Local Independence Testing for Point Processes

Nikolaj Thams and Niels Richard Hansen

***Abstract*— Constraint-based causal structure learning for point processes require empirical tests of local independence. Existing tests require strong model assumptions, e.g., that the true data generating model is a Hawkes process with no latent confounders. Even when restricting attention to Hawkes processes, latent confounders are a major technical difficulty because a marginalized process will generally not be a Hawkes process itself. We introduce an expansion similar to Volterra expansions as a tool to represent marginalized intensities. Our main theoretical result is that such expansions can approximate the true marginalized intensity arbitrarily well. Based on this, we propose a test of local independence and investigate its properties in real and simulated data.**

***Index Terms*— Causal discovery, local independence, neuroscience, point processes.**

I. Introduction

HAWKES processes are models of time-dynamic interacting point processes with applications in such diverse areas as finance [1], seismology [2], social science [3], and neuroscience [4]. Hawkes proposed himself that his model for self- and mutually exciting point processes could be applied as a model of epidemic spread and neuron firing among other things [5], and with reference to Hawkes' pivotal work the model has taken the name *Hawkes process* in the literature. Specifically, Hawkes introduced the multivariate linear Hawkes process, which together with its nonlinear extension [6] have become the most widely applied models of multivariate dynamic point processes.

It is straightforward to define—in purely mathematical terms—whether one event type in a Hawkes process affects another event type. This defines a network, and our main objective is to test hypotheses regarding network connectivity. Constraining the network structure to be sparse can have well-known statistical and computational benefits, e.g., a favorable bias-variance tradeoff for large networks and fast data-fitting algorithms [7], [8]. However, it is much less obvious if the network structure allows for a subject matter interpretation beyond the purely statistical one. In particular, if the network conveys causal information.

We will use Hawkes process models of neuron spike activity as a main motivating example, and we will discuss the question of causal discovery in this context; nonetheless, our proposed methodology does not rely on parametric assumptions, such as the process being Hawkes, and can equally well be applied to other point process model classes. Hawkes processes have a long history in neuron science with Brillinger using them some 45 years ago for the first time [9], [10]. Early applications relied on moment identities and spectral methods, but likelihood methods later became computationally feasible and widely used [4], [11], [12], [13], [14], [15]. The Hawkes processes have served several objectives, from a statistical characterization of dependencies among correlated neurons to a vehicle for sensory decoding from neuron ensembles, and, more recently, as a way to learn a sparse network structure among the neurons [16], [17]. According to [17] the Hawkes process can identify the *functional connectivity* of a neural network, but the network "cannot be directly interpreted as synaptic connections"—yet the model's attractiveness was from the very beginning tied to its physiological interpretability as representing synaptic integration [10]. Moreover, functional connectivity was interpreted in [17] as a *causal relation*, and understanding the extent to which this interpretation is justified was a main motivation for the work presented in this article.

The methods proposed by [17], as well as most methods in the statistical [7], [8], [18], [19], [20], [21], result in networks that only enables identification of intervention effects [22], [23], [24] when assuming that all variables of the system are observed. Causal structure learning algorithms like the causal analysis (CA) algorithm [25] likewise require all variables observed, but recent constraint-based learning algorithms [26], [27] do allow for an interventional causal interpretation of the resulting network even in the presence of latent confounders. The algorithm by [26] is related to FCI for acyclic causal structures [28], but it is adapted to cyclic graphs that can represent time-dynamic feedback mechanisms. Where FCI and other algorithms for acyclic graphs are relying on tests of conditional independence, cyclic graphs of time-dynamic systems are based on (conditional) *local independence* [22], [27], and causal discovery algorithms require empirical tests of this asymmetric independence relation. This has the additional benefit that, contrary to works exploiting the Hawkes assumption to learn causal graphs [8], [29], constraint-based methods apply to any model class for which the local independence test captures causal dependencies. Very recently, Bhattacharjya et al. [30] proposed a likelihood-ratio test in the model class of "proximal event" models, though until now, no test for local independence in neither Hawkes processes nor general point processes has been described in the literature.

In this article, we propose a test of local independence in point process data. Let j and k denote two types of events,

Manuscript received 22 October 2021; revised 25 July 2022 and 8 January 2023; accepted 4 November 2023. Date of publication 18 December 2023; date of current version 5 April 2024. This work was supported in part by the Novo Nordisk Foundation under Grant NNF20OC0062897 and in part by VILLUM FONDEN under Grant 18968. *(Corresponding author: Nikolaj Thams.)*

The authors are with the Department of Mathematics, University of Copenhagen, 2100 Copenhagen, Denmark (e-mail: thams@math.ku.dk; niels.r.hansen@math.ku.dk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2023.3335265, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2023.3335265

e.g., the firing of two different neurons, and let C denote a set of event types, e.g., a set of neurons. The hypothesis that k is locally independent of j given C is denoted $j \not\to k|C$. We test the hypothesis by testing whether events of type j contribute significantly to the intensity of k given events of type C. We approximate point process intensities by basis expansions and propose to use higher order interactions terms of events to fit intensities, such that the intensity does not only take into account single events (as is the case for Hawkes processes), but also pairs or triples of events. We show that higher order interactions can be captured through iterated integrals and that any intensity can be arbitrarily well approximated by including enough higher order terms, analogous to Volterra expansions in dynamical systems [31], [32].

Our main motivation for this nonparametric expansion is that Hawkes processes are not closed under marginalization, meaning that a subcollection of event types of a Hawkes process need not be a Hawkes process. Consequently, if some event types of a Hawkes process are unobserved, we may not be able to model it by a Hawkes process (that is, using only first-order terms of events). Even if all processes are observed, constraint-based learning algorithms [25], [26] construct a local independence graph by testing $j \not\to k|C$ within a (typically small) subcollection of event types, in effect corresponding to testing local independence when marginalizing away everything else than j, k and C. The model misspecification arising from the assumption that the marginalized processes are Hawkes may result in tests that do not have asymptotic level. By including higher order interactions in our tests, this model misspecification is reduced, such that the null hypotheses of local independence are rejected less often, resulting in sparser and more correct graphs.

Our main contributions are as follows.

1) In Section II, we discuss local independence testing in point processes, and in particular the challenge of model misspecification due to marginalization. We propose an approximation (5) that allows us to test local independence without model misspecification under the null hypothesis.
2) In Section III, we develop an expansion of point process intensities via iterated integrals, and show that it converges to the true point process intensity. The expansion does not rely on parametric assumptions, and can be used to approximate marginalized intensities from any model class including Hawkes processes.
3) Finally, we propose a concrete test based on a second-order approximation and basis splines, and show in real and synthetic data that compared to using first terms only the use of second-order interactions improves performance of the local independence tests.

A. Structure of This Article

In Section II, we outline the existing theory on Hawkes processes and local independence. Section III contains our main theoretical result, that intensities can be approximated arbitrarily well by including higher order interaction terms. We apply this approximation in Section IV to construct a test of local independence. In Section V, we evaluate the test in simulation studies, and in Section VI we apply the test in causal learning algorithms to learn network structure in a neuron spiking dataset.

II. Hawkes Processes and Local Independence

In this section, we first give a brief introduction to Hawkes processes (see [33] for a more thorough introduction). We then introduce local independence graphs and tests of local independence.

Let $V = \{1, \dots, d\}$ and let $N = (N^k)_{k \in V}$ denote a collection of point processes on $\mathbb{R}$ indexed by V. Each mark $k \in V$ represents a particular type of event, and N is also referred to as a marked point process [33]. More formally, if for each $k = 1, \dots, d$, we let $\{\dots, \tau_{-1}^k, \tau_0^k, \tau_1^k, \tau_2^k, \dots\}$ be a series of event times, the point process N^k is defined as the random measure

$$N^k(A) = \sum_i \delta_{\tau_i^k}(A)$$

where $\delta_t(A)$ is a Dirac measure. We associate with the kth point process the counting process $N_t^k = N^k((0, t])$. We will assume that N is simple and non-exploding, meaning that event times are distinct and any finite interval only has finitely many points.

For a point process N, the intensity $\lambda_t = (\lambda_t^1, \dots, \lambda_t^d)$ describes the conditional rate of new events at time t

$$\lambda_t^k = \mathbb{E}\big(N^k(\mathrm{d}t)|\mathcal{F}_{t-}^V\big)$$

where $\mathcal{F}_{t-}^V$ is the predictable filtration generated by $N^1, \dots, N^d$, i.e., $\mathcal{F}_{t-}^V$ is the history of events strictly prior to time t of any type $j \in V$.

Let $g^{jk}{:}[0, \infty) \to [0, \infty)$ for $j, k \in V$ be integrable functions, which we call kernels. We introduce the intensity process

$$\begin{aligned}\lambda_t^k &= \beta_0^k + \sum_{j \in V} \sum_{i:\tau_i^j < t} g^{jk}\left(t - \tau_i^j\right) \\ &= \beta_0^k + \sum_{j \in V} \int_{-\infty}^{t-} g^{jk}(t - s) N^j(\mathrm{d}s) \end{aligned} \tag{1}$$

where we call $\beta_0^k \geq 0$ baseline intensities.

Definition 1: A d-dimensional point process with intensity processes λ^k, $k \in V$, as defined by (1) is called a multivariate linear Hawkes process with kernels g^{jk} and baseline intensities β_0^k.

In this article, we only consider stationary Hawkes processes. If we define

$$\mathbf{g}^{jk} = \int_0^\infty g^{jk}(t)\mathrm{d}t \tag{2}$$

and define the matrix G as the matrix where the the (j, k)th entry is $\mathbf{g}^{jk}$, stationarity for the linear Hawkes process is achieved if the spectral radius of G is strictly smaller than 1, see [33, Ch. 6] and [34].

The linear Hawkes process can be extended to *the nonlinear Hawkes process* using a link function η

$$\eta\left(\lambda_t^k\right) = \beta_0^k + \sum_{j \in V} \int_{-\infty}^{t-} g^{jk}(t-s) N^j(\mathrm{d}s).$$

Useful alternatives to $\eta(x) = x$ are $\eta(x) = \log(x)$ or $\eta(x) = 1_{x \geq 1} \cdot x + 1_{x<1} \cdot (\log(x) + 1)$. In both cases, η^{-1} maps $\mathbb{R}$ into $[0, \infty)$, which ensures that $\lambda_t^k \geq 0$ even if the kernels are allowed to take negative values. In the following subsection, we only discuss marginalization in the linear Hawkes process. However, the approximation result in Section III extends readily to nonlinear processes, and so we state that result in generality.

A. Marginalization in Hawkes Processes

If we only observe events corresponding to marks in $V' \subset V$, the distribution of the V'-events is a marginalization of the distribution of V-events. Even if all event types of a system are observed, the local independence statement $j \not\to k|C$ relates to the marginal distribution of N^j, N^k and N^C, so when $\{j, k\} \cup C \neq V$, we test local independence in a marginalized distribution.

This creates a problem for testing, because many model classes, including Hawkes processes, are not closed under marginalization, i.e., the marginalized distribution need not be in the same model class as the original distribution. We explore the case of marginalized Hawkes processes in more detail.

For $C \subseteq V$ let $\mathcal{F}_{t-}^C := \sigma(\cup_{s<t} \mathcal{F}_s^C)$ denote the predictable filtration generated by N^j for $j \in C$, and let $E(\cdot|\mathcal{F}_{t-}^C)$ denote expectations given only information about events strictly prior to t of types C.[1] Suppose that $k \in C$, then by the innovation theorem (see [36]) the $\mathcal{F}_{t-}^C$-intensity of N^k is

$$\lambda_t^{k,C} = E\left(\lambda_t^k \middle| \mathcal{F}_{t-}^C\right).$$

We will refer to this as the C-intensity. For the linear Hawkes process, we have from (1) that

$$\lambda_t^{k,C} = \beta_0^k + \sum_{j \in C} \int_{-\infty}^{t-} g^{jk}(t-s) N^j(\mathrm{d}s) + \sum_{l \in C^c} \int_{-\infty}^{t-} g^{lk}(t-s) E\left(N^l \middle| \mathcal{F}_{t-}^C\right)(\mathrm{d}s) \quad (3)$$

thus for a complete computation of the C-intensity we need to compute $E(N^l|\mathcal{F}_{t-}^C)$, which is a classical filtering problem. The solution can be characterized via general filtration equations, see [37] and [38].

We could approximate the solution of the filtering problem by a linear filter

$$E\left(N^l \middle| \mathcal{F}_{t-}^C\right)(\mathrm{d}s) \simeq \left(\gamma_0^l + \sum_{j \in C} \int_{-\infty}^{t-} h^{jl}(t-u, t-s) N^j(\mathrm{d}u)\right) \mathrm{d}s$$

for a choice of kernels h^{jl}. Using this, we arrive at the following *approximate* C-intensity:

$$\tilde{\lambda}_t^{k,C} = \tilde{\beta}_0^{k,C} + \sum_{j \in C} \int_{-\infty}^{t-} \tilde{g}^{jk,C}(t-s) N^j(\mathrm{d}s) \quad (4)$$

where

$$\tilde{\beta}_0^{k,C} = \beta_0^k + \sum_{l \in C^c} \int_0^\infty \gamma_0^l g^{lk}(t) \mathrm{d}t$$

and

$$\tilde{g}^{jk,C}(t) = g^{jk}(t) + \sum_{l \in C^c} \int_{0+}^\infty h^{jl}(t,s) g^{lk}(s) \mathrm{d}s$$

for $j \in C$. We recognize (4) as being the intensity for a linear Hawkes process over event types indexed by C. However, this is only an approximation, and the marginalized process will generally not be a linear Hawkes process. Thus some effects of this *model misspecification* should be expected if we fit a model of the form (4) to marginalized data.

B. Local Independence Hypotheses

Following [27], we define local independence for a point process N by saying that N^k is locally independent of N^j given N^C if $\lambda^{k,C\cup\{j\}}$ has an $\mathcal{F}_t^C$-predictable version. Intuitively that means that $\lambda^{k,C\cup\{j\}}$ only depends on events in N^C and not N^j. In this case, we write $j \not\to k|C$, and else (if $\lambda_t^{k,C}$ is not a version of $\lambda_t^{k,C\cup\{j\}}$) we write $j \to k|C$.

Our goal is to test the local independence hypothesis

$$H_0 : j \not\to k|C.$$

In the approximate C-intensity from (4) this hypothesis corresponds to $\tilde{g}^{jk,C\cup\{j\}}$ being 0. However, a test of $\tilde{g}^{jk,C\cup\{j\}} = 0$ as a surrogate for H_0 comes with no guarantee on the level due to the model misspecification of $\tilde{\lambda}^{k,C}$.

Instead of relying on $\tilde{\lambda}^{k,C\cup\{j\}}$, the first-order approximation in (4) for $\lambda^{k,C\cup\{j\}}$, we propose to base the test on the approximation

$$\overline{\lambda}_t^{k,C\cup\{j\}} := \lambda_t^{k,C} + \int_{-\infty}^{t-} \overline{g}^{jk}(t-s) N^j(d_C s) \quad (5)$$

of the $C \cup \{j\}$-intensity $\lambda^{k,C\cup\{j\}}$. While $\tilde{\lambda}^{k,C\cup\{j\}}$ uses a first-order approximation for all processes, $\overline{\lambda}^{k,C\cup\{j\}}$ uses the actual $\lambda^{k,C}$ intensity along a first-order approximation of the contribution from N^j. Under H_0:$j \not\to k|C$, where $\lambda^{k,C\cup\{j\}}$ and $\lambda^{k,C}$ coincide,[2] $\overline{\lambda}^{k,C\cup\{j\}}$ is in fact correctly specified. A major practical and technical challenge is to approximate and fit $\lambda^{k,C}$ sufficiently well for the test to maintain level, and we dedicate Section III to developing methods for appropriately fitting $\lambda^{k,C}$.

[1] Technically, $E(\cdot|\mathcal{F}_{t-}^C)$ is the predictable projection operator to have regular sample paths of the resulting stochastic process. See the remark in [35] for a discussion of this.

[2] More formally $\lambda^{k,C}$ is a version of $\lambda^{k,C\cup\{j\}}$.

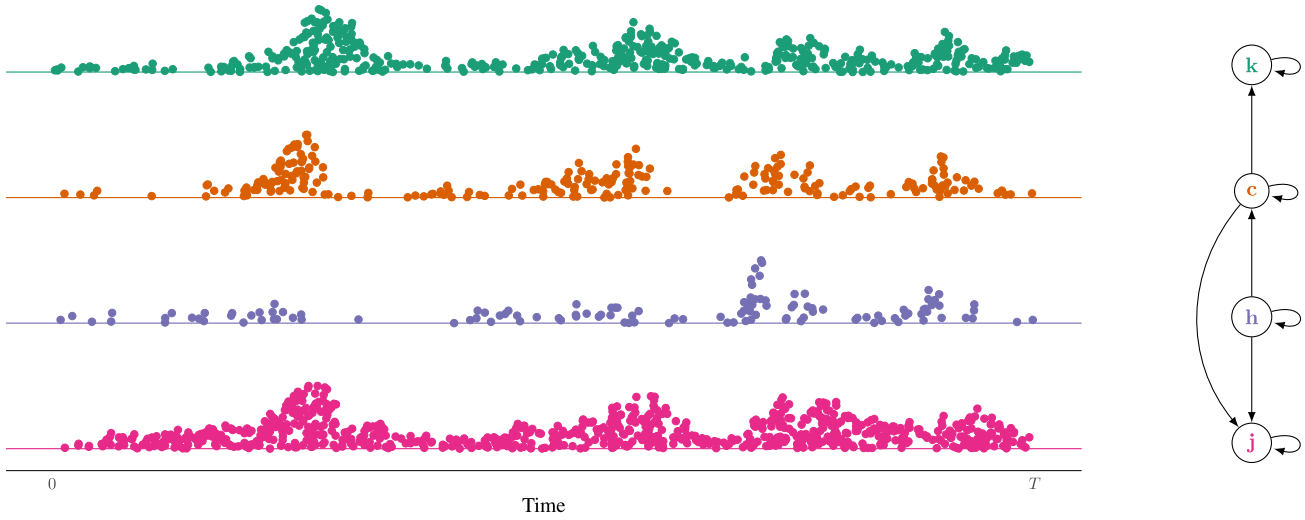

Fig. 1. (*left*): Data from a 4-D Hawkes process. The vertical position of points reflect the local frequency of points. (*right*): Local independence graph (see Section II-C) of the process that generated the data.

C. Local Independence Graphs

In this article, we consider tests for local independence, with the motivation of learning graphical representations of causal relations in point processes. In particular, we consider the local independence graph for point processes, introduced by Didelez in [22], where the absence of an edge $j \not\to k$ in the graph corresponds to the local independence $j \not\to k | V \setminus \{j, k\}$.

For the linear Hawkes process, the local independence graph is a graph with vertices V and an edge $j \to k$ if and only if $\mathbf{g}^{jk} > 0$, where $\mathbf{g}^{jk}$ is defined in (2). That is, there is an edge from j to k if and only if the kernel g^{jk} is not constantly equal to 0. Fig. 1 displays data from a Hawkes process and the underlying local independence graph that was used to generate the data.

III. Higher Order Expansions

A. Motivating Higher Order Interactions

In the following, we propose a general expansion of point process intensities, which we show to converge to the true intensity as the degree of the expansion approaches infinity. We intend to apply this to marginalized Hawkes processes, to remove the model misspecification discussed above, but the result does not rely on the process being Hawkes, and applies to any point process model.

The expansion utilizes iterated integrals, which already [9] used for specifying models with higher order interactions. Cohen [39] showed that the chaos expansion of point processes initiated at zero can approximate any measurable variable arbitrarily well, by integrals over random intervals. Similar to [39] our proof relies on martingale convergence, but uses integrals over deterministic intervals.

Iterated integrals are also used in the theory of Volterra series [31], where the dynamics of a time-homogeneous system over variables x and y is approximated by the Lth-*order* expansion

$$y_t \approx \beta^0 + \sum_{n=1}^{L} \int_{-\infty}^{t} \cdots \int_{-\infty}^{t} h^n \times (s_1, \ldots, s_n) x_{t-s_1} \cdots x_{t-s_n} \, \mathrm{d}s_1 \cdots \mathrm{d}s_n.$$

Under various regularity conditions, including continuity and finite memory of the system, this approximation will converge, that is, the right-hand side converges to y_t for all t when L tends to infinity [32], [40]. Although point process systems are very different in nature to continuous systems, we show a similar expansion for point processes below.

B. Intensity Representations

We consider a fixed subset $C \subseteq V$, and a stationary process N. Let C_n be the set of tuples $\alpha = (j_1, \ldots, j_n)$ of length n where $j_i \in C$ and $j_{i_1} \leq j_{i_2}$ for $i_1 < i_2$. Further define

$$E_n = \mathcal{L}\big([0, \infty)^n, \mathbb{R}\big)^{C_n}$$

where $\mathcal{L}([0, \infty)^n, \mathbb{R})$ is the set of measurable functions $h : [0, \infty)^n \to \mathbb{R}$. That is, every element $(h^\alpha)_{\alpha \in C_n}$ in E_n is a collection of functions, indexed by the distinct combinations of C. We introduce the notation φ_n^t as the functional that maps a collection of functions $(h^\alpha)_{\alpha \in C_n}$ to the integral over those functions with respect to N

$$\varphi_n^t : (h^\alpha)_{\alpha \in C_n} \mapsto \sum_{\alpha \in C_n} \int_{(-\infty, t)} h^\alpha\big(t - s^n\big) N^\alpha\big(\mathrm{d}s^n\big)$$

where

$$\int_{(-\infty, t)} h\big(t - s^n\big) N^\alpha\big(\mathrm{d}s^n\big) := \int_{-\infty}^{t-} \cdots \int_{-\infty}^{t-} h(t - s_1, \ldots, t - s_n) N^{j_1}(\mathrm{d}s_1) \cdots N^{j_n}(\mathrm{d}s_n).$$

Note that φ_n^t maps into $\mathcal{L}(\mathcal{F}_{t-}^C)$ because for any $h := (h^\alpha)_{\alpha \in C_n}$, the filter $\varphi_n^t(h)$ is $\mathcal{F}_{t-}^C$-measurable.

For example, if $V = \{1, 2\}$ and $n = 2$, we have $C_n = \{(1, 1), (1, 2), (2, 2)\}$, $E_n = \{(h^{(1,1)}, h^{(1,2)}, h^{(2,2)}) | h^{(i,j)}: [0, \infty)^2 \to \mathbb{R}$ measurable$\}$ and

$$\begin{aligned}\varphi_n^t\big(h^{(1,1)}, h^{(1,2)}, h^{(2,2)}\big) &= \int_{-\infty}^{t-}\int_{-\infty}^{t-} h^{(1,1)}(t-s_1, t-s_2)N^1(\mathrm{d}s_1)N^1(\mathrm{d}s_2) \\ &+ \int_{-\infty}^{t-}\int_{-\infty}^{t-} h^{(1,2)}(t-s_1, t-s_2)N^1(\mathrm{d}s_1)N^2(\mathrm{d}s_2) \\ &+ \int_{-\infty}^{t-}\int_{-\infty}^{t-} h^{(2,2)}(t-s_1, t-s_2)N^2(\mathrm{d}s_1)N^2(\mathrm{d}s_2)\end{aligned}$$

is the evaluation of the kernels $h^{(1,1)}, h^{(1,2)}$ and $h^{(2,2)}$ in all combinations of points in the respective event types N^1 and N^2.

We now show that we can approximate point process intensities by such sums of iterated integrals. We first show this for $t = 0$ and then extend the result to all $t \in \mathbb{R}$ using time homogeneity. At $t = 0$, we define the set W_n of all $\mathcal{F}_{0-}^C$-measurable random variables, that can be written as a n-fold iterated integral and are almost surely finite

$$W_n = \big\{X \in \varphi_n^0(E_n) || X| < \infty \text{ a.s}\big\}.$$

This allows us to state the following theorem, which is proven in the appendix in the Supplementary Material.

Theorem 1: With $\mathcal{F}_{0-}^C = \sigma(\cup_{s<0}\mathcal{F}_s^C)$ it holds that $\oplus_{n\in\mathbb{N}} W_n$ is dense in $\{X \in \mathcal{L}(\mathcal{F}_{0-}^C) || X| < \infty \text{ a.s.}\}$ in the topology of convergence in probability.[3]

That is, every finite $\mathcal{F}_{0-}^C$-measurable variable can be approximated arbitrarily well by iterated integrals, over the past events of the processes in C.

Consider now the case of a point process intensity $\lambda_t^{k,C}$, and let η denote a link function. Assume further that the intensity is time homogeneous: if $\eta(\lambda_t^{k,C})(\tau_1, \tau_2, \ldots)$ denotes the mechanism with which $\eta(\lambda_t^{k,C})$ depends on the event times prior to time t, we say that $\eta(\lambda_t^{k,C})$ is time-homogeneous if for $s \geq 0$

$$\eta\big(\lambda_t^{k,C}\big)(\tau_1, \tau_2, \ldots) = \eta\Big(\lambda_{t-s}^{k,C}\Big)(\tau_1 - s, \tau_2 - s, \ldots).$$

Corollary 1: If $\eta(\lambda_t^{k,C})$ is a time homogeneous point process intensity, $\eta(\lambda_t^{k,C})$ can at all times be arbitrarily well approximated by iterated integrals in the topology of convergence in probability.

Proof: Take $\epsilon > 0$ and any $t \in \mathbb{R}$. Since the intensity is $\mathcal{F}_t^C$-predictable, $\eta(\lambda_0^{k,C}) \in \mathcal{F}_{0-}^C$ at time $t = 0$. Thus take $\phi^0 \in \oplus_{n\in\mathbb{N}} W_n$ such that $P(|\eta(\lambda_0^{k,C}) - \phi^0| > \epsilon) < \epsilon$, which is possible by Theorem 1. Since $\oplus_{n\in\mathbb{N}} W_n$ is a sum of images, we can choose $h_1 \in E_1, h_2 \in E_2, \ldots, h_L \in E_L$ such that $\phi^0 = \sum_{n=1}^L \varphi_n^0(h_n)$. Let ϕ be the process $t \mapsto \sum_{n=1}^L \varphi_n^t(h_n)$, and observe that ϕ is time homogeneous.

Conclusively, the process $\eta(\lambda_t^{k,C}) - \phi^t$ is time homogeneous, and by the assumed stationarity, the distribution of $\eta(\lambda_t^{k,C}) - \phi^t$ is invariant over t. In particular $P(|\eta(\lambda_t^{k,C}) - \phi^t| > \epsilon) < \epsilon$ for all $t \in \mathbb{R}$. □

[3] i.e., the topology induced by the Ky Fan metric

$$d(X, Y) = \inf\{\epsilon > 0 | P(|X - Y| > \epsilon) \leq \epsilon\}.$$

Observe that it is the same kernels $h_1, \ldots, h_L$ that enter into the approximation of $\lambda_t^{k,C}$ for all t. In Theorem 1, there is nothing special about $t = 0$, and one could as well have proven that $\oplus_{n\in\mathbb{N}} \varphi_n^t(E_n)$ is dense in $\mathcal{L}(\mathcal{F}_{t-}^C)$. However, only by the time-homogeneity can one be ensured that the same kernels can be used for all t.

C. Approximate Intensities

The fully observed (nonlinear) Hawkes process has intensity given by sums of first-order terms

$$\eta\big(\lambda_t^k\big) = \eta\big(\lambda_t^{k,V}\big) = \beta^0 + \sum_{j\in V}\int_{-\infty}^{t-} g^{jk}(t-s)N^j(\mathrm{d}s).$$

As discussed in Section II-A, when $C \neq V$, $\lambda^{k,C}$ cannot in general be represented by sums of first-order terms. However, by Theorem 1 the intensity can be approximated by including interaction terms of higher orders, and so one could approximate $\lambda^{k,C}$ by the Lth-*order* expansion

$$\eta\big(\lambda_t^{k,C}\big) \approx \beta_0^k + \sum_{n=1}^L \sum_{\alpha\in C_n}\int_{(-\infty,t)} h_n^\alpha\big(t - s^n\big)N^\alpha\big(\mathrm{d}s^n\big)$$

for some sequence of kernels h_n^α, $1 \leq n \leq L$, $\alpha \in C_n$. For $L = 2$, we obtain the approximate intensity

$$\begin{aligned}\eta\big(\lambda_t^{k,C}\big) \approx \beta_0^k &+ \sum_{j_1\in C}\int_{-\infty}^{t-} h^{j_1}(t-s_1)N^{j_1}(\mathrm{d}s_1) \\ &+ \sum_{j_1,j_2\in C}\int_{-\infty}^{t-}\int_{-\infty}^{t-} h^{j_1,j_2}(t-s_1, t-s_2) \\ &\times N^{j_1}(\mathrm{d}s_1)N^{j_2}(\mathrm{d}s_2). \qquad (6)\end{aligned}$$

The class of models described by (6) contains the class of linear Hawkes processes (corresponding to $h^{j_1,j_2} = 0$) but also encompasses more complicated models, such as a model where the intensity boosts only when two events occur very close to each other.

IV. Testing Local Independence

We now return to the question of developing a test for local independence $j \not\to k | C$. We consider the approximation of $\lambda^{k,C\cup\{j\}}$ in (5), and use the higher order interactions from Section III together with basis splines to approximate $\lambda^{k,C}$. We fit this approximation from data and test significance of the contribution from j.

A. Approximating Kernel Functions With Basis Functions

We consider the question of approximating the terms $\int_0^{t-} \overline{g}^{jk}(t-s)N^j(\mathrm{d}s)$ and $\lambda_t^{k,C}$ from (5).

To approximate the intensity $\lambda_t^{k,C}$, we utilize the $W_0 \oplus W_1 \oplus W_2$-approximation from (6). We approximate the kernels $h^{j_1}(s_1)$ and $h^{j_1,j_2}(s_1, s_2)$ by spline expansions

$$h^{j_1} \approx \sum_i \beta_i^{j_1} b_i \quad \text{and} \quad h^{j_1,j_2} \approx \sum_{i_1,i_2} \beta_{i_1,i_2}^{j_1,j_2} b_{i_1} \otimes b_{i_2}$$

for some class of basis functions $\{b_i\}_i$. We propose to use B-splines [41], which is a flexible and frequently studied

model class, though one could use other classes if desired. Using a smaller number of basis functions is less expressive, but is less prone to over-fitting and makes the problem computationally easier (see Section IV-D), whereas choosing a larger number of basis functions increases the expressive power. For a given application, one can use cross validation to select an appropriate number of basis functions.

Due to the linearity in β, the coefficient terms can be collected into one vector β^C and we can write $\lambda_t^{k,C} \approx (\beta^C)^T x_t^C$. Each entry of x_t^C corresponds to one basis function integrated with respect to either a single event type or a pair of event types. For instance the entry corresponding to $\beta_{i_1,i_2}^{j_1,j_2}$ would be

$$\int_{-\infty}^{t-}\int_{-\infty}^{t-} b_{i_1}(t-s_1)b_{i_2}(t-s_2)N^{j_1}(\mathrm{d}s_1)N^{j_2}(\mathrm{d}s_2).$$

Similarly, we approximate the kernel $\overline{g}^{jk}$ by $\sum_i \overline{\beta}_i^j b_i$, and collect the coefficients to $\overline{\beta}^j$ and $\overline{x}_t^j$. Conclusively, the intensity (5) can be approximated by

$$\eta\left(\overline{\lambda}_t^{k,C\cup\{j\}}\right) = \left(\beta^C\right)^T x_t^C + \left(\overline{\beta}^j\right)^T \overline{x}_t^j =: \left(\beta^{C\cup\{j\}}\right)^T x_t^{C\cup\{j\}}$$

for some choice of β^C and $\overline{\beta}^j$.

B. Maximum Likelihood

Given an observation of a point process over the interval $[0, T]$, we compute maximum likelihood estimates $\hat{\beta}^{C\cup\{j\}}$ using the penalized log-likelihood

$$\int_0^T \log \overline{\lambda}_t^{k,C\cup\{j\}} N^k(\mathrm{d}t) - \int_0^T \overline{\lambda}_t^{k,C\cup\{j\}}\,\mathrm{d}t - \rho\left(\beta^{C\cup\{j\}}\right)$$

where $\rho(\beta) = \kappa_0 \beta^T \Omega \beta$ is a quadratic penalization, and where $\kappa_0 > 0$ and Ω is the roughness penalty matrix, which penalizes curvature of the kernel estimates (see [41, Ch. 5]). κ_0 is a hyper parameter, which needs to be chosen by the modeler; for example, this can be done by cross validation, choosing the value κ_0 which yields the largest likelihood on a held-out validation set. In practice, we find that model performance is not sensitive to the choice of κ_0.

Assuming that the true model belongs to the model class, with parameter $\beta_0^{C\cup\{j\}}$, it follows from [14] that the distribution of the maximum likelihood estimate $\hat{\beta}^{C\cup\{j\}}$ is approximately normal with mean

$$\mu = \left(I + 2\kappa_0 \hat{J}_T^{-1}\Omega\right)\beta_0^{C\cup\{j\}}$$

and covariance matrix

$$\Sigma = \hat{J}_T^{-1}\hat{K}_T\hat{J}_T^{-1}$$

where

$$\hat{K}_T = \int_0^T x_t^{C\cup\{j\}} x_t^{C\cup\{j\}T} \frac{\left(\left(\eta^{-1}\right)'\left(\hat{\beta}x_t^{C\cup\{j\}}\right)\right)^2}{\eta^{-1}\left(\hat{\beta}x_t^{C\cup\{j\}}\right)}\,\mathrm{d}t$$

$$\hat{J}_T = \hat{K}_T - 2\kappa_0\Omega.$$

If μ_j, Σ_j denotes the respective subvector and -matrix which corresponds to the entries of $\overline{\beta}^j$, the approximate distribution of the estimated parameter $\hat{\overline{\beta}}^j$ is known and can be used for testing.

C. Hypothesis Testing

We can now test the hypothesis $H_0{:}\overline{g}^{jk} = 0$ by testing whether $\overline{\beta}^j = 0$. In the setting of testing $g = 0$ for a function $g = \sum_i \beta_i b_i$, [42] shows that directly testing $\hat{\beta}^j = 0$ can lead to loss of power. Instead, [42] proposes to evaluate the function in a grid $\mathbb{X} = (u_1, \ldots, u_M)$ and perform the hypothesis test that the resulting vector $g(\mathbb{X}) := (g(u_m))_{1\le m\le M}$ is 0.

Let $\mathbb{B} = (b_i(u_m))_{m,i}$ be the matrix where the ith column is the evaluation of the ith basis function evaluated in $\mathbb{X}$. Then $\overline{g}^{jk}(\mathbb{X}) = \mathbb{B}\overline{\beta}^{jk}$ is the evaluation of $\overline{g}^{jk}$ in $\mathbb{X}$, which is then approximately $\mathcal{N}(\mathbb{B}\mu_j, \mathbb{B}\Sigma_j\mathbb{B}^T)$-distributed. This allows for testing the hypothesis $\overline{g}^{jk} = 0$ by the Wald-test statistic

$$T = \left[\mathbb{B}\hat{\overline{\beta}}^{jk}\right]^T \left(\mathbb{B}\Sigma_j\mathbb{B}^T\right)^{-1}\left[\mathbb{B}\hat{\overline{\beta}}^{jk}\right]$$

which is approximately $\chi^2_{(M)}$-distributed. By comparing T to the theoretical quantiles of $\chi^2_{(M)}$, we can test for significance of the contribution of j to the intensity $\lambda^{k,C\cup\{j\}}$. The test is implemented in `python` and is available online.[4]

D. Computational Complexity

The main complexity of the procedure is fitting the second-order estimate of $\lambda^{k,C}$. If $|C| = d_C$, this implies fitting d_C estimates of the first-order kernels h^{j_1} and $d_C(d_C+1)/2$ estimates of the second-order kernels h^{j_1,j_2} and so the complexity grows quadratically in the size d_C of the conditioning set. In general, for the nth order approximation, the number of pairs $(j_1, \ldots, j_n)$ (where repetitions $j_i = j_k, i \neq k$ is possible, but ordering does not matter) is $((d_C+n-1)!)/(n!(d_C-1)!)$.

If we use a basis $\{b_1, \ldots, b_K\}$ of K basis functions, each first-order kernel requires K parameters and each second-order kernel requires K^2 parameters, so the total number of parameters becomes $dK + d_C(d+1)K^2/2$.

For the datasets we consider, using the second-order approximation is tractable to run on a standard laptop with $K = 10$; in cases where it is not (typically if the number of processes N^j is very large), one can for example use decreasing granularity of the basis functions, $K_1 \ge K_2$, to make each individual h^{j_1,j_2} easier to approximate.

On the basis of Theorem 1, we could consider higher order approximations. The growth in the number of parameters means that although the model is more expressive, the increase in the number of parameters also increases variance of the estimator. In many cases, it is reasonable to assume that the parameter vector is sparse, and we can apply a sparsity inducing penalty during training to reduce the variance of the estimator; in this case it might be possible to include third- or higher order interactions.

V. Simulation Experiments

We evaluate our test using simulated data. First, we explore the level and power for several graphical structures. Second, we apply the test in a causal discovery algorithm to learn the local independence graph from an observed dataset. In both experiments, we compare our method to the first-order method in (4), where also the $\lambda^{k,C}$ intensity is approximated by basis expansions using only first-order interaction terms.

[4]Code available at https://github.com/nikolajthams/LIPP

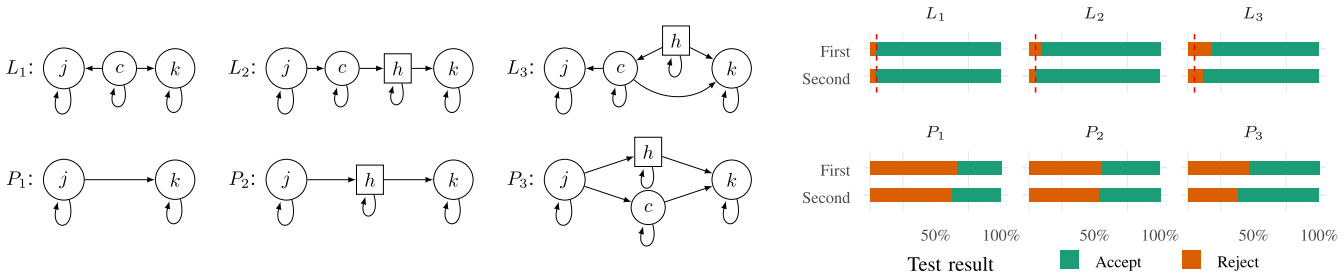

Fig. 2. (*left*) Graphical structures used for testing local independence. Square nodes indicate unobserved event types. For each, we simulate 500 samples from a Hawkes process with this true local independence graph, and evaluate the test $j \not\to k|C$ for with C being $\{c, k\}$ or (in the absence of a node c) $\{k\}$. (*right*) H_0 acceptance rates ($p < 0.05$ level) for the 500 repetitions of the test $j \to k|\{k, c\}$ in each of the structures using both a first- (1) and second- (2) order approximation of $\lambda^{k,C}$. The colors indicate the proportion of tests accepted and rejected, and the dashed line marks 5% rejection rate (only relevant for graphs L_1–L_3).

A. Level and Power

In Section II-A, we argued that the misspecification from using only first-order terms may lead to a loss of level. To validate this, for each of the graphs $\mathcal{G}$ in Fig. 2, we sample $n = 500$ point processes from the Hawkes process with kernel $g^{i_1 i_2}(s) = \alpha_{i_1 i_2}\beta_{i_1 i_2}e^{-\beta_{i_1 i_2}s}$ if $(i_1, i_2) \in \mathcal{G}$ and otherwise $g^{i_1 i_2}(s) = 0$. Simulation details are in the appendix in the Supplementary Material.

For each sample, we test the hypothesis H_0:$j \not\to k|C$ with $C = \{c, k\}$ (or $C = \{k\}$ in the graphs with no node c). The hypothesis H_0 is true in structures L_1–L_3 (and thus we here evaluate level) and false in structures P_1–P_3 (and so we here evaluate power).

The nodes h represent an unobserved event type, and so is not included in the conditioning set C. Due to the latent events, we expect the first-order test to loose level compared to the second-order test. We conduct the test of H_0 from Section IV on a nominal 5% level and display in Fig. 2 the proportion of p-values below 5% for each structure, with red indicating a rejected test of H_0.

In the structure L_1, we observe that the both the first- and second-order tests maintain level in the structure L_1. This is as expected, because the ground truth structure L_1 has no latent events, and so the effect $c \to k$ is truly a first-order interaction. In the structure L_2, our proposed second-order test has a rejection rate around 5%, while the first-order test exceeds the nominal level by rejecting in around 9% of the simulations. This indicates that due to the latent process N^h being marginalized out, the dependence between N^c and N^k is not fully captured by first-order interactions, and so when fitting only first-order interactions, there is some residual information which mistakenly is then captured in the fit kernel $\overline{g}^{jk}$. By introducing second-order interactions, this residual information is reduced, and the false negative link $j \to k$ becomes less likely. In L_3 both the first- and second-order tests reject in more than 5% of cases, however with the level of the second-order test being closer to the nominal 5% level. This indicates that the marginalization of h induces a model misspecification which is partly captured by the second-order interaction.

For the graphs P_1, P_2 and P_3, where truly $j \to k|C$, we observe that both the first- and second-order approaches have substantial power. For the structure P_3, we observe that the first-order test has more power than the second-order test, possibly due to the fewer parameters that need to be estimated to use the first-order test.

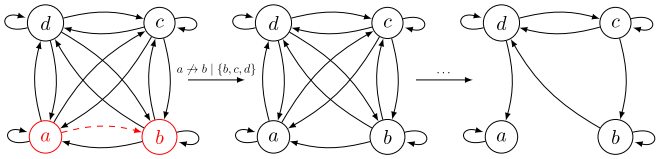

Fig. 3. Illustration of constrained-based learning algorithms like the CA algorithm [25] or the PC-algorithm [28]. The algorithm starts with the fully connected graph (*left*), and removes the edge $a \to b$ if there exist a set C of current parents of b, such that $a \not\to b|C$ (*middle*). This is then done repeatedly for all nodes and for sets C of increasing size. The algorithm terminates, when no more edges can be removed, that is when no more local independences can be found (*right*).

B. Causal Structure Learning

We also evaluate the proposed test in the context of the CA algorithm proposed by Meek [25], which is similar to the PC-algorithm [28] but applies to local independence graphs (see Fig. 3 for an illustration of the algorithm). For $d \in \{3, \ldots, 7\}$, we simulate $n = 60$ graphs of dimension d and with each edge occurring with a fixed probability of 0.2. We then simulate a Hawkes process with the simulated graph as causal graph. Simulation details are in the appendix in the Supplementary Material. Constrained-based causal learning algorithms, such as the CA-algorithm, estimate the causal graph by sequentially testing local independence $j \not\to k|C$ for nodes j, k given conditioning sets $C \subset V \setminus \{j\}$ of increasing size. If at some point, a local independence $j \not\to k|C$ is found, the edge $j \to k$ is removed from the graph.

For each simulated Hawkes process, we run the CA-algorithm using either the first- or the second-order tests and obtain a resulting estimated graph. We then compare the estimated graphs to the true graph that generated the Hawkes process by the structural hamming distance (SHD), which measures the number of edge additions, removals or flips that is needed to convert the estimated graph into the true graph. That is, the SHD measures how far the estimated graph is from the true graph. Fig. 4 shows the resulting SHDs for the different dimensions. We observe that for all dimensions, the second-order approach performs as well or better than the first-order approach. Notably, this is more outspoken as dimensions increase: In larger systems, more processes are marginalized away when testing $j \not\to k|C$, and so the effect of model misspecification is more severe for larger dimensions.

Further, we compare to a Lasso-based approach [7]: for each process N^j we fit the conditional intensity given all processes $N^1, \ldots, N^d$ (using the same basis expansion as for the independence test). We apply ℓ_1 penalization to the fit (the penalization parameter is chosen by cross validating

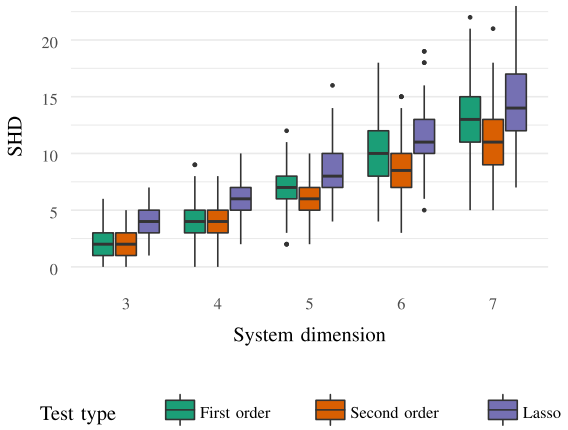

Fig. 4. SHDs between the true graph that simulated data and the graphs estimated using either first- or second-order tests or a Lasso-based approach in the experiment in Section V-B.

the predictive performance on held out time slices) and take as parents those processes whose fit parameters are non-zero. While this is computationally more efficient, there is no theory supporting consistency of this approach without further assumptions [7].

The average and median SHD of the Lasso estimate is larger than the SHDs of the constraint-based methods (Fig. 4) meaning that the Lasso estimates are further from the true graph than the constraint-based estimates. This shows that even in fully observed data (as in this experiment) where the Lasso is correctly specified, fitting intensities given all other processes simultaneously, may not be optimal. If the processes were only partially observed, this would only be more pronounced, since constraint-based methods can still estimate the Markov equivalence class of the mixed graph [27] while Lasso regression coefficients cannot represent latent variables.

VI. Neuron Firing Data

We employ a causal discovery algorithm using our proposed tests to a dataset of neuron firing in turtles.[5] The turtles were exposed to a stimuli in a period of 10 seconds, in which the activity of $d = 6$ channels were measured. The experiment was repeated 5 times.

For each repetition, we employ the CA algorithm from [25] to learn the causal structure, using either first- or second-order tests. Fig. 5 shows data from the first repetition of the experiment and the resulting learned graphs (repetitions 2–5 are shown in the appendix in the Supplementary Material). The graph estimated using second-order tests is sparser than the one using first-order tests. This concurs with our motivation for including second-order terms: when level is lost due to misspecification, the edge $j \to k$ will too often remain in the graph, even though $j \not\to k|C$ for some C. Using first-order tests results in a denser and less informative graph. This effect is more outspoken in the neuron firing data than in the simulated data in Section V: While the synthetic data were truly simulated from a Hawkes process, and so the misspecification would only be due to marginalization, there may be additional misspecification in the real data if the full process is not truly a Hawkes process.

[5]Data provided by Associate Professor Rune W. Berg, University of Copenhagen.

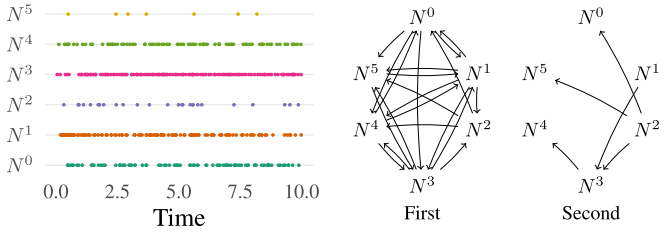

Fig. 5. (*left*) The first repetition of the experiment. Each point corresponds to one neuron firing. (*right*) Output of the CA algorithm on the first repetition, when the test of local independence either uses a first-order test ("First") or a second-order test ("Second"). In the appendix in the Supplementary Material, we show the similar plots and graphs for the repetitions two through five.

TABLE I
Consistency of Estimated Graphs From the Five Repetitions of the Stimulus Experiment

	Baseline	First	Second
Edges consistent in all 5 repetitions	6.25%	23.3%	**26.7%**
Edges consistent in a least 4 repetitions	37.5%	40.0%	**56.7%**
Number of edges present	50.0%	62.7%	30.7%

Since ground truth graphs for the neural connections are not available, we cannot directly evaluate which test provides estimated graphs closer to ground truth. Instead, we compare the first- and second-order tests by their consistency across the five repetitions, i.e., how similar the estimated graphs are from the five repetitions. For each repetition, a separate graph is learned using the CA-algorithm, with a test using either first- or second-order terms. In Table I we display the proportion of edges where either 1) all five graphs agree on the presence or absence of the edges and 2) at least 4 of 5 graphs agree. As a baseline, we include the theoretical proportions, if in each graph, an edge would appear randomly with a probability of $1/2$. Self-edges, which are easy to detect, and hence inflates consistency, are excluded from all numbers. We observe that the second-order approach is more consistent in terms of both agreement between all five repetitions and agreement between at least four repetitions.

VII. Discussion

In this article, we formulated a framework for testing local independence in point processes. We introduced a test of local independence that fits intensities using basis expansions and tests the local independence hypothesis $j \not\to k|C$ by testing significance of contributions of the process N^j to the intensity $\lambda^{k,C\cup\{j\}}$.

We addressed the issue of marginalization: Even if the full data-generating mechanism is a known and simple model class, such as Hawkes processes, a partially observed system with some event types unobserved cannot necessarily be modeled as a Hawkes process. This issue is native to (conditional) local independence testing, since the local independence $j \not\to k|C$ relates to the marginal distribution of $N^{\{j,k\}\cup C}$. To overcome this misspecification, we proved that, when facing marginalized variables, the intensity can be arbitrarily well approximated by expansions in terms of iterated integrals, and we have verified that including higher order interactions leads to an improved level of the test of $j \not\to k|C$.

The availability of an empirical local independence test is quintessential to constraint-based causal structure learning

algorithms for point processes, and we have validated in simulation studies that using our proposed test, one can from data obtain good estimates of the underlying graph. We applied our approach to a real-world dataset on neuron spiking in turtles, and found that including higher order interactions resulted in sparser, more informative estimated networks.

REFERENCES

[1] E. Bacry, I. Mastromatteo, and J.-F. Muzy, "Hawkes processes in finance," *Market Microstructure Liquidity*, vol. 1, no. 1, 2015, Art. no. 1550005.

[2] Y. Ogata, "A prospect of earthquake prediction research," *Statist. Sci.*, vol. 28, no. 4, pp. 521–541, Nov. 2013.

[3] K. Zhou, H. Zha, and L. Song, "Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes," in *Proc. 16th Int. Conf. Artif. Intell. Statist.*, 2013, pp. 641–649.

[4] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *J. Neurophysiol.*, vol. 93, no. 2, pp. 1074–1089, Feb. 2005.

[5] A. G. Hawkes, "Spectra of some self-exciting and mutually exciting point processes," *Biometrika*, vol. 58, no. 1, pp. 83–90, 1971.

[6] P. Brémaud and L. Massoulié, "Stability of nonlinear Hawkes processes," *Ann. Probab.*, vol. 24, no. 3, pp. 1563–1588, Jul. 1996.

[7] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard, "Lasso and probabilistic inequalities for multivariate point processes," *Bernoulli*, vol. 21, no. 1, pp. 83–143, 2015.

[8] S. Chen, D. Witten, and A. Shojaie, "Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process," *Electron. J. Statist.*, vol. 11, no. 1, pp. 1207–1234, Jan. 2017.

[9] D. R. Brillinger, "The identification of point process systems," *Ann. Probab.*, vol. 3, no. 6, pp. 909–924, Dec. 1975.

[10] D. R. Brillinger, H. L. Bryant, and J. P. Segundo, "Identification of synaptic interactions," *Biol. Cybern.*, vol. 22, no. 4, pp. 213–228, Dec. 1976.

[11] D. R. Brillinger, "Nerve cell spike train data analysis: A progression of technique," *J. Amer. Stat. Assoc.*, vol. 87, no. 418, pp. 260–271, Jun. 1992.

[12] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: State-of-the-art and future challenges," *Nature Neurosci.*, vol. 7, no. 5, pp. 456–461, May 2004.

[13] J. W. Pillow et al., "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, Aug. 2008.

[14] N. R. Hansen, "Nonparametric likelihood based estimation of linear filters for point processes," *Statist. Comput.*, vol. 25, no. 3, pp. 609–618, May 2015.

[15] R. Cai, S. Wu, J. Qiao, Z. Hao, K. Zhang, and X. Zhang, "THPs: Topological Hawkes processes for learning causal structure on event sequences," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 25, 2022, doi: 10.1109/TNNLS.2022.3175622.

[16] M. S. Masud and R. Borisyuk, "Statistical technique for analysing functional connectivity of multiple spike trains," *J. Neurosci. Methods*, vol. 196, no. 1, pp. 201–219, Mar. 2011.

[17] D. Song et al., "Identification of sparse neural functional connectivity using penalized likelihood estimation and basis functions," *J. Comput. Neurosci.*, vol. 35, no. 3, pp. 335–357, Dec. 2013.

[18] E. C. Hall and R. M. Willett, "Tracking dynamic point processes on networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4327–4346, Jul. 2016.

[19] M. Eichler, R. Dahlhaus, and J. Dueck, "Graphical modeling for multivariate Hawkes processes with nonparametric link functions," *J. Time Ser. Anal.*, vol. 38, no. 2, pp. 225–242, Mar. 2017.

[20] M. Achab, E. Bacry, S. Gaïffas, I. Mastromatteo, and J.-F. Muzy, "Uncovering causality from multivariate Hawkes integrated cumulants," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, Aug. 2017, pp. 1–10.

[21] S. Seth and J. C. Principe, "Assessing Granger non-causality using nonparametric measure of conditional independence," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 1, pp. 47–59, Jan. 2012.

[22] V. Didelez, "Graphical models for marked point processes based on local independence," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 70, no. 1, pp. 245–264, Feb. 2008.

[23] K. Røysland, "Counterfactual analyses with graphical models based on local independence," *Ann. Statist.*, vol. 40, no. 4, pp. 2162–2194, Aug. 2012.

[24] V. Didelez, "Causal reasoning for events in continuous time: A decision-theoretic approach," in *Proc. UAI Workshop Adv. Causal Inference*, 2015, pp. 40–45.

[25] C. Meek, "Toward learning graphical and causal process models," in *Proc. UAI Workshop Causal Inference, Learn. Predict.*, Jul. 2014, pp. 43–48.

[26] S. W. Mogensen, D. Malinsky, and N. R. Hansen, "Causal learning for partially observed stochastic dynamical systems," in *Proc. 34th Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2018, pp. 350–360.

[27] S. W. Mogensen and N. R. Hansen, "Markov equivalence of marginalized local independence graphs," *Ann. Statist.*, vol. 48, no. 1, pp. 539–559, Feb. 2020.

[28] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search* (Lecture Notes in Statistics), vol. 81. New York, NY, USA: Springer-Verlag, 1993.

[29] J. Etesami, N. Kiyavash, K. Zhang, and K. Singhal, "Learning network of multivariate Hawkes processes: A time series approach," in *Proc. 32nd Annu. Conf. Uncertainty Artif. Intell. (UAI)*, 2016, pp. 1–14.

[30] D. Bhattacharjya, K. Shanmugam, T. Gao, and D. Subramanian, "Process independence testing in proximal graphical event models," in *Proc. Conf. Causal Learn. Reasoning*, 2022, pp. 144–161.

[31] V. Volterra, *Theory of Functionals and of Integral and Integro-Differential Equations*. New York, NY, USA: Dover, 1959.

[32] M. O. Franz and B. Schölkopf, "A unifying view of Wiener and Volterra theory and polynomial kernel regression," *Neural Comput.*, vol. 18, no. 12, pp. 3097–3118, Dec. 2006.

[33] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes* (Probability and its Applications), vol. 1, 2nd ed. New York, NY, USA: Springer-Verlag, 2003.

[34] A. G. Hawkes and D. Oakes, "A cluster process representation of a self-exciting process," *J. Appl. Probab.*, vol. 11, no. 3, pp. 493–503, Sep. 1974.

[35] J.-P. Florens and D. Fougere, "Noncausality in continuous time," *Econometrica*, vol. 64, no. 5, pp. 1195–1212, Sep. 1996.

[36] M. Jacobsen, *Point Process Theory and Applications: Marked Point and Piecewise Deterministic Processes*. Berlin, Germany: Springer, 2006.

[37] E. Arjas, P. Haara, and I. Norros, "Filtering the histories of a partially observed marked point process," *Stochastic Processes Appl.*, vol. 40, no. 2, pp. 225–250, Mar. 1992.

[38] G. Last and A. Brandt, *Marked Point Processes on the Real Line: The Dynamical Approach*. Berlin, Germany: Springer, 1995.

[39] S. N. Cohen, "Chaos representations for marked point processes," *Commun. Stochastic Anal.*, vol. 6, no. 2, pp. 263–279, Jun. 2012.

[40] N. U. Ahmed, "Closure and completeness of Wiener's orthogonal set G_n in the class L^2 (Ω, B, μ) and its application to stochastic heriditary dif," *Inf. Control*, vol. 17, no. 2, pp. 161–174, Sep. 1970.

[41] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Series in Statistics). New York, NY, USA: Springer, 2001.

[42] S. N. Wood, "On p-values for smooth components of an extended generalized additive model," *Biometrika*, vol. 100, no. 1, pp. 221–228, Mar. 2013.

Nikolaj Thams received the B.Sc. degree in mathematics, the B.Sc. degree in financial mathematics, and the M.Sc. degree in statistics from the University of Copenhagen, Copenhagen, Denmark, in 2016, 2017, and 2019, respectively, and the Ph.D. degree from the Copenhagen Causality Laboratory, University of Copenhagen, in 2022, supervised by Prof. Jonas Peters.

He is currently a Quantitative Researcher with G-Research, London, U.K.

Niels Richard Hansen is a Professor of computational statistics with the University of Copenhagen, Copenhagen, Denmark. He currently serves as the Head of the Section of Statistics and Probability Theory, Department of Mathematical Sciences, and a Co-Founder of Copenhagen Causality Laboratory. His research interests include causal models and Bayesian networks, causal structure learning, identifiability, stochastic processes, predictive modeling, and double machine learning.

APPENDIX

A. Proof of Theorem 1

In this appendix, we prove Theorem 1. The proof first shows the result for a univariate process ($|C| = 1$), and then argues that result can easily be extended to the multivariate setting.

We stress that the motivation for the theorem is to show convergence of the representation. In practice, many other kernel functions than those appearing in the proof, could also be used to describe the system, and so our interest lies very little in the concrete functional forms used.

Let $\tau_1, \tau_2, \ldots$ be the jumps of N starting at 0 and moving backwards in time. That is $\ldots < \tau_2 < \tau_1 < 0$.

Definition 2. *For $s < 0$, let $\mathcal{F}_s$ denote the σ-algebra generated by events in $[s, 0)$. That is*

$$\mathcal{F}_s = \sigma(\tau_1 \vee s, \tau_2 \vee s, \ldots),$$

where $\tau \vee s = \max(\tau, s)$. Define also $\mathcal{F}_{0-} = \sigma(\cup_{s<0} \mathcal{F}_s)$.

Proposition 1. *$\mathcal{F}_0$ equals $\mathcal{F}_{-\infty} := \sigma(\tau_1, \tau_2, \ldots)$.*

Proof. For all i, $\tau_i \vee s$ is $\sigma(\tau_i)$-measurable, and in particular, $\mathcal{F}_{-\infty}$-measurable. Therefore $\mathcal{F}_s = \sigma(\tau_1 \vee s, \ldots) \subseteq \mathcal{F}_{-\infty}$ and so $\mathcal{F}_{0-} = \sigma(\cup_{s<0} \mathcal{F}_s) \subseteq \mathcal{F}_{-\infty}$.

Reversely, τ_n is $\mathcal{F}_0$-measurable for each n. $\mathcal{F}_{-\infty}$ is the smallest σ-algebra making all τ_n's measurable, so $\mathcal{F}_{-\infty} \subseteq \mathcal{F}_0$ will follow. To see that τ_n is $\mathcal{F}_0$-measureable, consider any n. $(\tau_n \vee s) \to \tau_n$ for $s \to -\infty$ (potentially with $\tau_n = -\infty$), and so since $(\tau_n \vee s)$ is $\mathcal{F}_0$-measurable for each s, τ_n is $\mathcal{F}_0$-measurable. □

Proposition 2. *The union of function spaces $\cup_{s<0} \mathcal{L}^1(\mathcal{F}_s)$ is dense in $\mathcal{L}^1(\mathcal{F}_{0-})$.*

Proof. Take any $\lambda \in \mathcal{L}^1(\mathcal{F}_{0-})$. By the tower property, $\lambda_s := E[\lambda \mid \mathcal{F}_s] \in \mathcal{L}^1(\mathcal{F}_s)$ and further from the martingale convergence theorem, $(\lambda_s)_{s<0}$ is a martingale (in $-s$) and $E[\lambda \mid \mathcal{F}_s]$ converges in $\mathcal{L}^1$ to $E[\lambda \mid \mathcal{F}_{0-}] = \lambda$ as $s \to -\infty$.

Because each $\lambda_s \in \mathcal{L}^1(\mathcal{F}_s) \subseteq \cup_{s<0} \mathcal{L}^1(\mathcal{F}_s)$, it follows that $\cup_s \mathcal{L}^1(\mathcal{F}_s)$ is dense in $\mathcal{L}^1(\mathcal{F}_{0-})$. □

We now show that for any $\lambda \in \mathcal{L}^1(\mathcal{F}_s)$ and for each $M \in \mathbb{N}$ that $\lambda 1_{N([s,0))=M}$ can be written as a sum of integrals of deterministic functions. These integrands will play a role similar to Volterra kernels, but only given the count $N([s, 0))$. We then sum over these terms, to obtain a general representation of λ.

It is well known that if $Y \in \mathcal{L}^1(\sigma(X_1, X_2, \ldots))$ for some random variables $X_1, \ldots$, then there exists a measurable map f such that $Y = f(X_1, X_2, \ldots)$. In the case of event times truncated at s, $\tau_n \vee s$, this corresponds to that if $\lambda \in \mathcal{L}^1(\mathcal{F}_s)$ there exists a function f such that

$$\lambda = f(\tau_1 \vee s, \tau_2 \vee s, \ldots)$$

To obtain an integral representation of λ, we can utilize this function. Define $f^n_s(t_1, \ldots, t_n) = f(t_1, \ldots, t_n, s, s, \ldots)$ as the evaluation of f in $(t_1, \ldots, t_n)$ and then the s in all other entries of the function. We will write f^n if s is clear from the context or even $f(t_1, \ldots, t_n)$.

As a motivation for the below proof, suppose that we knew that exactly one event occurred in the interval $A := [s, 0)$, i.e. $\tau_1 \in A, \tau_n \notin A$ for $n \geq 2$. Then one could write:

$$\begin{aligned}\lambda &= f(\tau_1 \vee s, \tau_2 \vee s, \ldots) = f(\tau_1, s, s, \ldots) \\ &= \int_s^{0-} f(t, s, s, \ldots) N(\mathrm{d}t) = \int_s^{0-} f^1 N(\mathrm{d}t)\end{aligned}$$

This however depends heavily on the assumption that $N(A) = 1$. If instead the interval contained m events, then $\int_s^{0-} f^1(t) N(\mathrm{d}t) = f^1(\tau_1) + \ldots + f^1(\tau_m)$ which is not equal to λ (because in this case $\lambda = f^m(\tau_1, \ldots, \tau_m)$).

The following proposition devices a procedure, such that one can obtain $f(\tau_1)$ exactly if $N(A) = 1$ and else 0, using only integrals of deterministic functions. For a function $h(t_1, \ldots, t_n)$, we use the shorthand notation

$$\int_A h \, \mathrm{d}N(t^n) := \int_A \cdots \int_A h(t_1, \ldots, t_n) \, \mathrm{d}N(t_1) \cdots \mathrm{d}N(t_n).$$

Proposition 3. *Assume N is a simple, non-exploding point process. Let $\lambda \in \mathcal{L}^1(\mathcal{F}_s)$ and $A = [s, 0)$. Then*

$$\sum_{n=1}^{L} \beta_n \int_A f(t_1) 1_{D_n} \, dN(t^n) \xrightarrow{a.s.} \lambda 1_{N(A)=1} \tag{7}$$

for $L \longrightarrow \infty$ where $\beta_n = \frac{(-1)^{n-1}}{(n-1)!}, n \geq 1$, and:

$$D_n = \{(t_1, \ldots, t_n) \in [-s, 0)^n \mid t_i \neq t_j \text{ for } i \neq j\}$$

Proof. Observe that while we integrate over sequences $(t_1, \ldots, t_n)$, we evaluate only the function $f^1(t_1)$ in t_1. The indicator function 1_{D_n} still is evaluated in $(t_1, \ldots, t_n)$. For this reason

$$\begin{aligned}&\int_A f(t_1) 1_{D_n}(t_1, \ldots, t_n) \, \mathrm{d}N(t^n) \\ &= \left[\int_A f(t_1) \, \mathrm{d}N(t_1)\right] \binom{N(A) - 1}{n - 1} (n - 1)!\end{aligned}$$

This follows because for each event time $\tau \in A$, there are exactly $\binom{N(A) - 1}{n - 1} (n - 1)!$ tuples $(\tau, t_2, \ldots, t_n)$ where τ is the first element and no elements are identical.

It then follows that:

$$\begin{aligned}&\sum_{n=1}^{N(A)} \beta_n \int_A f(t_1) 1_{D_n} \, \mathrm{d}N(t^n) \\ &= \left[\int_A f(t) \, \mathrm{d}N(t)\right] \sum_{n=1}^{N(A)} (-1)^{n-1} \binom{N(A) - 1}{n - 1} \\ &= \begin{cases} f(\tau_1) & N(A) = 1 \\ 0 & \text{else} \end{cases} = \lambda 1_{N(A)=1}\end{aligned}$$

This last step is utilizes that for $M = 1$, $\sum_{n=1}^{M}(-1)^{n-1}\binom{M-1}{n-1} = 1$, and for $M > 1$, the binomial formula implies that

$$\begin{aligned} 0 &= (1+(-1))^{M-1} \\ &= \sum_{n=0}^{M-1}(-1)^n \binom{M-1}{n} \\ &= \sum_{n=1}^{M}(-1)^{n-1} \binom{M-1}{n-1}. \end{aligned}$$

Since the integrand 1_{D_n} is 0 for $n \geq N(A)$, and $P(N(A) < \infty) = 1$, it follows that

$$\sum_{n=1}^{L} \beta_n \int_A f(t_1) 1_{D_n} \,\mathrm{d}N(t^n) \xrightarrow{\text{a.s.}} \lambda 1_{N(A)=1} \text{ for } L \to \infty$$

□

This extends to the following corollary:

Corollary 2. *Let $\lambda \in \mathcal{L}^1(\mathcal{F}_s)$. For $M \in \mathbb{N}$, one has:*

$$\sum_{n=M}^{L} \beta_n^M \int_A f(t_1,\ldots,t_M) 1_{D_n} 1_{O_M} \,dN(t^n) \xrightarrow{a.s.} \lambda 1_{N(A)=M}$$

with $\beta_n^M = \frac{(-1)^{n-M}}{(n-M)!}$ for $n \geq M$ and

$$O_M = \{(t_1,\ldots,t_M) \in [-s,0)^n \mid t_1 < t_2 \ldots < t_M\}$$

Proof. The case $M = 1$ is covered in Proposition 3. For $M \geq 2$, the result essentially is the same, with the additional requirement that the first M jumps should be ordered, which is handled by 1_{O_M}.

Apart from this, combinatorics of how many tuples $(t_1,\ldots,t_n)$ with $t_1 < \ldots < t_M$ ordered (as fixed by O_M) and all t's distinct (by D_n) remains the same, in particular

$$\begin{aligned} &\int_A f(t_1,\ldots,t_M) 1_{D_n} 1_{O_M} \,\mathrm{d}N(t^n) \\ &= \left[\int_A f(t_1,\ldots,t_M) 1_{O_M} \,\mathrm{d}N(t^M)\right] \binom{N(A)-M}{n-M} (n-M)! \end{aligned}$$

Consequently, the proof from Proposition 3 also applies in the case of $1_{N(A)=M}$. □

Extending further on Proposition 3 and Corollary 2, we may include the base-rate $\lambda 1_{N(A)=0}$. Let h^0 be the value of λ on the set $\{N(A) = 0\}$ (that is $h^0 = f(s,s,\ldots)$). Now $\sum_{n=1}^{L}\int_A \left[f(t_1) - h^0\right] 1_{D_n} \,\mathrm{d}N(t^n)$ will return the *additional to base-rate* intensity $f(\tau_1) - h^0$ if $N(A) = 1$ and 0 else. We combine the above:

Proposition 4. *Assume N is a non-exploding point process, and assume $\lambda \in \mathcal{L}^1(\mathcal{F}_s)$. Then*

$$\sum_{M=1}^{L}\sum_{n=M}^{L} \beta_n^M \int_A f(t_1,\ldots,t_M) 1_{D_n} 1_{O_M} \,dN(t^n) \xrightarrow{a.s.} \lambda,$$

for $L \longrightarrow \infty$.

Proof. As above, the almost sure convergence follows simply by decomposing $\lambda = \lambda 1_{N(A)=0} + \sum_{M\in\mathbb{N}} \lambda 1_{N(A)=M}$, and again observing that since $P(N(A) < \infty)$, for every ω, the left hand side will arrive at the true value for some finite L. □

Finally we are able to prove the main result.

Proof of Theorem 1. Observe that each function $\beta_n^M f 1_{D_n} 1_{O_M} \in E_n$, and so

$$\sum_{M=1}^{L}\sum_{n=M}^{L} \beta_n^M \int_A f(t_1,\ldots,t_M) 1_{D_n} 1_{O_M} \,\mathrm{d}N(t^n)$$

is in $\bigoplus_{n=0}^{L} W_n$. Be reminded that by Proposition 2, $\cup_{s<0}\mathcal{L}^1(\mathcal{F}_s)$ is ($\mathcal{L}^1$-)dense in $\mathcal{L}^1(\mathcal{F}_{0-})$, and for every element λ of $\cup_{s<0}\mathcal{L}^1(\mathcal{F}_s)$ there exists a sequence in $\bigoplus_{n\in\mathbb{N}} W_n$ converging almost surely to λ. Consequently, as both $\mathcal{L}^1$ and almost sure convergence implies convergence in probability, for any $\lambda \in \mathcal{L}^1(\mathcal{F}_{0-})$ there exist a sequence in $\bigoplus_{n\in\mathbb{N}} W_n$ converging to λ in probability.

Consider now any $\lambda \in \mathcal{L}(\mathcal{F}_{0-}^C)$ with $|\lambda| < \infty$ a.s. Trivially $\lambda_k := 1_{|\lambda|<K}\lambda$ converges in probability to λ for $k \to \infty$. Further each $\lambda_k \in \mathcal{L}^1(\mathcal{F}_{0-})$, and hence there exists a sequence there exists a sequence in $\bigoplus_{n\in\mathbb{N}} W_n$ converging almost surely to λ, completing the proof in the case without marks. □

The above framework is readily extended to marked point processes. Remember that with $V = \{1,\ldots,d\}$ and $C \subseteq V$, one has for any Borel measurable set A that:

$$N(A \times C) = \sum_{v\in C} N(A \times \{v\}) = \sum_{v\in C} N^v(A)$$

When integrating, this factorizes:

$$\begin{aligned} &\int_{A\times C} f(x,v) N(\,\mathrm{d}x, \,\mathrm{d}v) \\ &= \int_A f(x,v) \sum_{v\in C} N^v(\,\mathrm{d}x) \\ &= \sum_{v\in C} \int_A f^v(x) N^v(\,\mathrm{d}x) \end{aligned}$$

where we let $f^v(x) := f(x,v)$. Similarly in higher dimensions:

$$\begin{aligned} &\int_{A\times C} f(x_1,v_1,\ldots,x_n,v_n) N(\,\mathrm{d}x^n \times \,\mathrm{d}v^n) \\ &= \sum_{|\alpha|=n} \int_{A\times C} f^\alpha(x_1,\ldots,x_n) \underbrace{N^{\alpha_1}(x^1)\cdots N^{\alpha_n}(x^n)}_{=:N^\alpha(\,\mathrm{d}x^n)} \end{aligned}$$

where $f^\alpha(x_1,\ldots,x_n) = f(x_1,\alpha_1,\ldots x_n,\alpha_n)$ and $\alpha \in V^n$ is some tuple of length n.

Thus, the combinatorics of the one-dimensional case apply also in the marked setting and the result thus directly transfers to the multi-dimensional case: In the marked setting, the generated σ-field becomes $\mathcal{F}_s = \sigma((\tau_1 \vee s, v_1 1_{\tau_1>s}),\ldots)$. A multivariate version of Proposition 1 follows because $(\tau_1 \vee s, v_1 1_{\tau_1>s}) \to (\tau_1, v_1 1_{\tau_1>-\infty})$[6], and so denseness of $\cup_s \mathcal{L}^1(\mathcal{F}_s)$ also follows in the marked case. Thus the function f could have been written:

$$\lambda = f\left((\tau_1 \vee s, v_1 1_{\tau_1>s}),\ldots\right)$$

[6]Which is the desired limit, with the convention that $v_n = 0$ if $\tau_n = -\infty$.

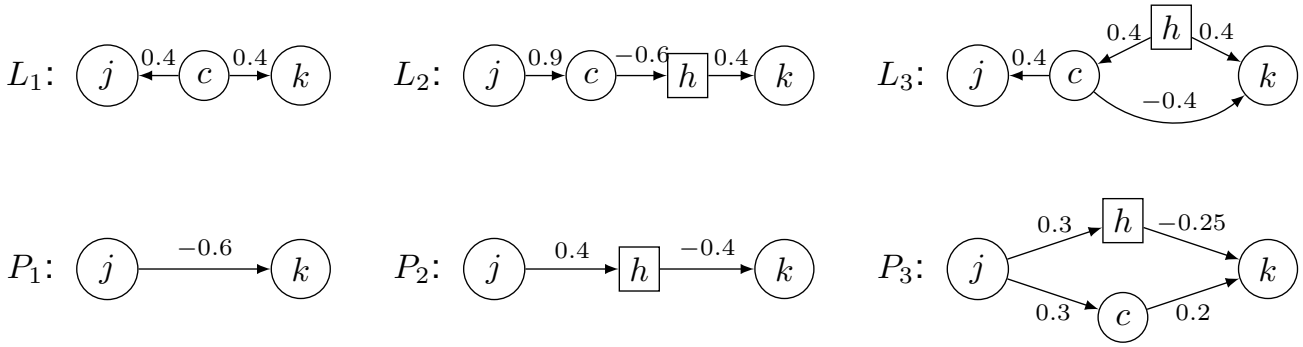

Fig. 6. Simulation parameters for the experiment in Section V-A

In Proposition 3, one could have proceeded in exactly the same way, but using integrals $\int_{A\times V} f(t_1, v_1)N(\mathrm{d}t_1 \times v_1)$ instead.

Therefore also Proposition 4 generalizes such that any $\lambda \in \mathcal{L}^1(\mathcal{F}_s)$ can be approximated by an almost surely converging sequence, and combined with the denseness result, the result extends to the multivariate case.

B. Simulation details

In this section, we provide simulation details for the experiments in Section V.

1) Details from Section V-A: Recall that from each structure, we sampled point processes with kernels $g^{i_1 i_2}(s) = \alpha_{i_1 i_2}\beta_{i_1 i_2}e^{-\beta_{i_1 i_2}s}$ if $(i_1, i_2) \in \mathcal{G}$ and otherwise $g^{i_1 i_2}(s) = 0$. We simulated data using the link-function $\eta(x) = 1_{x\geq 1}\cdot x + 1_{x<1}\cdot(\log(x)+1)$.

For all structures and edges, the decay parameter $\beta_{i_1 i_2}$ is 0.8, the baseline intensity is $\beta_0 = 0.25$ and the rate parameter on self-edges is $\alpha_{i_1 i_1} = 0.4$. The remaining rate parameters $\alpha_{i_1 i_2}$ are given in Fig. 6.

2) Details from Section V-B: All graphs are sampled randomly with all self-edges present and all other edges sampled with a probability of an edge occuring at $p = 0.2$. Given the graph, Hawkes processes are sampled with kernels $g^{i_1 i_2}(s) = \alpha_{i_1 i_2}\beta_{i_1 i_2}e^{-\beta_{i_1 i_2}s}$ if $(i_1, i_2) \in \mathcal{G}$ and otherwise $g^{i_1 i_2}(s) = 0$, and again using the link function $\eta(x) = 1_{x\geq 1}\cdot x + 1_{x<1}\cdot(\log(x)+1)$. The decay parameter is $\beta^{i_1 i_2} = 0.8$, the baseline intensities $\beta_0^{i_1} = 0.25$, and for self-edges the rate parameter is $\alpha_{i_1 i_1} = 0.3$. The rate parameters between two different nodes is $s\cdot 0.4$ where $P(s = 1) = P(s = -1) = 1/2$.

C. Estimated graphs for remaining 4 experiments

Figure 5 in Section VI we displayed data and resulting estimated graphs from the first repetition in an experiment that was repeated 5 times. This section contains plots similar to Fig. 5, but for the other 4 repetitions. These are displayed in Figs. 7 to 10.

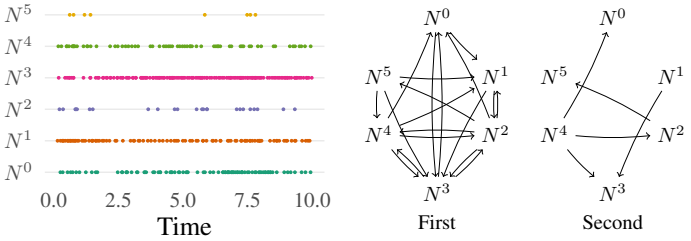

Fig. 7. Data and estimated graphs from repetition 2

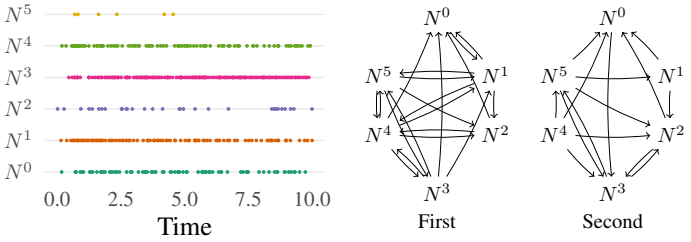

Fig. 8. Data and estimated graphs from repetition 3

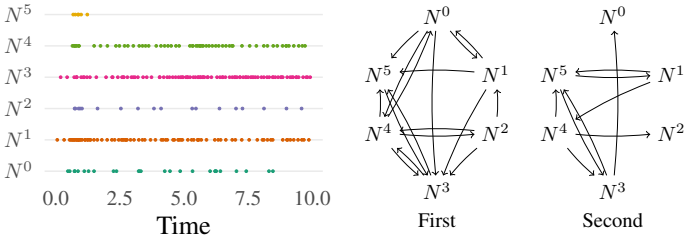

Fig. 9. Data and estimated graphs from repetition 4

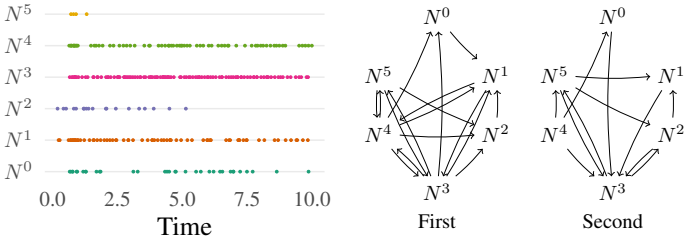

Fig. 10. Data and estimated graphs from repetition 5