Cumulants of Hawkes Processes are Robust to Observation Noise

William Trouleau [*1] **Jalal Etesami** [*2] **Matthias Grossglauser** [1] **Negar Kiyavash** [2] **Patrick Thiran** [1]

Abstract

Multivariate Hawkes processes (MHPs) are widely used in a variety of fields to model the occurrence of causally related discrete events in continuous time. Most state-of-the-art approaches address the problem of learning MHPs from perfect traces without noise. In practice, the process through which events are collected might introduce noise in the timestamps. In this work, we address the problem of learning the causal structure of MHPs when the observed timestamps of events are subject to random and unknown shifts, also known as random translations. We prove that the cumulants of MHPs are invariant to random translations, and therefore can be used to learn their underlying causal structure. Furthermore, we empirically characterize the effect of random translations on state-of-the-art learning methods. We show that maximum likelihood-based estimators are brittle, while cumulant-based estimators remain stable even in the presence of significant time shifts.

1. Introduction

From modeling the price change in financial markets (Bacry et al., 2015), to analyzing epidemic pathways in global outbreaks of infectious diseases (Colizza et al., 2007), or yet uncovering the dynamics of information on social media (Gomez-Rodriguez et al., 2011), extracting the diffusion patterns of sequences of discrete events is a problem that is of interest in many fields. Event data typically consists of a set of marked timestamps, where the mark of an occurrence denotes its type, such as its geographical location.

To model such event data, temporal point processes, which model the probability of occurrence of a set of events in continuous time, are rising as a popular framework. In finance, they are used to model the stochastic time evolution of limit order books (Da Fonseca & Zaatour, 2014; Abergel & Jedidi, 2015). In neuroscience, they are used to model networks of stochastic spiking neurons (Truccolo et al., 2005; Reynaud-Bouret et al., 2013; Gerhard et al., 2017). In epidemiology, they enable to get beyond the mean-field assumptions of classic epidemic models based on differential equations, and to capture the self and mutual excitation nature of disease spread across heterogeneous social systems (Kim et al., 2019).

In practice, the process through which the sequences of events are collected often introduces noise in the observed timestamps. For instance, in neuroscience, the activity of neurons is typically collected by measuring a continuous signal coming from the action potential of neurons using electrode micro-arrays. The signal is then converted into a discrete sequence of events of firing neurons, called spike trains, which are the times when the action potential exceeds a threshold. This procedure is inherently noisy and prone to introduce inaccuracies in the measured timestamps. Another example is in epidemiology, where the reported times of infection have an approximate granularity, and do not account for the latent incubation period. This could lead to inaccuracies in the measured timestamps. As a result, a secondary case might be reported before the primary case, which could interfere with learning the true causation structure.

The current literature for learning temporal point-processes assumes perfect information regarding the observation. In this work, we consider inferring the causal network of a popular class of temporal point-processes, called multivariate Hawkes processes (MHPs), when the observations are subject to a particular form of noise, called *random translation*. In a randomly translated point process, every event within a dimension is shifted randomly and independently in time according to a fixed but unknown distribution. We show that the cumulants of an MHP are invariant with respect to random translations. Therefore, any inference method that can obtain the causal network of an MHP from its cumulants can also be used to learn its causal network under random translation noise.

[*]Equal contribution [1]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland [2]College of Management of Technology, EPFL, Lausanne, Switzerland. Correspondence to: William Trouleau <william.trouleau@epfl.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

2. Related Works

Learning the excitation matrix of an MHP from a set of observations has been the focus of recent work (Xu et al., 2016; Yang et al., 2017; Salehi et al., 2019). The main approaches for inferring the excitation matrix of MHPs are of two flavors: maximum likelihood-based approaches such as (Ozaki, 1979; Zhou et al., 2013; Yang et al., 2017; Salehi et al., 2019); or moment-based approaches that learn the parameters of interest by solving a set of equations obtained from first, second, or third-order moments of the MHP (Hawkes, 1971b; Bacry et al., 2012; Bacry & Muzy, 2016; Etesami et al., 2016; Achab et al., 2017). All the aforementioned approaches assume that the observations are noiseless, that is, the arrival times of the events are recorded accurately without any delay.

A recent study addressed the case where events are synchronized (Trouleau et al., 2019). This is a special case of the random translation framework that we study in this work. More precisely, in our general random translation noise model, the events of a dimension are shifted independently according to some unknown distribution. In the synchronized noise model, all events within a dimension have the exact same delay. See Section 4 for more details.

The inference of temporal point processes in the presence of noisy observations has also been studied for other types of point processes, such as spatial Poisson processes (Cucala, 2008; Bar-Hen et al., 2013). However, these studies mostly focus on the special case of independent and known noise, which is not applicable to MHPs. Another line of research that tackles the inference problem in Hawkes processes without perfect observations appears in (Xu et al., 2017; Shelton et al., 2018), where the inference problem with missing data is considered. In this work, data are not missing, but timestamps are inaccurately measured.

Hoffmann & Caramanis (2019) consider a type of temporal noise in the context of disease modeling. In particular, they study the inference of epidemic pathways for a discrete-time epidemic model spreading over a network of individuals, when the infection times are not known exactly. The approaches developed in that work are designed for a discrete-time model where each dimension has at most one event, *i.e.*, the infection time of an individual. Hence, these methods are not applicable to our setting. In the context of univariate Hawkes processes, Deutsch & Ross (2021) have studied a similar type of noise, referred to as "*data distortion*". They propose an approach to estimate the parameters of the process based on Approximate Bayesian Computation (ABC) and Markov Chain Monte Carlo (MCMC). However, the method is limited to the univariate setting.

3. Preliminaries

We begin by introducing some notations. Plain letters denote scalar values, while boldface letters denote column vectors, matrices, and tensors. We denote the Dirac function by $\delta(t)$. For a given function $f(t)$, we denote its time reversed version $\underline{f}(t) \coloneqq f(-t)$, and we define its convolution with a function $g(t)$ by $f * g(t) \triangleq \int_{\mathbb{R}} f(t-x)g(x)dx$. We use $f^{*n}(t)$ to denote the convolution of $f(t)$ with itself n times. The n-dimensional Laplace transform of a function $f(\boldsymbol{x})$ is given by $\mathcal{L}[f](\boldsymbol{s}) \triangleq \int_{\mathbb{R}^n} f(\boldsymbol{x}) \exp(-\boldsymbol{s}^T\boldsymbol{x})d\boldsymbol{x}$. Finally, the Laplace transform of a matrix function $\boldsymbol{G}(t) = [G_{i,j}(t)]$, denoted by $\mathcal{L}[\boldsymbol{G}](s) \triangleq [\mathcal{L}[G_{i,j}](s)]$, is done element-wise.

3.1. Temporal Point Process

Consider a sequence $\boldsymbol{t} = \{t_k\}_{k\geq 0}$ of positive random variables representing the times of random occurrence of a set of events. Let $N(t)$ denote the number of events occurring before time $t \in \mathbb{R}$. The conditional probability, given the past activity, of a new event occurring in the interval $(t, t+dt)$ is specified by the conditional intensity function $\lambda(t)$. Additionally, we assume that the probability of two or more events arriving simultaneously is negligibly small. More specifically, up to first order, we have

$$\mathbb{P}(dN(t) = 1|\mathcal{H}_t) = \lambda(t)dt,$$
$$\mathbb{P}(dN(t) > 1|\mathcal{H}_t) = o(dt)$$

where $\mathcal{H}_t$ describes the history of the point process up to time t.

3.2. Multivariate Hawkes process

Formally, a d-dimensional multivariate Hawkes process (MHP) is a collection of d univariate temporal point-processes $N_i(t)$, $i = 1, \dots, d$, also called dimensions, with conditional intensity functions taking the form

$$\lambda_i(t|\mathcal{H}_t) = \mu_i + \sum_{j=1}^{d} \int_0^t G_{i,j}(t-\tau)dN_j(\tau), \qquad (1)$$

where $\mathcal{H}_t = \bigcup_{i=1}^d \mathcal{H}_t^i$ and $\mathcal{H}_t^i$ is the history of the i-th process up to time t. The constant μ_i is the exogenous part of the intensity of the i-th process. The excitation function $G_{i,j}(t)\colon \mathbb{R} \mapsto \mathbb{R}_+$ is causal, non-negative, and captures the endogenous dynamics of influence of the arrivals in the j-th dimension on the intensity of the i-th dimension[1]. The matrix $\boldsymbol{G}(t) \coloneqq [G_{i,j}(t)]$ is called the excitation matrix. The process is called stable if and only if the spectral radius $\rho(\boldsymbol{G})$ of the *integrated* excitation matrix

$$\boldsymbol{G} \triangleq \mathcal{L}[\boldsymbol{G}](0) = \int_{\mathbb{R}} \boldsymbol{G}(t)dt$$

[1]The function $G_{i,j}(\cdot)$ is causal in the sense that $G_{i,j}(t) = 0$ for $t < 0$.

is strictly less than 1, in which case the process is said to have *stationary increments*.

It has been shown that the support of the excitation matrix encodes the causal structure of the MHP in terms of Granger causality, *i.e.*, process j does not Granger-cause process i if and only if $G_{ij}(t) = 0$ (Etesami et al., 2016; Eichler et al., 2017). The Granger-causal graph of a d-dimensional MHP is therefore a directed graph on d nodes (each dimension is denoted by a node), with a directed edge from node j to node i if and only if $G_{ij}(t) \neq 0$. For more details on MHPs, we refer the interested reader to (Liniger, 2009).

3.3. Poisson Cluster Representation

There exists an equivalent definition of MHPs based on the *Poisson cluster* process, generated by a certain branching structure. The cluster process representation is defined as follows:

- Let I^k be a realization, on the interval $[0, T]$, of a homogeneous Poisson process with rate μ_k. We call the points in I^k *immigrants* of type k.
- For every k, each immigrant $x \in I^k$ generates a cluster of points C_x^k . All such clusters are mutually independent.
- The clusters C_x^k are generated according to the following branching structure:
  - Each cluster C_x^k consists of generations of offspring of all types, where the immigrant x itself belongs to generation 0.
  - Recursively, given the immigrant x and the offspring of generation $1, 2, \ldots, n$ of all types, every "child" y of generation n and type j, produces its own offspring of generation $n + 1$ and type i, $\forall i$, by generating a realization of an inhomogeneous Poisson process with rate $\lambda(t) := G_{i,j}(t - y)$.

An illustration of the cluster representation is shown in Figure 1.

3.4. Cumulants of a Hawkes Process

Consider an arbitrary n-dimensional random vector $\boldsymbol{x} = (x_1, \ldots, x_n)$. The cumulant of order n, denoted by $K(\boldsymbol{x})$, is a measure of statistical dependence of the components of $\boldsymbol{x}$ and is defined as

$$K(\boldsymbol{x}) := \sum_{\pi} (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{C \in \pi} \mathbb{E}\left[\prod_{c \in C} x_c\right], \quad (2)$$

where the sum is over all partitions π of the set $\{1, \ldots, n\}$, and where $|\pi|$ denotes the number of blocks of a given

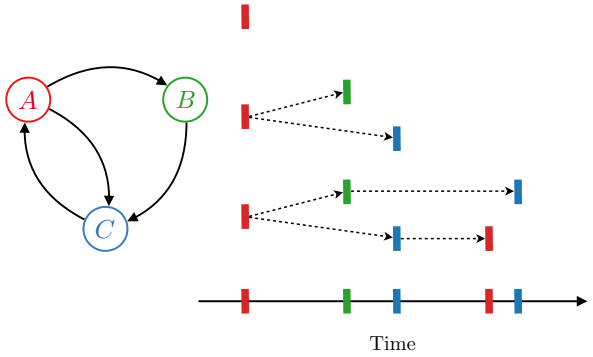


Figure 1: Illustration of the evolution of a Poisson cluster on a network of 3 nodes and 4 directed links. Types (dimensions) are coded by color. The immigrant is of type A (in red). The first generations are one from type B (in black) and one from type C (in blue). The evolution is shown up to the second generation (*i.e.*, tree structure for generations 0, 1 and 2, and the resulting point process over time).

partition (Lukacs, 1970). For example, for $n = 1$, the first-order cumulant density $K(x) = \mathbb{E}[x]$ is the expected value; for $n = 2$, $K(x_1, x_2) = \mathbb{E}[x_1 x_2] - \mathbb{E}[x_1]\mathbb{E}[x_2] = \mathrm{Cov}(x_1, x_2)$ is the covariance; and for $n = 3$, the third-order cumulant is the skewness.

For a given time vector $\boldsymbol{t} = (t_1, \ldots, t_m)$ and a multi-index $\boldsymbol{i} = (i_1, \ldots, i_m)$, the m-th order cumulant density and integrated cumulant of the Hawkes process are defined respectively by

$$K_{\boldsymbol{i}}(\boldsymbol{t}) := \frac{K(dN_{i_1}(t_1), \ldots, dN_{i_m}(t_m))}{dt_1 \ldots dt_m},$$
$$\overline{K}_{\boldsymbol{i}} := \mathcal{L}\left[K_{\boldsymbol{i}}\right](\boldsymbol{0}),$$

where $K(\cdot)$ is the cumulant function defined in (2). For more comprehensive details, we refer the reader to (Jovanović et al., 2015).

4. Methodology

This section introduces a particular form of noise for point processes, called *random translations*. We then characterize the cumulants of a randomly translated MHP and characterize the robustness of cumulant-based estimators for learning their excitation matrix.

4.1. Random Translations

In a randomly translated point process, all the events are shifted randomly in time according to an unknown distribution (Daley & Jones, 2003). More precisely, suppose $\boldsymbol{t} = \left\{\left\{t_k^i\right\}_{k=0}^{n_i}\right\}_{i=1}^{d}$ is a sequence of discrete events, where t_k^i denotes the k-th event in the i-th dimension. A random

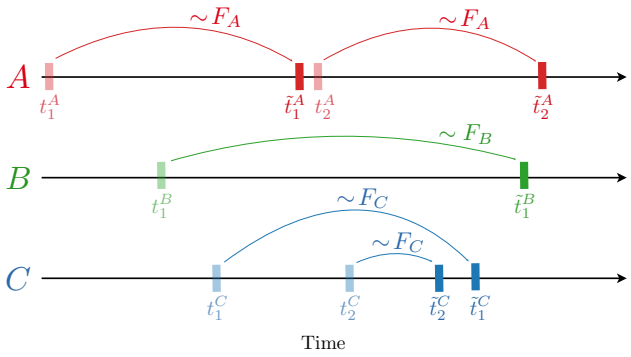


Figure 2: An example of events in a three-dimensional point process and their translations. Events in dimension A, (resp., B and C) are translated randomly by F_A (resp., F_B and F_C).

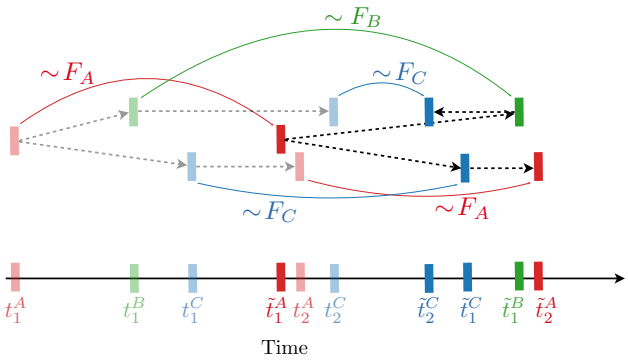


Figure 3: The cluster of Figure 1, with the immigrant of type A and its four descendants translated according to distributions $\{F_A, F_B, F_C\}$.

translation of $\boldsymbol{t}$ is denoted by $\tilde{\boldsymbol{t}}$ and is defined by

$$\tilde{\boldsymbol{t}} = \{\{\tilde{t}_k^i\}_{k=0}^{n_i}\}_{i=1}^d := \{\{t_k^i + x_k^i\}_{k=0}^{n_i}\}_{i=1}^d, \tag{3}$$

where $\{x_k^i\}_{k=0}^{n_i}$ are independent, identically distributed random variables with distribution function $F_i(\cdot)$, for all $1 \leq i \leq d$. Figure 2 demonstrates a simple MHP in three dimensions, in which events are translated according to distribution functions $\{F_A, F_B, F_C\}$. Note that the synchronization noise model proposed by Trouleau et al. (2019) is a special case of the random translation, when all the distributions are Dirac delta functions, *i.e.*, for every i, $dF_i(x) = \delta(x - z_i)dx$, where $z_i \in \mathbb{R}_+$.

Among the approaches to learn the randomly translated MHP, a first candidate is a maximum-likelihood based estimation, such as expectation maximization. However, as discussed in Trouleau et al. (2019), such method results in a non-convex objective function, has a high computational complexity, and fails even for the synchronized translations as the noise power increases. For the sake of completeness, we will demonstrate the similar shortcomings of the maximum-likelihood estimator for the random translation setting through empirical experiments.

4.2. Cumulants of Randomly Translated MHP

Jovanović et al. (2015) showed that the cumulant densities of an MHP can be calculated analytically through their cluster representation. This result establishes the relationships between the integrated cumulants of an MHP and its excitation matrix. Achab et al. (2017) used this relationship to develop an algorithm called NPHC to learn the causal network of an MHP given its integrated cumulants. They also provided an estimator for the first, second, and third-order integrated cumulants given a set of observations.

We will compute the cumulant densities of a randomly translated MHP using its cluster representation and show how they relate to the causal structure of the underlying MHP. To do so, we have to study the effect of random translations on the clusters of an MHP. We observe two key properties, which we discuss in the context of a simple example illustrated in Figure 3. As the random translations occur after the realization of the process, the Poisson cluster representation still holds. Hence, although the events within this cluster are randomly displaced, the tree structure (*i.e.*, the parent-children relationships) of the cluster is unaffected. Moreover, the clusters do not mix, *i.e.*, two separate clusters remain separated after translation. The next theorem follows from these properties and expresses the cumulant densities of a randomly translated MHP as functions of the translation distributions and the parameters of the MHP.

Theorem 1. *Consider an MHP with excitation matrix function $\boldsymbol{G}(t)$ and exogenous intensity vector $\boldsymbol{\mu} \in \mathbb{R}_+^d$. After random translation of the event set $\boldsymbol{t}$ with distributions $\{F_1(\cdot), \dots, F_d(\cdot)\}$, the resulting event set $\tilde{\boldsymbol{t}}$ has the following cumulants.*

$$K_i = \sum_{m=1}^d \mu_m \int_{\mathbb{R}} \widetilde{R}_{i,m}^{(x)} dx, \tag{4}$$

$$K_{i,j}(\tilde{t}_1, \tilde{t}_2) = \sum_{m=1}^d K_m \int_{\mathbb{R}} \widetilde{R}_{i,m}^{(\tilde{t}_1 - x)} \widetilde{R}_{j,m}^{(\tilde{t}_2 - x)} dx, \tag{5}$$

$$\begin{aligned} K_{i,j,k}(\tilde{t}_1, \tilde{t}_2, \tilde{t}_3) = \\ &\sum_{m,n=1}^d K_n \iint_{\mathbb{R}} \left(\widetilde{R}_{i,n}^{(\tilde{t}_1 - x)} \widetilde{R}_{j,m}^{(\tilde{t}_2 - y)} \widetilde{R}_{k,m}^{(\tilde{t}_3 - y)} \widetilde{\Psi}_{m,n}^{(y-x)} \right) dy dx \\ &+ \sum_{m,n=1}^d K_n \iint_{\mathbb{R}} \left(\widetilde{R}_{j,n}^{(\tilde{t}_2 - x)} \widetilde{R}_{i,m}^{(\tilde{t}_1 - y)} \widetilde{R}_{k,m}^{(\tilde{t}_3 - y)} \widetilde{\Psi}_{m,n}^{(y-x)} \right) dy dx \\ &+ \sum_{m,n=1}^d K_n \iint_{\mathbb{R}} \left(\widetilde{R}_{k,n}^{(\tilde{t}_3 - x)} \widetilde{R}_{i,m}^{(\tilde{t}_1 - y)} \widetilde{R}_{j,m}^{(\tilde{t}_2 - y)} \widetilde{\Psi}_{m,n}^{(y-x)} \right) dy dx \\ &+ \sum_{m=1}^d K_m \int_{\mathbb{R}} \left(\widetilde{R}_{i,m}^{(\tilde{t}_1 - x)} \widetilde{R}_{j,m}^{(\tilde{t}_2 - x)} \widetilde{R}_{k,m}^{(\tilde{t}_3 - x)} \right) dx, \end{aligned} \tag{6}$$

where $\widetilde{\boldsymbol{R}}^{(t)} \coloneqq \sum_{n\geq 0} \widetilde{\boldsymbol{G}}^{*n}(t)$, $\widetilde{\boldsymbol{\Psi}}(t) \coloneqq \widetilde{\boldsymbol{R}}^{(t)} - \boldsymbol{I}\delta(t)$, *and*

$$\begin{aligned}\widetilde{G}_{i,j}(t) &= f_i * G_{i,j} * \underline{f}_j(t) \\ &= \iint_{\mathbb{R}} f_i(t+x-s) G_{i,j}(s) f_j(x) ds dx.\end{aligned} \tag{7}$$

where $f_i(x)dx = dF_i(x)$.

A proof of the theorem is provided in the Appendix. This result shows the relationships between the first, second, and third-order cumulant densities of a randomly translated MHP, the noise distributions, and the parameters of the underlying MHP. Note that Equation (7) implies that the matrices $\boldsymbol{G}(t)$ and $\widetilde{\boldsymbol{G}}(t)$ have the same support. In the next corollary, we further show that their integrated versions, namely $\boldsymbol{G} \coloneqq \mathcal{L}[\boldsymbol{G}](0)$ and $\overline{\boldsymbol{G}} \coloneqq \mathcal{L}[\widetilde{\boldsymbol{G}}](0)$, are equal.

Corollary 1. *Consider an MHP with stationary increments. After a random translation, its corresponding matrix function* $\widetilde{\boldsymbol{R}}^{(t)}$ *given in Theorem 1 is bounded, and*

$$\overline{\boldsymbol{R}} = \left(\boldsymbol{I} - \overline{\boldsymbol{G}}\right)^{-1}, \tag{8}$$

$$\overline{\boldsymbol{G}} = \boldsymbol{G}, \tag{9}$$

where $\overline{\boldsymbol{R}} \coloneqq \mathcal{L}[\widetilde{\boldsymbol{R}}](0)$, *and* $\overline{\boldsymbol{G}} \coloneqq \mathcal{L}[\widetilde{\boldsymbol{G}}](0)$.

We use this equivalence to learn the support of $\boldsymbol{G}$. Note that, given a realization $\tilde{\boldsymbol{t}} = \left\{\left\{\widetilde{t}_k^i\right\}_{k=0}^{n_i}\right\}_{i=1}^{d}$ of a randomly translated MHP, we can estimate the integrated cumulants[2]. In the remainder of this section, we transform the equations (4)-(6) into their integrated forms by evaluating their Laplace transform at $s = 0$ and solve for $\overline{\boldsymbol{R}}$. Corollary 1 can then be applied to obtain $\boldsymbol{G}$. More precisely, let

$$\begin{aligned}\overline{\Psi}_{i,j} &= \mathcal{L}[\widetilde{\Psi}_{i,j}](0), \\ \overline{K}_{i,j} &= \mathcal{L}[K_{i,j}](0), \\ \overline{K}_{i,j,k} &= \mathcal{L}[K_{i,j,k}](0).\end{aligned}$$

Then, the integrated cumulant of a randomly translated MHP can be computed from (4)-(6) as follows.

$$K_i = \sum_{k=1}^{d} \mu_k \overline{R}_{i,k}, \tag{10}$$

$$\overline{K}_{i,j} = \sum_{m=1}^{d} K_m \overline{R}_{i,m} \overline{R}_{j,m}, \tag{11}$$

$$\begin{aligned}\overline{K}_{i,j,k} = &\sum_{m,n=1}^{d} K_n \overline{R}_{i,n} \overline{R}_{j,m} \overline{R}_{k,m} \overline{\Psi}_{m,n} \\ &+ \sum_{m,n=1}^{d} K_n \overline{R}_{j,n} \overline{R}_{i,m} \overline{R}_{k,m} \overline{\Psi}_{m,n} \\ &+ \sum_{m,n=1}^{d} K_n \overline{R}_{k,n} \overline{R}_{i,m} \overline{R}_{j,m} \overline{\Psi}_{m,n} \\ &+ \sum_{m=1}^{d} K_m \overline{R}_{i,m} \overline{R}_{j,m} \overline{R}_{k,m},\end{aligned} \tag{12}$$

where $\overline{\boldsymbol{\Psi}} = \overline{\boldsymbol{R}} - \boldsymbol{I}$.

We would like to emphasize that the above equations are analogous to those of an MHP without random translations given in Jovanović et al. (2015). Together with the fact that $\overline{\boldsymbol{G}} = \boldsymbol{G}$, this implies that the integrated cumulants are invariant with respect to random translations, a key result that will enable to estimate them consistently.

In Equations (10)-(12), the first-order cumulants $\{K_i\}$ and the integrated cumulants $\{\{\overline{K}_{i,j}\}, \{\overline{K}_{i,j,k}\}\}$ can be empirically estimated from the data. These estimates are then used to solve for $\overline{\boldsymbol{R}} = [\overline{R}_{i,j}]$, which yields the underlying causal structure (*i.e.*, the support of $\boldsymbol{G}$) of the randomly translated MHP, via Corollary 1. In the next section, we review two approaches for learning MHPs based on their cumulants and show how exactly they can be adapted to infer the underlying causal structures of randomly translated MHPs.

4.3. Cumulant-based Estimation Methods

4.3.1. The NPHC Algorithm

Achab et al. (2017) proposed the NPHC algorithm, a non-parametric approach inspired by the generalized method of moments. First, note that (11) and (12) provide $(d^2+d)/2$ and $(d^3+3d^2+2d)/6$ independent equations, respectively. The number of unknowns, $\{\{\mu_i\}_{i=1}^d, \overline{\boldsymbol{R}}\}$, is only $d + d^2$. Achab et al. (2017) then select a subset of size d^2 equations out of the group of equations in (12), namely, $\overline{K}_{i,i,j}$ for $1 \leq i, j \leq d$, and use $d^2 + (d^2+d)/2$ equations to obtain the unknowns. The NPHC algorithm works in two steps. First, the integrated cumulants are estimated from the data. Let $\widehat{\overline{\boldsymbol{C}}} \coloneqq [\widehat{\overline{K}}_{i,j}]$ and $\widehat{\overline{\boldsymbol{S}}} \coloneqq [\widehat{\overline{K}}_{i,i,j}]$ denote the estimators of the integrated covariance matrix and skewness matrix,

[2] See Appendix B for the estimators.

respectively. Details of these estimators are provided in Appendix B.

The NPHC estimator for $\overline{\boldsymbol{R}}$ is then defined as the solution of a polynomial optimization problem

$$\widehat{\overline{\boldsymbol{R}}} \in \arg\min_{\overline{\boldsymbol{R}}} \ell_\alpha(\overline{\boldsymbol{R}}),$$

where the objective function is defined as

$$\ell_\alpha(\overline{\boldsymbol{R}}) \coloneqq (1-\alpha)\|\boldsymbol{S}(\overline{\boldsymbol{R}}) - \widehat{\overline{\boldsymbol{S}}}\|_2^2 + \alpha\|\boldsymbol{C}(\overline{\boldsymbol{R}}) - \widehat{\overline{\boldsymbol{C}}}\|_2^2.$$

The weight $\alpha = \|\widehat{\overline{\boldsymbol{S}}}\|_2^2 \Big/ \left(\|\widehat{\overline{\boldsymbol{S}}}\|_2^2 + \|\widehat{\overline{\boldsymbol{C}}}\|_2^2\right)$ balances between the two terms matching the integrated covariance matrix $\boldsymbol{C}(\overline{\boldsymbol{R}}) = [\overline{K}_{i,j}]$ and the integrated skewness matrix $\boldsymbol{S}(\overline{\boldsymbol{R}}) = [\overline{K}_{i,i,j}]$. The authors prove that the NPHC estimator is consistent[3]. Corollary 1 then implies that the NPHC estimator is also consistent for randomly translated MHPs. Therefore, applying the NPHC algorithm to a randomly translated sequence of events will recover the matrix $\overline{\boldsymbol{R}}$ and consequently the integrated excitation matrix $\boldsymbol{G}$.

4.3.2. The Wiener-Hopf Formulation

Another cumulant-based approach for learning MHPs is based on the second-order statistics (Bacry & Muzy, 2016). More precisely, we define the covariance density matrix of an MHP, $\boldsymbol{\Sigma}(t_1, t_2) = [\Sigma_{i,j}(t_1, t_2)]$ as

$$\Sigma_{i,j}(t_1, t_2) \coloneqq K_{i,j}(t_1, t_2) - \frac{\mathbb{E}[dN_i(t_1)]}{dt_1}\epsilon_{i,j}\delta(t_1 - t_2),$$

where $\epsilon_{i,j}$ is the Kronecker symbol, which is always 0 except when $i = j$, in which case it is 1. Hawkes (1971a) proved that $\boldsymbol{\Sigma}(t) \coloneqq \boldsymbol{\Sigma}(t, 0)$ is directly related to the excitation matrix $\boldsymbol{G}(t)$ through the equation

$$\boldsymbol{\Sigma}(t) = \big(\boldsymbol{I}\delta + \underline{\boldsymbol{\Psi}}\big) * \boldsymbol{\Lambda}\big(\boldsymbol{I}\delta + \boldsymbol{\Psi}\big)^T(t) - \boldsymbol{\Lambda}\delta(t),\ \forall t, \quad (13)$$

where $\boldsymbol{\Lambda} \coloneqq \mathrm{diag}([K_1, \dots, K_d])$ is the mean intensity of the stationary process and $\boldsymbol{\Psi}(t) \coloneqq \sum_{n\geq 1} \boldsymbol{G}^{*n}(t)$. Note that this equation does not admit a unique[4] solution with respect to $\boldsymbol{G}(t)$.

Bacry & Muzy (2016) derived the following d^2-dimensional Wiener-Hopf system of equations from (13):

$$\boldsymbol{X}(t) = \boldsymbol{G}(t) + \boldsymbol{G} * \boldsymbol{X}(t),\ \forall t > 0, \quad (14)$$

where $\boldsymbol{X}(t) = \boldsymbol{\Sigma}^T(t)\boldsymbol{\Lambda}^{-1}$ can be estimated from data. The interesting aspect of this equation is that using the fact that $\boldsymbol{G}(t)$ is causal results in a unique solution with respect to $\boldsymbol{G}(t)$. It can therefore be used to infer the excitation matrix $\boldsymbol{G}(t)$ of an MHP from data as done by Bacry & Muzy (2016).

Similar to the aforementioned approach, we can use Theorem 1 to define the covariance density matrix of a randomly translated MHP, and explicit its relation to $\widetilde{\boldsymbol{G}}(t)$ which was defined in (7).

Corollary 2. *Let $\widetilde{\boldsymbol{\Sigma}}(t)$ denotes the covariance density matrix of a randomly translated MHP, defined as*

$$\widetilde{\Sigma}_{i,j}(\tilde{t}_1, \tilde{t}_2) \coloneqq K_{i,j}(\tilde{t}_1, \tilde{t}_2) - \frac{\mathbb{E}[dN_i(\tilde{t}_1)]}{d\tilde{t}_1}\epsilon_{i,j}\delta(\tilde{t}_1 - \tilde{t}_2).$$

Then, for all $t \in \mathbb{R}$,

$$\widetilde{\boldsymbol{\Sigma}}(t) = \big(\boldsymbol{I}\delta + \widetilde{\underline{\boldsymbol{\Psi}}}\big) * \boldsymbol{\Lambda}\big(\boldsymbol{I}\delta + \widetilde{\boldsymbol{\Psi}}\big)^T(t) - \boldsymbol{\Lambda}\delta(t), \quad (15)$$

where $\boldsymbol{\Lambda} = \mathrm{diag}([K_1, \dots, K_d])$ and $\widetilde{\boldsymbol{\Psi}}(t)$ is defined as in Theorem 1.

Similar to (13), Equation (15) does not admit a unique solution with respect to $\widetilde{\boldsymbol{G}}(t)$, but unlike $\boldsymbol{G}(t)$, $\widetilde{\boldsymbol{G}}(t)$ is not a causal function. This is evident from (7) because $\tilde{G}_{i,j}(t)$ is obtained by convolving the causal function $G_{i,j}(t)$ with functions $\{\underline{f}_j(t), f_i(t)\}$ in which at least one is an anti-causal function. This is a major hurdle that was not present in the noiseless case but comes with any non-zero amount of noise. Indeed, it prevents us from obtaining a Wiener-Hopf system of equations from (15) that, like (14), admits a unique solution. Nevertheless, for a small amount of noise, experiments show that we can successfully apply the Wiener-Hopf approach in (Bacry & Muzy, 2016) to randomly translated MHPs and solve the following system to learn $\widetilde{\boldsymbol{G}}(t)$

$$\widetilde{\boldsymbol{X}}(t) = \widetilde{\boldsymbol{G}}(t) + \widetilde{\boldsymbol{G}} * \widetilde{\boldsymbol{X}}(t),\ \forall t > 0, \quad (16)$$

where $\widetilde{\boldsymbol{X}}(t) = \widetilde{\Sigma}^T(t)\Lambda^{-1}$. However, because $\widetilde{\boldsymbol{G}}(t)$ increasingly departs from being causal as the noise power increases, this approach fails to learn the causal structure accurately.

5. Experimental Results

To illustrate the result of Theorem 1 and to characterize the effect of random translations on the estimation of MHPs, we carry out two sets of experiments. First, we simulate a synthetic dataset from an MHP and quantify the ability of two maximum likelihood-based and two cumulant-based approaches for learning the ground-truth excitation matrix under varying levels of noise power. Second, we evaluate the stability of each approach to random translations on a real dataset pertaining to Bund Future traded at Eurex. For reproducibility, a detailed description of the experimental setup is provided in Appendix C. In addition, the open-source code is publicly available on GitHub[5].

[3] For more comprehensive details on the algorithm and its relation to the generalized method of moments, we refer the reader to (Achab et al., 2017).

[4] See the Appendix A for a proof.

[5] https://github.com/trouleau/noisy-hawkes-cumulants

We evaluate the effect of random translation on the following four state-of-the-art approaches.

- NPHC. (Achab et al., 2017) This non-parametric approach is based on matching the empirical integrated cumulants of the events, as discussed in Section 4.3.1.
- WH. (Bacry & Muzy, 2016) This method is a non-parametric approach based on solving a set of Wiener-Hopf equations for learning the excitation functions of the process, as discussed in Section 4.3.2.
- ADM4. (Zhou et al., 2013) This method is a parametric approach that maximizes the log-likelihood function with a sparse and low-rank regularization. It assumes an exponential excitation function of the form $G_{i,j}(t) = \alpha_{i,j}\kappa(t)$, where $\kappa(t) = \beta \exp(-\beta t)$. The exponential decay β is a given hyper-parameter.
- Desync-MLE. (Trouleau et al., 2019) This method is a parametric approach, which maximizes the log-likelihood function of an MHP under synchronization noise, *i.e.*, a particular type of random translation where the noise is assumed to be distributed as $dF_i(x) = \delta(x - z_i)dx$, $\forall i$, such that all events within a dimension are shifted by a constant. This method jointly learns the parameters of the MHP as well as the noise value $\{z_i\}$ using stochastic gradient descent. Similar to ADM4, this approach assumes exponential excitation functions where the exponential decay β is a given hyper-parameter.

5.1. Synthetic Data

We first apply the result of Theorem 1 to a synthetic 10-dimensional MHP ($d = 10$). We considered a non-symmetric block-matrix $\boldsymbol{G}^\star$ depicted in Figure 4(a), with exponential excitation functions $G^\star_{i,j}(t) = \alpha_{i,j}\beta \exp(-\beta t), \forall i, j$, with $\beta = 1$, and baseline intensity $\mu_i = 0.01, \forall i$.

We simulated 20 datasets, each comprised of 5 realizations of 10^5 events. We then randomly translated each dataset with distributions[6] $F_i \sim \mathcal{N}(0, \sigma^2)$, $1 \leq i \leq d$, for varying noise powers σ^2, and we estimated the excitation matrix for the aforementioned approaches. All reported values are averaged over the 20 simulated datasets ($\pm$ standard error).

Figure 4 depicts the estimated integrated excitation matrices for a fixed noise level $\sigma^2 = 5$ for a qualitative visualization of the results. We observe that, while the cumulant-based NPHC method is able to accurately recover the excitation matrix, the maximum likelihood-based ADM4 suffers from both false positives and misses true positives. The covariance-based WH approach is doing a better job than ADM4 but tends to suffer from false positives. This is expected from Corollary 2, as WH wrongly assumes that $\widetilde{\boldsymbol{G}}(t)$ is a causal function.

[6] We also ran experiments with other noise distributions (*i.e.*, exponential and uniform) and observed similar results.

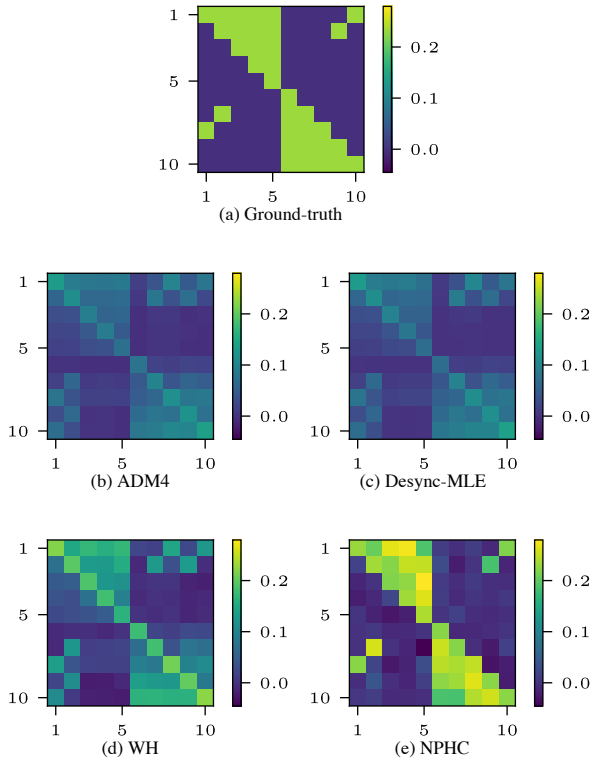


Figure 4: Comparison of the integrated kernel matrix estimated by several methods under noisy observations.

To verify the findings of Theorem 1, we evaluated the sensitivity of the estimators of the integrated cumulants used in NPHC. This pertains to the estimation of the left-hand-side of (5) and (6). In Figure 5, we report the squared $L_{2,2}$ distance of the estimated integrated covariance and skewness matrices to their corresponding ground-truth. As expected, the cumulant estimators remain stable over a large range of noise levels.

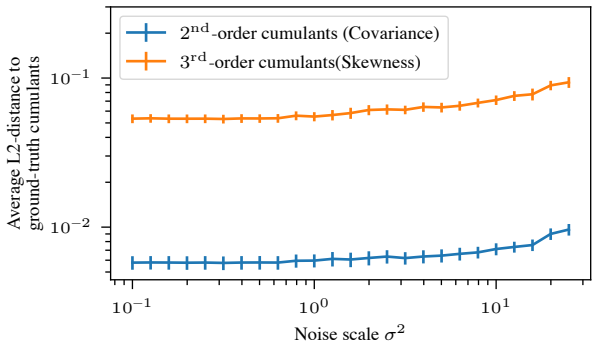


Figure 5: Sensitivity analysis of the integrated cumulants estimators with respect to the noise scale.

To further quantitatively evaluate the sensitivity of each approach to increasing noise levels, we also measured their performance against several metrics for a large range of noise variances σ^2. More specifically, we considered the following metrics.

- **Relative error.** To evaluate the distance between the estimated and the ground-truth values, we computed the averaged relative error defined as

$$\begin{cases} |\widehat{G}_{i,j} - G^\star_{i,j}|/|G^\star_{i,j}|, & \text{if } G^\star_{i,j} > 0, \\ |\widehat{G}_{i,j} - G^\star_{i,j}|/\min_{G^\star_{m,n}\neq 0} |G^\star_{m,n}|, & \text{otherwise.} \end{cases}$$

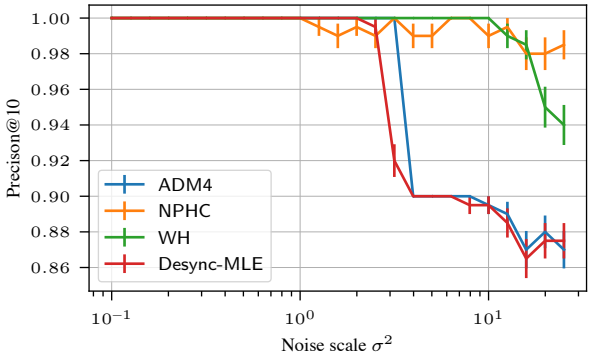
  This metric is more sensitive to errors in small values and therefore penalizes methods with large false positive entries learned in the excitation matrix (Zhou et al., 2013; Figueiredo et al., 2018).

- **Precision@n.** To assess the performance of the approaches at recovering the top entries in $\boldsymbol{G}^\star$, we used precision@n, which is defined as the average fraction of correctly identified entries in the top n largest estimated values. We reported this metric for $n = 10$ (Figueiredo et al., 2018; Salehi et al., 2019).

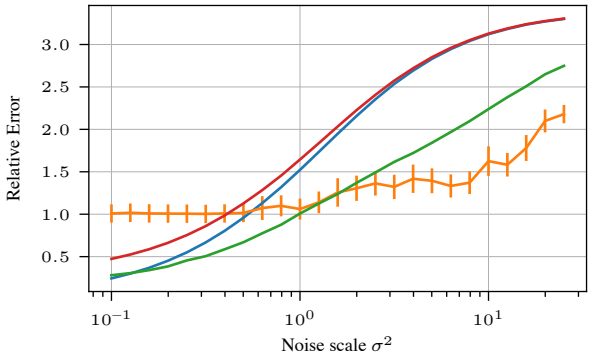
- **PR-AUC.** Considering that there is a Granger-causal link between two dimensions if the learned value $\widehat{G}_{i,j} > \eta$, we evaluate the performance of the resulting binary classification problem using the area under the precision-recall curve over all thresholds $\eta > 0$. Methods that accurately uncover the excitation patterns from the randomly translated data will have a PR-AUC close to 1.

- **$L_{2,2}$ Norm.** We also measured the squared $L_{2,2}$ norm of the estimated excitation matrices, defined as $\|\widehat{\boldsymbol{G}}\|^2_{2,2} = \sum_{i,j} \widehat{G}^2_{i,j}$. Methods that fail to uncover the excitation patterns from the randomly translated data tend to learn an almost-zero matrix with small $L_{2,2}$ norm.

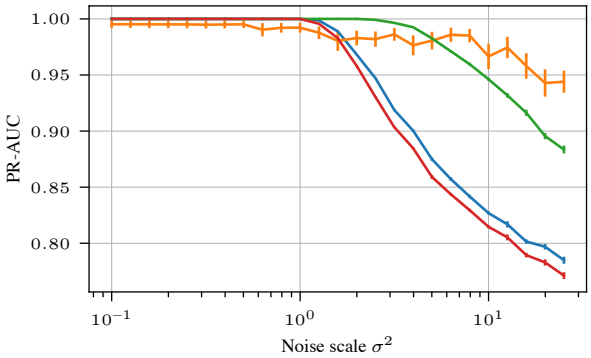
The results are shown in Figure 6. As expected from Corollary 1, the NHPC estimator provides stable estimates for a large range of noise levels. On the other hand, Figure 6d shows that the norm of the matrices estimated by the other approaches tends to zero with increasing σ^2. This is particularly obvious for ADM4 and Desync-MLE. This result is consistent with the findings of (Trouleau et al., 2019) for the special case of synchronized noise. The WH method performs well only for low noise. This is consistent with the observation discussed in Section 4.3.2. This is because, as expected, the non-causal property of $\widetilde{\boldsymbol{G}}(t)$ in randomly translated MHP violates the assumption of WH and hence introduces a bias in the estimation. In smaller noise regime, $\widetilde{\boldsymbol{G}}(t)$ is closer to being causal and as a result the WH method does a better job at learning it.



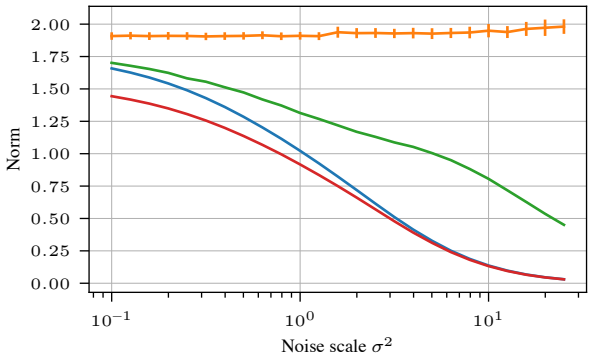
(a) Precision@10



(b) Relative Error



(c) PR-AUC



(d) $L_{2,2}$ Norm

Figure 6: Sensitivity analysis of the estimation methods with respect to the noise scale for synthetic datasets.

5.2. Real Data

We also evaluated the effect of random translations on a publicly available real-world dataset of Bund Futures traded at Eurex[7]. This dataset was already modeled using MHPs in (Bacry et al., 2016) using the WH algorithm. It contains trades performed over 20 days in April 2014. Each event corresponds to one of $d = 4$ types corresponding to the following cases:

- mid-price movement up,
- mid-price movement down,
- buyer initiated trades that do not move the mid-price,
- seller initiated trades that do not move the mid-price.

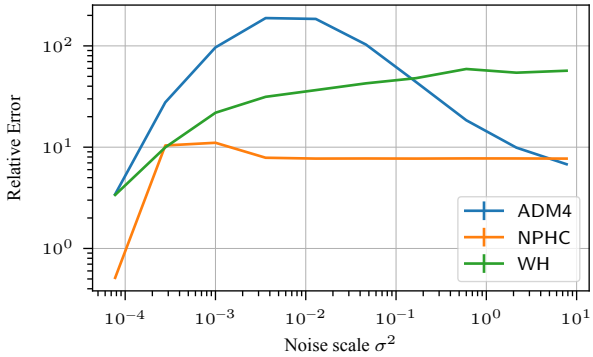
Since there is no ground-truth available for this dataset, we focus our experiments on evaluating the stability of the estimates when a random translation is added to the observations. More precisely, for a large range of noise levels σ^2, we randomly shifted the observed timestamps with distributions $F_i \sim \mathcal{N}(0, \sigma^2)$, and compared the resulting estimated $\widehat{\boldsymbol{G}}_\sigma$ to the noise-free estimate $\widehat{\boldsymbol{G}}_0$ based on the dataset without random translation.

We show the results in Figure 7. We observe that they are consistent with the conclusions reached on the synthetic datasets. ADM4 converges to a zero excitation matrix as the noise scale increases, whereas the cumulant-based approaches, NPHC and WH, remain stable for a wider range of noise levels.

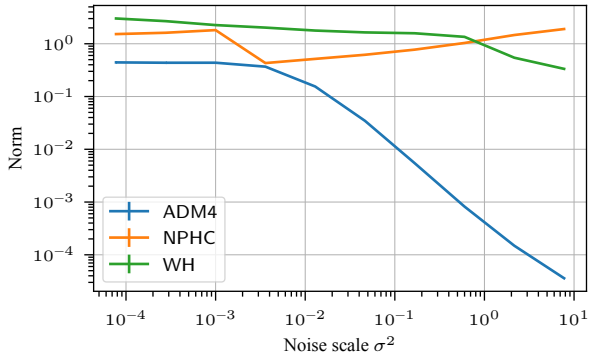
6. Conclusion

In this work, we studied the inference problem of multivariate Hawkes processes from noisy observations. We introduced a general form of observation noise called random translation and proved that the cumulants of MHPs are invariant to such noise. We derived a set of equations for the first, second, and third-order cumulants of a randomly translated MHP with respect to its underlying parameters, namely, the exogenous intensities and the excitation matrix. Using these findings, we showed than cumulant-based estimators are robust to random translations and can accurately learn the causal structure of randomly translated MHPs. In particular, the NPHC estimator remains consistent under randomly translated observations. Because no sample complexity bound was derived for the NPHC estimator, this result only holds asymptotically. However, through extensive experiments, we validated our results and demonstrated that the state-of-the-art inference methods based on maximum-likelihood fail to capture the structure when the observations are affected by random translations.

[7]The dataset is publicly available at: https://github.com/X-DataInitiative/tick-datasets/



(a) Relative Error



(b) $L_{2,2}$ Norm

Figure 7: Sensitivity analysis of the estimation to noise scale for the Bund Futures traded at Eurex.

References

Abergel, F. and Jedidi, A. Long-time behavior of a hawkes process–based limit order book. *SIAM Journal on Financial Mathematics*, 6(1):1026–1043, 2015. doi: 10.1137/15M1011469. URL https://doi.org/10.1137/15M1011469.

Achab, M., Bacry, E., Gaïffas, S., Mastromatteo, I., and Muzy, J.-F. Uncovering causality from multivariate hawkes integrated cumulants. *The Journal of Machine Learning Research*, 18(1):6998–7025, 2017.

Bacry, E. and Muzy, J.-F. First-and second-order statistics characterization of hawkes processes and non-parametric estimation. *IEEE Transactions on Information Theory*, 62(4):2184–2202, 2016.

Bacry, E., Dayri, K., and Muzy, J.-F. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):1–12, 2012.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1 (01):1550005, 2015.

Bacry, E., Jaisson, T., and Muzy, J. Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, 16(8):1179–1201, 2016. doi: 10.1080/14697688.2015.1123287. URL https://doi.org/10.1080/14697688.2015.1123287.

Bar-Hen, A., Chadœuf, J., Dessard, H., and Monestiez, P. Estimating second order characteristics of point processes with known independent noise. *Statistics and Computing*, 23(3):297–309, May 2013. ISSN 1573-1375. doi: 10.1007/s11222-011-9311-7. URL https://doi.org/10.1007/s11222-011-9311-7.

Colizza, V., Barrat, A., Barthélemy, M., and Vespignani, A. Predictability and epidemic pathways in global outbreaks of infectious diseases: the sars case study. *BMC medicine*, 5(1):1–13, 2007.

Cucala, L. Intensity estimation for spatial point processes observed with noise. *Scandinavian Journal of Statistics*, 35:322–334, 06 2008. doi: 10.1111/j.1467-9469.2007.00583.x.

Da Fonseca, J. and Zaatour, R. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *Journal of Futures Markets*, 34(6):548–579, 2014.

Daley, D. J. and Jones, D. V. *An Introduction to the Theory of Point Processes: Elementary Theory of Point Processes*. Springer, 2003.

Deutsch, I. and Ross, G. J. Abc learning of hawkes processes with missing or noisy event times, 2021.

Doob, J. L. *Stochastic processes*, volume 101. New York Wiley, 1953.

Eichler, M., Dahlhaus, R., and Dueck, J. Graphical modeling for multivariate Hawkes processes with nonparametric link functions. *Journal of Time Series Analysis*, 38(2):225–242, 2017.

Etesami, J., Kiyavash, N., Zhang, K., and Singhal, K. Learning network of multivariate hawkes processes: A time series approach. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pp. 162171, Arlington, Virginia, USA, 2016. AUAI Press. ISBN 9780996643115.

Figueiredo, F., Borges, G. R., de Melo, P. O. V., and Assunção, R. Fast estimation of causal interactions using wold processes. In *Advances in Neural Information Processing Systems*, pp. 2971–2982, 2018.

Gerhard, F., Deger, M., and Truccolo, W. On the stability and dynamics of stochastic spiking neuron models: Nonlinear hawkes process and point process glms. *PLOS Computational Biology*, 13(2):1–31, 02 2017. doi: 10.1371/journal.pcbi.1005390. URL https://doi.org/10.1371/journal.pcbi.1005390.

Gomez-Rodriguez, M., Balduzzi, D., and Schölkopf, B. Uncovering the temporal dynamics of diffusion networks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 561568, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

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

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

Hoffmann, J. and Caramanis, C. Learning graphs from noisy epidemic cascades. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(2), June 2019. doi: 10.1145/3341617.3326155. URL https://doi.org/10.1145/3341617.3326155.

Jovanović, S., Hertz, J., and Rotter, S. Cumulants of hawkes point processes. *Physical Review E*, 91(4):042802, 2015.

Kim, M., Paini, D., and Jurdak, R. Modeling stochastic processes in disease spread across a heterogeneous social system. *Proceedings of the National Academy of Sciences*, 116(2):401–406, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1801429116. URL https://www.pnas.org/content/116/2/401.

Liniger, T. J. *Multivariate Hawkes processes*. PhD thesis, Eidgenössische Technische Hochschule ETH Zürich, 2009.

Lukacs, E. *Characteristic functions*. Griffin, 1970.

Ozaki, T. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155, 1979.

Reynaud-Bouret, P., Rivoirard, V., and Tuleau-Malot, C. Inference of functional connectivity in neurosciences via hawkes processes. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 317–320, 2013. doi: 10.1109/GlobalSIP.2013.6736879.

Salehi, F., Trouleau, W., Grossglauser, M., and Thiran, P. Learning hawkes processes from a handful of events. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 12715–12725. Curran Associates, Inc., 2019.

Shelton, C. R., Qin, Z., and Shetty, C. Hawkes process inference with missing data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Trouleau, W., Etesami, J., Grossglauser, M., Kiyavash, N., and Thiran, P. Learning Hawkes processes under synchronization noise. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6325–6334. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/trouleau19a.html.

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, 93(2):1074–1089, 2005. doi: 10.1152/jn.00697.2004. URL https://doi.org/10.1152/jn.00697.2004. PMID: 15356183.

Xu, H., Farajtabar, M., and Zha, H. Learning Granger causality for Hawkes processes. *International Conference on Machine Learning*, 48:1717–1726, 2016.

Xu, H., Luo, D., and Zha, H. Learning hawkes processes from short doubly-censored event sequences. *arXiv preprint arXiv:1702.07013*, 2017.

Yang, Y., Etesami, J., He, N., and Kiyavash, N. Online learning for multivariate Hawkes processes. *Neural Information Processing Systems*, 2017.

Zhou, K., Zha, H., and Song, L. Learning triggering kernels for multi-dimensional Hawkes processes. In *International Conference on Machine Learning*, volume 28, pp. 1301–1309, 2013.