $t = T$. The coefficients $\theta = (W_1, W_2, b)$ are sequentially optimized by gradient descent solution of the log-likelihood process and back-propagated through time.

an activation function $f(u)$ (e.g., softmax function, typically $\mathbb{R} \to [0, 1]$) maps it to a normalized value to be interpreted as a probability, which is then embedded to y_{i+1}. Typical choices of embeddings are nonlinear $(y, x) \mapsto f(W_1^T y + W_2^T x + b)$, with the weight matrices W's and bias vector b being learned, and $\theta = (W_1, W_2, b)$, $y = y_{i-1}$, and possibly $x = x_i$. All the weights connecting the layers are the MLE solution of Eq. (9.17), obtained by stochastic gradient descent, and back-propagated through time (BPTT) to minimize the prediction error. Notice that the weights play an important role, insofar as they tailor the possible inhibitory or excitatory effects of the past events on the current event. The estimated NN can be used to predict the value of the modeled $\lambda_t(\theta_n)$ at the final point $t \in (t_n, T]$, based on a new data sample. Multiple models have been proposed; they differ through the choices of the activation functions, characteristics of the weights, and optimization algorithm in particular. For a thorough review on neural TPP, we refer to the excellent works of Shchur et al. (2021) for a general overview, and to Rommel et al. (2022) when applied to EEG data in particular. We summarize the procedure in Fig. 9.1. Du et al. (2016) proposed a recurrent NN for MTPP, for fixed number of discrete marks, and motivated for learning the underlying graph (DAG) in particular. The hidden layer y_i is then computed as in Fig. 9.1, wherein the marks are generated according to a multinomial distribution, conditioned on the current hidden layer; and $f(u) = u_+$. The conditional intensity is of the form $\lambda_t(\theta) = \exp(\theta_1^T y_i + \theta_2(t - t_i) + \theta_3)$, for all $t \in [t_i, t_{i+1})$. The network is estimated by a truncated BPTT procedure, based on a loss inducing a sparse model for the marks, but dense with regard to time-occurrences.

Recurrent models, however, suffer from the long memory of processes through the sequential nonlinear computations during training. It numerically results in either very large effects or in vanishing gradients after many

events. A common work-around is by either truncating the network that provides an approximation of the gradients, or by using long short-time memory NN models (LSTM), see, e.g., (Hochreiter and Schmidhuber, 1997). For instance, Mei and Eisner (2017) proposed a LSTM NN to approximate multivariate HPs of real-valued μ_0's and h's, decomposed over the exponential basis functions, with rates depending on the layers and coordinates. The estimated function is then mapped using a scaled softplus function of the form $f : \theta \mapsto s\log(1 + \exp(\lambda_t(\theta)/s))$, with $s > 0$, which approximates the ReLU function when $s \to 0$, and models nonlinear intensity processes (Eq. (9.12)). It still requires to approximate the integrals using a Monte-Carlo procedure. Mei and Eisner (2017) extended the architecture to continuous-time by including hidden memory cells between two consecutive events, and of exponential decay, in the estimation of the hidden states (γ_i's). See the references therein for related models. Wu et al. (2018) proposed a model for approximating factorial MTPP, wherein ℓ_t is formulated in terms of the conditional p.d.f.'s, and $\pi(\rho, \theta)$ is a linear combination of the ℓ_1-norm for the overall underlying network, and the group-level ℓ_2-norm. For Hawkes processes modeled with exponential interaction function, Nickel and Le (2020) decomposed the log-likelihood with regard to active and inactive event times, resulting in a closed-form with exact computation of the gradients. By considering a random sample of observed processes of size, say m, the runtime complexity is of order $\mathcal{O}(m + n\kappa)$, with κ being the number of the m active *entities*. Chen et al. (2021) introduced a new method for modeling spatial TPP with continuous distribution of the marks by combining continuous-time NN with either jump of attentive continuous-time normalizing flows; see, e.g., Jia and Benson (2019) for neural jump SDE. For spatial TPP, Yuan et al. (2023) proposed an advanced deep model, allowing for learning the joint conditional p.d.f. of (t, x) directly, avoiding parameterizing the density $f(x \mid t)$, which can exhibit complex dependence structure.

Intensity-free models. We lastly expose a different approach highly motivated by recent advances in deep architectures for neural density estimation. Indeed, approximating the intensity process with NN can fail in defining a valid p.d.f., and obtaining a closed-form of the expectation of the compensator process. In Shchur et al. (2020), an intensity-free framework models the p.d.f. of the τ_i's by normalizing flows. It consists in defining a complex p.d.f. $p(t)$ as a (forward) transformation g of a simple one $q(z)$, yielding $p(t) = q(g^{-1}(t))|\partial g^{-1}(t)/\partial t|$ after a change of variable, see, e.g., (Tabak and Turner, 2013). This transformation g should be a diffeomorphism allowing for differentiation. But, its inverse mapping is not

necessarily analytically available. Shchur et al. (2020) considered inverse mappings, based on either sigmoids or polynomials (Jaini et al., 2019), and of parameters trained using NN (Huang et al., 2018). The proposed p.d.f. of the TPP is modeled by a finite log-normal mixture, of parameters (means and variances) modeled by functions of linear combination of y_{t-1} and the current (additional) inputs, whereas the weights of the mixture are obtained by computing the softmax image of those parameters. The related loss is penalized by the ℓ_2-norm. We further refer to Shchur et al. (2020), Section 4 therein, for an account of recent references.

The next section presents recent methods for learning processes of approximation models being the solution of the least-squares penalized loss. As we shall see, this approach differs by nature, as it relies on the martingale structure of the target processes, and focuses on their stochastic characterization through their intensity solely.

9.3.3 Quadratic loss functions

In this section, we review statistical methods estimating approximation models for intensity processes, when formulated as solution of penalized quadratic losses. We consider the same notations and framework exposed in Section 9.3.2.

Formulation of quadratic loss functions

We suppose, having observed the counting process, N_t over $[0, T]$, with $T > 0$ finite, under a probability measure $\mathbb{P}_T$, of $\mathcal{F}_t$-predictable intensity λ_t, compatible with the abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the generic filtration $(\mathcal{F}_t)_{t\geq 0}$. We suppose the compensator process Λ_t to be bounded on $[0, T]$ for simplicity. Let $(\lambda_t(\theta))_{\theta\in\Theta}$ be the set of $\mathcal{F}_t$-predictable processes to model the unknown λ_t. The associated generic least-squares contrast function is then

$$\ell_t(\theta) := -\frac{2}{t}\int_0^t \lambda_u(\theta)\mathrm{d}N_u + \frac{1}{t}\int_0^t \lambda_u(\theta)^2\mathrm{d}u \ . \tag{9.20}$$

For linear models, the loss has the concise formulation $\ell_t(\theta) = -2\theta^T H_t + \theta^T G_t\theta$, with $(H_t)_{t\geq 0}$ being a $\mathcal{F}_t$-predictable vector process and $(G_t)_{t\geq 0}$ being the Gram matrix defined as a stochastic quadratic form, both being generated by the transformations of the basis functions of $\mathcal{H}$. For instance, the Gram matrix associated with linear multivariate HP modeled by Eq. (9.12), equals to $G_t = \left(\langle \int g_{i,\cdot}\mathrm{d}N_i, \int g_{i',\cdot}\mathrm{d}N_{i'}\rangle\right)_{i,i'\leq d}$, for a given coordinate $j \leq d$, when the inner product is well defined and the integrals are

defined on $[t-T, t)$. A wide literature has been devoted to contrasts of that form, under various assumptions on the eigenspace of G. The strongest assumption is for G being invertible. Weaker assumptions focus on *local* neighborhoods of θ^*, usually based on its support, and are known as restricted isometry property; see (Koltchinskii, 2011), Chapter 7 therein, and restricted eigenvalue (RE) condition; see (Bickel et al., 2009) in particular. Under such assumptions, it is possible to analyze the expected risk of an empirical minimizer of ℓ_t, wherein the martingale structure of N_t plays a key role. Indeed, Doob–Meyer's decomposition theorem ensures that the process $M_t = N_t - \Lambda_t$ is a (local) $\mathcal{F}_t$-martingale, and implies

$$\mathbb{E}[\ell_t(\theta)] = -\frac{2}{t}\int_0^t \lambda_u(\theta)\lambda_u \mathrm{d}u + \frac{1}{t}\int_0^t \lambda_u(\theta)^2 \mathrm{d}u = \|\lambda(\theta) - \lambda\|_t^2 - \|\lambda\|_t^2 \,, \quad (9.21)$$

which is minimized for $\lambda_t(\theta) = \lambda_t$ a.s. and for almost all t. We expect that the optimal solution of the least-squares results in a *good* approximation model. When decomposing the approximation processes over the class $\mathcal{H}$ of orthonormal basis, indexed by a finite set of parameters Θ, the *best* approximation model $\lambda(\theta)$ is the orthogonal projection of the true intensity λ induced by a parameter θ^*, which we would like to estimate, i.e., $\theta^* \in \arg\min_\theta \mathbb{E}\|\lambda(\theta) - \lambda\|_T^2$. We expect that the minimizer θ_n of the penalized empirical risk, Eq. (9.17), performs similarly to the unknown solution θ^*, expressed by

$$\mathbb{E}\|\lambda(\theta_n) - \lambda\|_T^2 \le C\,\mathbb{E}\|\lambda(\theta^*) - \lambda\|_T^2 \,, \quad (9.22)$$

with the constant $C > 1$, which is supposed to be small. In fact, the condition (9.22) can be extended and used for traditional estimation methods, such as adaptive and model selection methods, as argued in the excellent (Reynaud-Bouret, 2014). By the derivation of (9.21), nonasymptotic statistical guarantees of the empirical θ_n results in the study of $\|\lambda(\theta_n) - \lambda\|_t^2$. It is upperbounded as in Eq. (9.22) with an additive term of rate depending on the penalization function, and where the constant C might increase with T, while ideally tending to one. To achieve such (sharp) oracle inequalities, it has been shown for linear stochastic models for λ_t of the form of Eq. (9.9), that there is an essential relation between optimal data-driven penalty weights $\hat{\rho}$, and the random fluctuations of (the sup-norm of) the (local) martingale process, roughly taking the form

$$Z_t = \int_0^t H_u \left(\mathrm{d}N_u - \lambda_u \mathrm{d}u\right) \,, \quad (9.23)$$

with $(H_t)_{t\geq 0}$ being a predictable process, resulting from the decomposition of $\lambda_t(\theta)$ over a class $\mathcal{H}$, cf. the quadratic formulation for linear models. Recent works proved nonasymptotic probabilistic control for $(Z_t)_{t\geq 0}$ at a $t = T$, $t \leq T$, or stopped if M_t is a local martingale, by means of concentration inequalities. For model selection and adaptive statistics, inequalities following that of Talagrand's for empirical processes are used; see (Reynaud-Bouret, 2003) for inhomogeneous Poisson processes. For proving oracle inequalities for penalized risk functions (Eq. (9.17)), Bernstein-type inequalities are required. To provide some intuition, those allow a probabilistic control of the stochastic deviation of Z_t in terms of its exact variance process and the sup-norm variations of H_t, which we loosely write as follows:

$$\mathbb{P}\left(Z_t \geq \sqrt{2\kappa x} + \frac{\kappa' x}{3}, \quad \int_0^t H_u^2 \lambda_u \mathrm{d}u \leq \kappa, \quad \sup_{u\leq t} |H_u| \leq \kappa'\right) \leq e^{-x}\,, \quad (9.24)$$

for all $x > 0$, as soon as it is well defined; see, e.g., (van de Geer, 1995), Sections 2 and 3 therein, and up to normalization with T. Additional derivations using the peeling technique, aim at replacing the exact variance (or bracket) process with its empirical counterpart $\int_0^t H_u^2 \mathrm{d}N_u =: \widehat{V}_t$, which can directly be estimated based on a data sample. A series of works have been devoted to proving extensions of Eq. (9.24), depending on mainly, the type of TPP, its related (local) assumptions, and of those related to the Gram matrix of the quadratic loss, such as in Hansen et al. (2015); Bacry et al. (2018); Howard et al. (2020); Ost and Reynaud-Bouret (2020); Reynaud-Bouret (2006) for instance. As we will present next, these results yield explicit optimal data-driven penalty weights $\hat{\rho}$ guaranteeing nonasymptotic control of the estimation error $\mathbb{E}\|\lambda(\theta_n) - \lambda\|_T^2$. There are multiple advantages related to those properties. Indeed, many applications face limited number of observed individuals m, and sometimes even only one, compared to the very large number of observed events n, particularly the case for biomedical applications. On the contrary, in survival analysis, there is no interest for large observational time T, as no additional information can be gained after failure. In addition, even though m can be small, providing optimal data-driven weights for the penalized loss prevents from additional computational and storage costs. This would typically be the case with cross-validation and sample-splitting to numerically estimate $\hat{\rho}$.

In the following paragraphs, we review models falling into that line of research, either for specific models of TPP, and usually of Poisson or Hawkes, and highlight the underlying assumptions related to the derivation of such guarantees.

Approximation models for Hawkes processes

A rich literature has been established for (multivariate) HPs for fixed function $\varphi(u)$ by proposing approximation models for the intensity process based on Eq. (9.12). The goal is to estimate the optimal decomposition of the interaction functions h over a class of basis functions, minimizing Eq. (9.17). Reynaud-Bouret and Schbath (2010) solve the lasso problem for linear HPs ($\varphi(u) = u$), with $\mathcal{H}$ being a L_2-dictionary of weighted intervals of the form (9.18). Reynaud-Bouret and Schbath (2010) proved the nonasymptotic control of the expected risk, with convergence rate of smaller order than $\mathcal{O}(\log(T)/T)$; see Proposition 1 therein. The penalty function is obtained of order $\mathcal{O}(|n|\log(T)^2/T)$ by model selection, resulting in a sharp oracle inequality bounding the expected risk, with both μ_0 and the h's to be truncated outside a fixed compact. It recovers classic well-known results, wherein the penalty is related to the variance of the underlying process, cf. Eq. (9.24). However, the authors did not consider weighted penalization function, due to the complexity of the resulting formulation induced by the dependence structure of HPs. Those new improvements were later obtained in Hansen et al. (2015), which were used in Reynaud-Bouret et al. (2013) in the context of dictionary learning for linear multivariate HP, where the h's are of bounded support, applied to spike trains analysis. The nonasymptotic bound is obtained on the high probability event that the spectral eigenspace of the Gram matrix based on the class $\mathcal{H}$ is lower bounded by a fixed positive constant, i.e., implying that it is invertible. In this direction and for multivariate HPs, Bacry et al. (2020) derived sharp oracle inequalities for the estimation error, wherein the weighted penalization function is of the form $\pi(\rho, \theta) = \rho_1\|\mu_0\|_1 + \rho_2\|H\|_1 + \rho_3\|H\|_*$, where H is the *self-excitement* tensor encoding the integrated kernels, being positive-valued functions with unit ℓ_1-norm (see Eq. (9.12)) and the ρ's are data-driven optimal weights. The consideration of the operator norm for matrix martingales required an additional Bernstein inequality for matrix martingales, proved in Bacry et al. (2018), and extended in Bacry et al. (2020) to remove the boundness conditions on $\widehat{V}_T$ and $\sup_t |H_t|$; see Theorem 4 therein. The nonasymptotic control of the estimation error relies on the restricted eigenvalue (RE) condition, introduced and analyzed in Bickel et al. (2009); Koltchinskii (2009a,b). When considering the estimation based on a single observed process for very large T, they proposed an algorithmic procedure achieving faster convergence ($\mathcal{O}(|I|^2 d^3)$) than likelihood-based methods ($\mathcal{O}(n|I|d)$), as $n >> |I|d^2$, but as well with regard to memory, test, and gradient considerations. As pointed out in Staerman

et al. (2023), choosing exponential basis functions specifically, results in a very efficient procedure of precomputational complexity in $\mathcal{O}(n)$. They developed a fast algorithm for learning multivariate linear HP as solution of the lasso penalization, where $\mathcal{H}$ is composed of smooth functions with bounded support, and relying on ℓ_2-gradient based solvers after discretization of the sequence of events. They proved asymptotic consistency (a.s.) of the empirical estimator, as soon as the discretization length tends to zero. We further refer to the interesting application to goodness-of-fit testing for Poisson and Hawkes processes applied to neuronal spike trains in Reynaud-Bouret et al. (2014), relying on the adaptive estimation of the best intensity model decomposed on histograms of Haar intervals in particular; see Section 9.3.1.

Approximation models for TPP and spatial TPP

One of the first works to derive oracle inequalities with data-driven weights for solutions of Eq. (9.17) with lasso penalization is, to the best of our knowledge, in Gaïffas and Guilloux (2012). In the specific application of survival analysis, they supposed having observed an i.i.d. sample of marked counting processes $\{N_j,\, j \leq m\}$, satisfying the Aalen multiplicative model. The resulting rate of convergence is of order $\mathcal{O}_{\mathbb{P}}((\log|I|/m)^{1/2})$ under no constraints on the Gram matrix. If it fulfills the RE assumption, however, faster rates are obtained $\mathcal{O}_{\mathbb{P}}(\log|I|/m)$. The empirical weights are of order $\mathcal{O}_{\mathbb{P}}(((x+\log d)\widehat{V}_{T,j}/m)^{1/2})$, for each coordinate $j \leq d$. Those results rely on a new Bernstein inequality for martingales resulting from counting processes of the form of Eq. (9.24), proved in Theorem 3 therein, where its bracket is replaced with the empirical variance process using a peeling method. Theorem 3 was used in Alaya et al. (2015) to show oracle guarantee for fused lasso based on the model $\lambda_t(\theta) = \sum_{i=1}^{n} \theta_i 1_{((i-1)/n, i/n]}(t)$; cf. (9.18). The penalization function is the weighted total-variation norm: $\pi(\rho, \theta) = \sum_{i=1}^{n} \rho_i |\theta_{i-1} - \theta_i|$, wherein the vector of weights ρ is optimally chosen based on the i.i.d. sample of observed N_j's, similar to Gaïffas and Guilloux (2012), resulting from the oracle inequality proved in Theorem 1 in Alaya et al. (2015). The optimal penalization weights are, in this case, of order $\mathcal{O}_{\mathbb{P}}(((ne_{j,n}\log n)/m)^{1/2})$, with $e_{j,n}$ being the empirical average of events on the m sample within the window $((i-1)T/n, T]$. The authors extended it to change-point detection method for the intensity process. The procedure is asymptotically consistent if the event-times of the process are at least c/n far apart, with $c > 8$; see Theorem 3 therein. In addition, if the number of events is overestimated, then they are asymptotically close to the

true ones, when $m \to \infty$, with explicit order of convergence. In parallel, Hansen et al. (2015) proposed a nonparametric model for multivariate TPP by modeling λ_t using a first-order interaction kernel, like Eq. (9.12). The interaction kernels are decomposed over a generic dictionary $\mathcal{H}$ of functions with bounded support. They importantly proved a new data-driven Bernstein inequality for bounding local martingales from counting processes; see Section 7 in particular. Hansen et al. (2015) obtained, similar to Gaïffas and Guilloux (2012), oracle inequalities for recovering multivariate TPP using lasso risk, with explicit constants, and of optimal data-driven penalty weights of order $\mathcal{O}_{\mathbb{P}}(\max((x\widehat{V}_T)^{1/2}, \kappa' x))$, where κ' depend on the interaction functions h's. However, their results rely on a stronger assumption on the Gram matrix, assumed to be invertible, and consider $m = 1$ with possibly very large T. We refer to Lambert et al. (2018) for an application with multivariate HPs for learning the functional connectivity graph of neuronal spike activity. We lastly refer to Qiu and Yang (2023) for spatio-temporal process monitoring, wherein the process is linearly decomposed $X_{t_i}(x_{ij}) = \mu_{t_i}(x_{ij}) + \varepsilon_{t_i}(x_{ij})$ for all time t_i, $i \leq n$, all locations x_{ij}, $j \leq m_i$ at time t_i. The first process μ is the mean of X, and ε the centered error random variable. By estimating the covariance function $\mathrm{Cov}(X_t(s), X_{t'}(s'))$, the authors proposed a (weighted) lasso objective function with an exponentially weighted kernel smoothing to estimate the empirical decentralized and decorrelated residuals (see Eq. 4) for *sequential monitoring*. We further refer to the literature cited therein.

9.4. Conclusion

In this chapter, we presented a review of recent advances in the statistical literature for modeling and learning (marked) temporal point processes. The focus was on optimizing a penalized loss function, in particular with the loss function being either based on the log-likelihood or on the quadratic contrast. These loss functions are in general given in terms of the intensity process describing the model. We presented a general framework for comparing those methods, and highlighted their ability to efficiently perform the statistical procedure when based on large datasets.

Acknowledgments

The work was supported by Novo Nordisk Foundation Grant NNF20OC0062897.

References

Alaya, M.Z., Gaïffas, S., Guilloux, A., 2015. Learning the intensity of time events with change-points. IEEE Transactions on Information Theory 61 (9), 5148–5171.

Bacry, E., Muzy, J.-F., 2014. Hawkes model for price and trades high-frequency dynamics. Quantitative Finance 14 (7), 1147–1166.

Bacry, E., Mastromatteo, I., Muzy, J.-F., 2015. Hawkes processes in finance. Market Microstructure and Liquidity 01 (01), 1550005.

Bacry, E., Gaïffas, S., Muzy, J.-F., 2018. Concentration inequalities for matrix martingales in continuous time. Probab. Theory Relat. Fields 170, 525–553.

Bacry, E., Bompaire, M., Gaïffas, S., Muzy, J.-F., 2020. Sparse and low-rank multivariate Hawkes processes. Journal of Machine Learning Research 21 (50), 1–32.

Banerjee, S., Carlin, B., Gelfand, A., 2014. Hierarchical Modeling and Analysis for Spatial Data, 2nd ed. Chapman and Hall/CRC.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of lasso and Dantzig selector. The Annals of Statistics 37 (4), 1705–1732.

Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3 (1), 1–122.

Brémaud, P., 2020. Point Process Calculus in Time and Space: An Introduction with Applications. Probability Theory and Stochastic Modelling. Springer International Publishing.

Brémaud, P., Massoulié, L., 1996. Stability of nonlinear Hawkes processes. Annals of Probability 24 (3), 1563–1588.

Brillinger, D.R., 2004. Maximum likelihood analysis of spike trains of interacting nerve cells. Biological Cybernetics 59, 189–200.

Brillinger, D.R., Segundo, J.P., 1979. Empirical examination of the threshold model of neuron firing. Biological Cybernetics 35, 213–220.

Chen, R.T.Q., Amos, B., Nickel, M., 2021. Neural spatio-temporal point processes. In: International Conference on Learning Representations.

Christgau, A.M., Petersen, L., Hansen, N.R., 2023. Nonparametric conditional local independence testing. The Annals of Statistics 51 (5), 2116–2144.

Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Daley, D.J., Vere-Jones, D., 1998. An Introduction to the Theory of Point Processes, vol. I, first edition. Springer Series in Statistics. Springer-Verlag, New York. General theory and structure.

Du, N., Song, L., Yuan, M., Smola, A., 2012. Learning networks of heterogeneous influence. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems, vol. 25. Curran Associates, Inc.

Du, N., Wang, Y., He, N., Sun, J., Song, L., 2015. Time-Sensitive Recommendation from Recurrent User Activities. In Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L., 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, KDD'16. Association for Computing Machinery, New York, NY, USA, pp. 1555–1564.

Gaïffas, S., Guilloux, A., 2012. High-dimensional additive hazards models and the lasso. Electronic Journal of Statistics 6, 522–546.

Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B., 2011. Uncovering the temporal dynamics of diffusion networks. In: Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11. Omnipress, pp. 561–568.

Gomez-Rodriguez, M., Leskovec, J., Krause, A., 2012. Inferring networks of diffusion and influence. ACM Transactions on Knowledge Discovery from Data 5 (4).
González, J.A., Rodríguez-Cortés, F.J., Cronie, O., Mateu, J., 2016. Spatio-temporal point process statistics: a review. Spatial Statistics 18, 505–544.
Hansen, N.H., 2013. Penalized maximum likelihood estimation for generalized linear point processes.
Hansen, N.R., 2015. Nonparametric likelihood based estimation of linear filters for point processes. Statistics and Computing 25 (3), 609–618.
Hansen, N.R., Reynaud-Bouret, P., Rivoirard, V., 2015. Lasso and probabilistic inequalities for multivariate point processes. Bernoulli 21 (1), 83–143.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York Inc., New York, NY, USA.
Hawkes, A.G., 1971a. Spectra of some self-exciting and mutually exciting point processes. Biometrika 58 (1), 83–90.
Hawkes, A.G., 1971b. Point spectra of some mutually exciting point processes. Journal of the Royal Statistical Society, Series B, Methodological 33 (3), 438–443.
Hawkes, A.G., Oakes, D., 1974. A cluster process representation of a self-exciting process. Journal of Applied Probability 11 (3), 493–503.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9 (8), 1735–1780.
Howard, S.R., Ramdas, A., McAuliffe, J., Sekhon, J., 2020. Time-uniform Chernoff bounds via nonnegative supermartingales. Probability Surveys 17, 257–317.
Huang, C.-W., Krueger, D., Lacoste, A., Courville, A., 2018. Neural autoregressive flows. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80. PMLR, pp. 2078–2087.
Jacobsen, M., 2005. Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes. Probability and Its Applications. Birkhäuser, Boston.
Jaini, P., Selby, K.A., Yu, Y., 2019. Sum-of-squares polynomial flow. In: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 97. PMLR, pp. 3009–3018.
Jia, J., Benson, A.R., 2019. Neural jump stochastic differential equations. In: Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc.
Koltchinskii, V., 2009a. The Dantzig selector and sparsity oracle inequalities. Bernoulli 15 (3), 799–828.
Koltchinskii, V., 2009b. Sparsity in penalized empirical risk minimization. Annales de L'I.H.P. Probabilités et Statistiques 45 (1), 7–57.
Koltchinskii, V., 2011. Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer.
Kutoyants, Y.A., 1984. Parameter Estimation for Stochastic Processes. Heldermann.
Lambert, R.C., Tuleau-Malot, C., Bessaih, T., Rivoirard, V., Bouret, Y., Leresche, N., Reynaud-Bouret, P., 2018. Reconstructing the functional connectivity of multiple spike trains using Hawkes models. Journal of Neuroscience Methods 297, 9–21.
Lemonnier, R., Vayatis, N., 2014. Nonparametric Markovian learning of triggering kernels for mutually exciting and mutually inhibiting multivariate Hawkes processes. In: Machine Learning and Knowledge Discovery in Databases. Springer, Berlin Heidelberg, pp. 161–176.
Lemonnier, R., Scaman, K., Kalogeratos, A., 2017. Multivariate Hawkes processes for large-scale inference. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (1).

Lewis, E., Mohler, G., 2011. A nonparametric em algorithm for multiscale Hawkes processes. Journal of Nonparametric Statistics.
Liu, Y., Yan, T., Chen, H., 2018. Exploiting graph regularized multi-dimensional Hawkes processes for modeling events with spatio-temporal characteristics. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. In: International Joint Conferences on Artificial Intelligence Organization, vol. 7, pp. 2475–2482.
Massart, P., Picard, J., 2007. Concentration Inequalities and Model Selection. Lecture Notes in Mathematics. Springer, Berlin New York.
Mei, H., Eisner, J.M., 2017. The neural Hawkes process: a neurally self-modulating multivariate point process. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc.
Mogensen, S.W., Hansen, N.R., 2020. Markov equivalence of marginalized local independence graphs. The Annals of Statistics 48 (1), 539–559.
Mogensen, S.W., Malinsky, D., Hansen, N.R., 2018. Causal learning for partially observed stochastic dynamical systems. In: Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, pp. 350–360.
Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E., 2011. Self-exciting point process modeling of crime. Journal of the American Statistical Association 106 (493), 100–108.
Myers, S.A., Leskovec, J., 2010. On the convexity of latent social network inference. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems – Volume 2, NIPS'10. Red Hook, NY, USA. Curran Associates, Inc., pp. 1741–1749.
Nesterov, Y., Nemirovskii, A.S., Ye, Y., 1994. Interior Point Polynomial Algorithms in Convex Programming, vol. 13. SIAM.
Nickel, M., Le, M., 2020. Learning multivariate Hawkes processes at scale.
Ogata, Y., 1978. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. Annals of the Institute of Statistical Mathematics 30, 243–261.
Ogata, Y., 1988. Statistical models for earthquake occurrences and residual analysis for point processes. Journal of the American Statistical Association 83 (401), 9–27.
Ogata, Y., Matsu'ura, R.S., Katsura, K., 1993. Fast likelihood computation of epidemic type aftershock-sequence model. Geophysical Research Letters 20 (19), 2143–2146.
Ost, G., Reynaud-Bouret, P., 2020. Sparse space–time models: concentration inequalities and lasso. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques 56 (4), 2377–2405.
Ozaki, T., 1979. Maximum likelihood estimation of Hawkes' self-exciting point processes. Annals of the Institute of Statistical Mathematics 31 (1), 145–155.
Paninski, L., 2004. Maximum likelihood estimation of cascade point-process neural encoding models. Network Computation in Neural Systems 15, 243–262.
Park, J., Schoenberg, F.P., Bertozzi, A.L., Brantingham, P.J., 2021. Investigating clustering and violence interruption in gang-related violent crime data using spatial–temporal point processes with covariates. Journal of the American Statistical Association 116 (536), 1674–1687.
Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.
Protter, P., 1992. Stochastic Integration and Differential Equation, second edition. Springer-Verlag, Berlin, Heidelberg.
Qiu, P., Yang, K., 2023. Spatio-temporal process monitoring using exponentially weighted spatial lasso. Journal of Quality Technology 55 (2), 163–180.
Reinhart, A., 2018. A review of self-exciting spatio-temporal point processes and their applications. Statistical Science 33 (3), 299–318.

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods in Ecology and Evolution 6 (4), 366–379.

Reynaud-Bouret, P., 2003. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. Probability Theory and Related Fields 126, 103–153.

Reynaud-Bouret, P., 2006. Compensator and exponential inequalities for some suprema of counting processes. Statistics & Probability Letters 76 (14), 1514–1521.

Reynaud-Bouret, P., 2014. Concentration inequalities, counting processes and adaptive statistics. ESAIM. Proceedings 44, 79–98.

Reynaud-Bouret, P., Schbath, S., 2010. Adaptive estimation for Hawkes processes, application to genome analysis. The Annals of Statistics 38 (5), 2781–2822.

Reynaud-Bouret, P., Rivoirard, V., Tuleau-Malot, C., 2013. Inference of functional connectivity in neurosciences via Hawkes processes. In: 2013 IEEE Global Conference on Signal and Information Processing, pp. 317–320.

Reynaud-Bouret, P., Rivoirard, V., Grammont, F., Tuleau-Malot, C., 2014. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. Journal of Mathematical Neuroscience 4 (3).

Rommel, C., Paillard, J., Moreau, T., Gramfort, A., 2022. Data augmentation for learning predictive models on EEG: a systematic comparison. Journal of Neural Engineering 19 (6).

Salehi, F., Trouleau, W., Grossglauser, M., Thiran, P., 2019. Learning Hawkes Processes from a Handful of Events. In Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc.

Schoenberg, F.P., 2013. Facilitated estimation of etas. Bulletin of the Seismological Society of America 103 (1), 601–605.

Shchur, O., Biloš, M., Günnemann, S., 2020. Intensity-free learning of temporal point processes. In: International Conference on Learning Representations.

Shchur, O., Türkmen, A.C., Januschowski, T., Günnemann, S., 2021. Neural temporal point processes: a review. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. In: International Joint Conferences on Artificial Intelligence Organization, vol. 8, pp. 4585–4593.

Silverman, B., 1986. Density Estimation for Statistics and Data Analysis, vol. 26. CRC Press.

Sokol, A., Hansen, N.R., 2015. Exponential martingales and changes of measure for counting processes. Stochastic Analysis and Applications 33 (5), 823–843.

Staerman, G., Allain, C., Gramfort, A., Moreau, T., 2023. Fadin: fast discretized inference for Hawkes processes with general parametric kernels. In: Proceedings of the 40th International Conference on Machine Learning, ICML'23.

Sulem, D., Rivoirard, V., Rousseau, J., 2023. Scalable and adaptive variational bayes methods for hawkes processes.

Sulem, D., Rivoirard, V., Rousseau, J., 2024. Bayesian estimation of nonlinear Hawkes processes. Bernoulli 30 (2), 1257–1286.

Tabak, E.G., Turner, C.V., 2013. A family of nonparametric density estimation algorithms. Communications on Pure and Applied Mathematics 66 (2), 145–164.

van de Geer, S., 1995. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. The Annals of Statistics 23 (5), 1779–1801.

Veen, A., Schoenberg, F.P., 2008. Estimation of space–time branching process models in seismology using an em–type algorithm. Journal of the American Statistical Association 103 (482), 614–624.

Wu, W., Yan, J., Yang, X., Zha, H., 2018. Decoupled learning for factorial marked temporal point processes. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Xu, H., Farajtabar, M., Zha, H., 2016. Learning granger causality for Hawkes processes. In: Proceedings of the 33rd International Conference on Machine Learning. New York, New York, USA. In: Proceedings of Machine Learning Research, vol. 48, pp. 1717–1726. PMLR.

Yan, J., 2019. Recent advance in temporal point process: from machine learning perspective.

Yuan, B., Li, H., Bertozzi, A.L., Brantingham, P.J., Porter, M.A., 2019. Multivariate spatiotemporal Hawkes processes and network reconstruction. SIAM Journal on Mathematics of Data Science 1 (2), 356–382.

Yuan, Y., Ding, J., Shao, C., Jin, D., Li, Y., 2023. Spatio-temporal diffusion point processes. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD'23. Association for Computing Machinery, New York, NY, USA, pp. 3173–3184.

Zhou, F., Li, Z., Fan, X., Wang, Y., Sowmya, A., Chen, F., 2020. Efficient inference for nonparametric Hawkes processes using auxiliary latent variables. Journal of Machine Learning Research 21 (241), 1–31.

Zhou, K., Zha, H., Song, L., 2013. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics. Scottsdale, Arizona, USA. In: Proceedings of Machine Learning Research, vol. 31, pp. 641–649. PMLR.