UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Master Thesis in Statistics

Nikolaj Theodor Thams

Causal Structure Learning in Multivariate Point Processes

Abstract

In this thesis, we consider the problem of learning the causal structure of a multivariate point process through constrained-based learning. We introduce temporal point processes and study local independence graphs as a means of describing the local independence structure of a stochastic process. Our main contribution is a non-parametric test for local independence in multivariate point processes. The test uses higher-order interactions between process events to approximate the intensity, an approximation we show to converge under regularity assumptions. We discuss the computational challenge of the proposed test, and evaluate the performance in simulation studies.

Advisor: Niels Richard Hansen

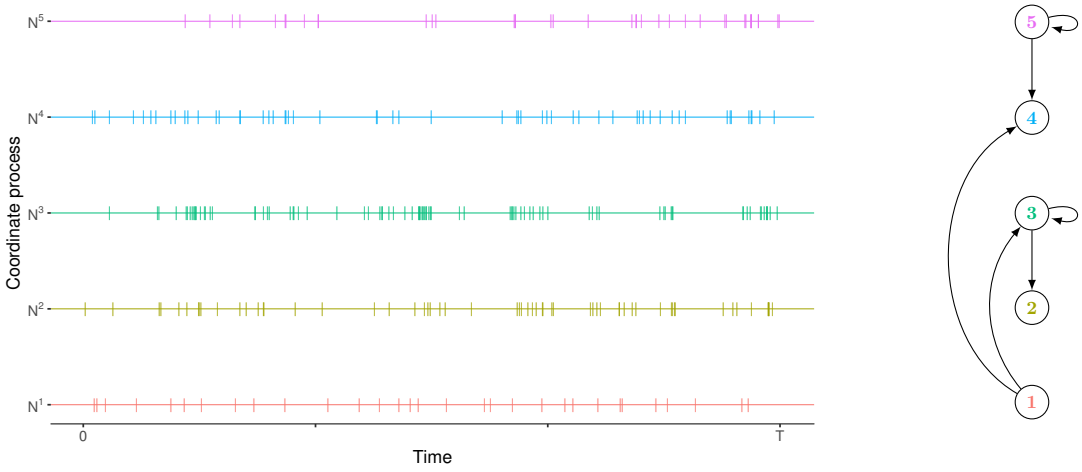
July 18, 2019

Table of Contents

1 Introduction

In this thesis, we develop causal methods for point processes. As a motivation, consider the data set in Figure 1a, where 5 different types of events are observed in an interval $[0, T]$, each tick representing an individual event. One could imagine this to be 5 different neurons spiking, companies in 5 different markets defaulting or the use of 5 different words in social media posts. The data is a realization of a **counting process**, a stochastic process modelling repeated observations of events over time.

In what way did events trigger further events? Clearly there is some dependence between points: While the events of N^1 are rather uniformly spread out across the interval, events of N^3 clearly exhibit a clustering. Also, what would happen, if we intervened in the system, say by blocking events from occurring in N^5?



(a) (b)

Figure 1: (a): A simulated data set from a 5-dimensional point process. Each tick represents one point in a coordinate process. (b): Local independence graph of the underlying process.

These questions call for a **causal model**. A causal model not only fits the joint distribution of the data, but also fits a causal structure, which one assumes entails the right predictions under interventions to the system.

One characterization of causal models builds on **local independence**. Roughly, we say that the coordinate process N^b is locally independent of N^a given N^C, if N^a does not affect the rate of occurrence of events in N^b, once the events of N^C are already known, and in that case we write $a \not\to b \mid C$.

From this we could draw a graph, with each node corresponding to a coordinate process, and draw an edge $a \to b$, whenever N^a affects N^b, even when we

are given all other coordinates. We call this the **local independence graph**, and it is an important characterization of a causal model.

Figure 1b shows the local indepence graph of the underlying process that generated the data set in Figure 1a. Indeed we observe several arrows into 3 and none into 1, which supports our observations about the respective clustering and non-clustering of the two. The local independence graph is a simple object, and does not contain the sign nor size of the effect mediated through an arrow. Yet, it illuminates the fundamental causal structure of the process. On the question of blocking events of N^5, we see from Figure 1b that while N^4 is potentially altered, N^1, N^2 and N^3 are not.

Algorithms for learning such local independence graphs are known as **structure learning** algorithms, amongst which several build the graph by assuming that a test of $a \not\to b \mid C$ is available. In this thesis, we propose an empirical test of local independence. We do so by first showing a general expansion of any intensity as:

$$h^0 + \sum_{n=1}^{\infty} \sum_{(\alpha_1,\ldots,\alpha_n)} \int_{-\infty}^{t} \cdots \int_{-\infty}^{t} h^{\alpha}(t - s_1, \cdots t - s_n) N^{\alpha_1}(\mathrm{d}s_1) \cdots N^{\alpha_n}(\mathrm{d}s_n) \quad (1)$$

We then propose a non-parametric approximation of this expansion by truncating the outer sum and approximating the integrands with a spline basis. Further, we suggest how to use the resulting estimates from this procedure to test empirically whether $a \not\to b \mid C$.

In some sense, Figures 1a and 1b represent respectively the input and the output of the methods presented in this thesis. The road between the two is paved by a vast existing theory, some of which we present below, alongside our own contribution, the local empirical independence test.

Local independence was first introduced by (Schweder 1970) and extended by (Aalen 1987) to all processes which to some extend allow for an intensity. More recently local independence has been used in the context of graphical models (Didelez 2000; Didelez 2008). We employ this graphical model framework as well, amongst others for the applicability in causality (Pearl 2009; Peters et al. 2017).

Several algorithms exist for structure learning in the non-temporal case, such as the PC- or FCI-algorithms (Spirtes et al. 2000). Further these have been extended to the dynamic settings (Meek 2014; Mogensen, Malinsky, et al. 2018).

While only few methods exist for testing local independence in dynamic systems, several independence tests have been developed in the non-temporal case (Ramsey 2014; Gretton et al. 2008; Fukumizu et al. 2008).

In this thesis, Section 2 gives an introduction to causal inference and in Section 3 we present the graphical model theory needed. Section 4 introduces point processes and in Section 5, we show the intensity expansion from Equation (1).

In Section 6 we define local independence for point processes, which leads to the local empirical independence test which we propose in Section 7. In Sections 8 and 9 we evaluate computational complexity and experimental results, and Section 10 contains a discussion of our methods and results.

2 A primer in Causal Inference

The main goal of this thesis is developing conditional independence tests to be used in point processes. The stimuli for doing so is the use of independence tests in causal inference, which we thus introduce here.

Although our overall focus is on causal inference in stochastic processes (dynamical systems), the elementary definitions are easier understood in the case of static distributions, and so we use this as an introduction to the field. Our presentation follows a Structural Causal Model approach, similar to that of for instance (Pearl 2009).

2.1 Structural Causal Models

A central task in causal inference is inferring the causal structure of a given model, a structure which exceeds that of the joint probabilistic structure. In particular, by a causal model we mean a SCM, which has the benefit of inducing interventional distributions, both of which are to be defined below:

Definition 2.1. A **structural causal model (SCM)** $\mathfrak{C} = (f_j, \epsilon_j)_{1 \leq j \leq d}$ on a set of variables $(X_1, \ldots, X_d)$ is a collection of independent noise variables ϵ_j with a specified joint distribution P_ϵ, and functional assignments f_j such that for each j:

$$X_j = f_j(\{X_i\}_{i<j}, \epsilon_j)$$

(with $X_1 = f_1(\epsilon_1)$). While f_j may potentially depend on all $j-1$ variables, often-times a number of these enter trivially. We let $\mathrm{pa}(j)$ denote the set of non-trivial entries of f_j[1]. ∘

It is clear from this definition, that a causal model entails a joint distribution over the variables $X_1, \ldots, X_d$, which we refer to as the probabilistic distribution, and denote $P^{\mathfrak{C}}$.

Obviously there are several other ways than the SCM framework to specify a probabilistic model. The quintessential strengthening of a SCM over a probabilistic model is the fact that causal models entail **interventional distributions**. I.e. a causal model describes not only how variables of the model appear together (the probabilistic distribution), but also how an intervention affects the system.

Definition 2.2. Consider a SCM $\mathfrak{C} = (f_j, \epsilon_j)_{1 \leq j \leq d}$, and let $\tilde{f}, \tilde{\epsilon}$ be some (other) assignment and noise, which is independent of $\epsilon_1, \ldots, \epsilon_d$. We perform an **intervention** and obtain an intervened SCM $\tilde{\mathfrak{C}}$ by replacing one of the assignments $X_j = f_j(\{X_i\}_{i<j}, \epsilon_j)$ by the assignment $X_j = \tilde{f}(\{X_i\}_{i<j}, \tilde{\epsilon}_j)$.

The joint distribution on $X_1, \ldots, X_d$ entailed by $\tilde{\mathfrak{C}}$ is called the **interventional distribution**. Commonly, one considers *hard* interventions, where

[1] For instance suppose $f_3(x_1, x_2, \epsilon_3) = 3x_2 + \epsilon_3$. Then we say x_1 enters f_3 trivially, and $\mathrm{pa}(3) = \{2\}$.

$\tilde{f}(\{x_i\}_{i<j}, y) = y$ and $\tilde{\epsilon}_j$ is the degenerate distribution $\tilde{\epsilon}_j = c$ for a constant c. Thus the hard intervention is $X_j = c$. ∘

We elaborate the above definitions with a very simple example.

Example 2.3. Consider the following 3 models of random variables on $\mathbb{R}^2$.

- The joint distribution P_1 of the vector
$$\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix} \sim \mathcal{N}\Big(0, \underbrace{\begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix}}_{:=A}\Big)$$
- The SCM $\mathfrak{C}_2$ over variables (X_2, Y_2) with $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0,1)$ and where $X_2 = f_1(\epsilon_1) = \epsilon_1, Y_2 = f_2(X_2, \epsilon_2) = 3X_2 + \epsilon_2$. This amounts to
$$\begin{pmatrix} X_2 \\ Y_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}}_{:=B} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$
- The SCM $\mathfrak{C}_3$ over variables (X_3, Y_3) with $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0,1)$ and where $Y_3 = f_1(\epsilon_1) = \sqrt{10}\epsilon_1, X_3 = f_2(Y_3, \epsilon_2) = \frac{3Y_3 + \epsilon_2}{\sqrt{10}}$. This amounts to
$$\begin{pmatrix} X_3 \\ Y_3 \end{pmatrix} = \underbrace{\frac{1}{\sqrt{10}}\begin{pmatrix} 1 & 3 \\ 0 & 10 \end{pmatrix}}_{:=C} \begin{pmatrix} \epsilon_2 \\ \epsilon_1 \end{pmatrix}$$

Obviously P_1 is a probabilistic model, while $\mathfrak{C}_2$ and $\mathfrak{C}_3$ are SCMs. It turns out however that the 3 models entail exactly the same probabilistic distribution, namely a Gaussian distribution with mean 0 and covariance $BB^T = CC^T = \begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix} = A$.

However, the interventions specified from the models are not the same. In the case of P_1, an intervention is not even specified. We have no a priori assumption of the origin of the joint distribution, and so we have no idea what would happen for instance to the distribution of Y_1, if we made the hard intervention $X_1 = 0$.

The intervention $X_2 = 0$ in $\mathfrak{C}_2$ would mean that $Y_2 = 3 \cdot 0 + \epsilon_2$, and so the joint distribution $P^{\tilde{\mathfrak{C}}_2}$ of this intervention is one where $X_2 = 0$ and $Y_2 \sim \mathcal{N}(0,1)$. The intervention $X_3 = 0$ in $\mathfrak{C}_3$ on the contrary will not affect the distribution of Y_3, and so $P^{\tilde{\mathfrak{C}}_3}$ is the distribution where $X_3 = 0$, $Y_3 \sim \mathcal{N}(0, 10)$.

Hence, although $P^{\mathfrak{C}_2} = P^{\mathfrak{C}_3}$, making the same hard intervention $X_i = 0$ in the two models yield different interventional distributions $P^{\tilde{\mathfrak{C}}_2} \neq P^{\tilde{\mathfrak{C}}_3}$. In fact this also shows, that interventions are a different concept than conditioning. Obviously, since $P^{\mathfrak{C}_2} = P^{\mathfrak{C}_3}$, they have the same conditional distribution of $X = 0$ - a bit of calculation show that $Y \mid X = 0 \sim \mathcal{N}(0,1)$. However, this is not the interventional distribution of $X = 0$ in $P^{\tilde{\mathfrak{C}}_3}$. (The fact that they coincide in $\mathfrak{C}_2$ relates to the fact that X_2 is parentless in this SCM. See (Peters et al. 2017, Section 6.6) for cases where intervention and conditioning coincides.) ∘

From this example it is clear that structural causal models contain more information about the distribution than the probabilistic distribution, since the model also describes what happens under interventions.

We should however be aware that this extra structure doesn't come for free. Whenever one uses a SCM to model actual data, one, implicitly or explicitly, *assumes* that the SCM encapsules the correct interventional distributions. But just as well as one could fit an incorrect statistical model, leading for instance to incorrect predictions $\mathbb{E}[Y \mid X]$, one can fit a wrong SCM, leading to incorrect interventional distributions.

2.2 Structure learning

As demonstrated above, due to the extra structure, a SCM is an appealing framework to perform (causal) inference in. In real applications however, one is not given the SCM, but rather some data set sampled from a SCM. In order to draw interventional conclusions, one needs to learn the structure of the SCM, concretely the parents of each variable. We refer to this task as **structure learning**.

The essence of causal structure learning is thus to identify from some given data, which SCMs are likely to have generated this data. This is a difficult task - Example 2.3 above even show that it may be impossible. Suppose one obtained data sampled from either $\mathfrak{C}_2$ or $\mathfrak{C}_3$ and tried to determine which of the two would be the true data-generating mechanism. Even for an infinitely large sample size, one would not be able to quantify which of the two is more likely, since they entail the same observational distribution, and so generate data sets from the same distribution.

Nonetheless, in several other cases, much of the structure can be inferred. In fact, as discussed in (Peters et al. 2017, Chapters 4 and 7), the case of a linear Gaussian SCM is in some sense the *hardest* structure to identify (contrary to the role of the linear Gaussian case almost everywhere else in statistics).

We stress that structure learning (or at least the part we consider here) focuses on learning the causal parents $\mathrm{pa}(j)$ of each variable j. Having done so, one has not learned the full SCM - the typical structure learning algorithm does not regard the concrete noise distribution P_ϵ or the concrete functional form of f_j - but only which entries of f_j are non-trivial. In (the several) cases where the concrete forms f_j are of interest, one can then compose the structure learning algorithm with for example a regression algorithm, regressing each variable only on its parents.

Graphs representing SCMs

Working with SCMs, a useful tool is representing the parent structure as a graph. Simply, given an SCM $\mathfrak{C}$ over nodes $V = \{1, \dots, d\}$, the **causal graph** is the directed graph $\mathcal{G}_{\mathfrak{C}} = (V, E)$ where the edge set E has an edge from node i to node j if $i \in \mathrm{pa}(j)$ (and so the parent set defined in the SCM context above coincide with the graph theoretic concept of a parent set in $\mathcal{G}_{\mathfrak{C}}$).

In the static case[2], $\mathcal{G}_{\mathfrak{C}}$ is a directed acyclic graph (DAG), due to the inequality $i < j$ in the assignments $X_j = f_j(\{X_i\}_{i<j}, \epsilon_j)$. For instance, the SCMs considered in Example 2.3 have causal graphs:

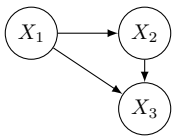
Due to the large theory of conditional independence and Markov kernels, it is of interest to ask: Can conditional independence statements relating to $P^{\mathfrak{C}}$ be read off the graph $\mathcal{G}_{\mathfrak{C}}$? The degrees to which this is possible is summarized in various **Markov properties**. Very heuristically, a distribution P satisfies the so called **global Markov property** with respect to a graph $\mathcal{G}$, if separation of A and B by C in $\mathcal{G}$, implies conditional independence $A \perp\!\!\!\perp B \mid C$ in the distribution P.

If $P^{\mathfrak{C}}$ is absolutely continuous, it follows from (Peters et al. 2017, Proposition 6.31) that for any SCM, $P^{\mathfrak{C}}$ satisfies the global Markov property with respect to its causal graph $\mathcal{G}_{\mathfrak{C}}$. How this separation is to be defined and several more details are formally presented in the dynamical case in Section 3.3. For a comprehensive introduction in the static case, see (Lauritzen 1996).

When forming causal graphs, one should exercise care: It is not in general true that the distribution of X_j changes under interventions on X_k, even if f_j is non-trivial in x_k (i.e. even if $k \in \mathrm{pa}(j)$). As an example consider the following linear Gaussian SCM:

$$\begin{aligned} \epsilon_1, \epsilon_2, \epsilon_3 &\sim \mathcal{N}(0,1) \\ X_1 &= \epsilon_1 \\ X_2 &= X_1 + \epsilon_2 \\ X_3 &= X_1 - X_2 + \epsilon_3 \end{aligned}$$

The causal graph for this is:



Clearly the assignment f_3 depends both on X_1 and X_2. Nonetheless, by inserting $X_2 = X_1 + \epsilon_2$ into the expression for X_3, we obtain $X_3 = \epsilon_3 - \epsilon_2$, which is completely independent of X_1 - and any intervention on X_1 would be unnoticeable at X_3. Intuitively this resonates badly with the arrow $X_1 \to X_3$.

This example captures the essence of **faithfulness** (or rather a lack of the same), which is the assumption that if variables are conditionally independent, then they are also separated in the graph. Again Section 3.3 gives a formal treatment in the dynamical case.

[2]but not necessarily in the temporal, as we will see below

The PC-algorithm

An example of a structure learning algorithm is the PC-algorithm (Spirtes et al. 2000), which assumes that data originates from a SCM $\mathfrak{C}$ such that $P^{\mathfrak{C}}$ is faithful with respect to $\mathcal{G}_{\mathfrak{C}}$. The algorithm outputs a graph $\mathcal{G}_{\text{PC}}$ by starting with a fully connected graph, and sequentially removes edges $a \to b \mid C$ if $a \perp\!\!\!\perp b \mid C$, with the size of conditioning set C gradually increasing throughout the algorithm. While the algorithm *does not* guarantee $\mathcal{G}_{\mathfrak{C}} = \mathcal{G}_{\text{PC}}$, $\mathcal{G}_{\mathfrak{C}}$ and $\mathcal{G}_{\text{PC}}$ are guranteed to induce the same conditional independences under the global Markov property (this is known as Markov equivalence).

The PC-algorithm supposes that we have an **oracle independence test**, i.e. for collections of variables A, B and C, a (completely true) test of whether $A \perp\!\!\!\perp B \mid C$. Though this is (potentially) simple mathematics if the joint distribution $P^{\mathfrak{C}}$ is known, given a real world sample, one needs to develop a way to test this, and only a noisy answer can be expected.

An important computational feature of the PC-algorithm is, that most of the independence tests $a \perp\!\!\!\perp b \mid C$ are done with relatively small sets C. As we shall later learn, conditional independence testing can be very computationally expensive, the cost growing rapidly in $|C|$. This motivates 'bottom-up' type algorithms, as opposed for instance to a method starting by regressing variables on the entire system at once.

2.3 Causality in dynamical systems

Let us now consider the case of a dynamical system, i.e. a number of variables $(X_t^j)_{j=1,\dots,d}$ that evolves over time. Can we define a framework similar to the static SCM that enables us to consider e.g. causal parents and interventions?

The answer is definitely *Yes* if our system evolves only over a finite number of timepoints $t_1, \dots, t_M$. In this setting, one could disregard the temporal understanding and treat a such system as a static SCM, with noise variables ϵ_t^j and assignments $X_{t_i}^j = f(\{X_{t_l}^v\}_{t_l < t_i}^{v \in V}, \epsilon_{t_i}^j)$ introducing only the temporal structure through the condition that $\text{pa}(X_{t_l}^j) \subseteq \{X_{t_j}^i \mid t_j < t_i\}$. An example of such a system is represented in the left side Figure 2, with arrows indicating the parent sets of the SCM.

However, in a continuous time system, one would need to include uncountably many nodes, which would neither be theoretically nor practically fruitful - how should one for instance define a PC-algorithm over an uncountable graph?

Instead, one could imagine collapsing the representation into a graph which omits the time axis, as is for instance the case in the right hand side of Figure 2. Here we collapse the edges $X_{t_l}^1 \to X_{t_i}^1$ to a self-loop $X^1 \to X^1$. Contrary to the static case, the graph is not a DAG, as is evident from the cycle $X^1 \to X^2 \to X^1$.

The following Sections 3 and 6 formalize this idea by local independence, an independence concept for stochastic processes, and local independence graphs, which are exactly such collapsed graphs.

While the intuition of a SCM is very helpful in dynamical settings, the concept does not directly carry over, and so in dynamical systems, we build the

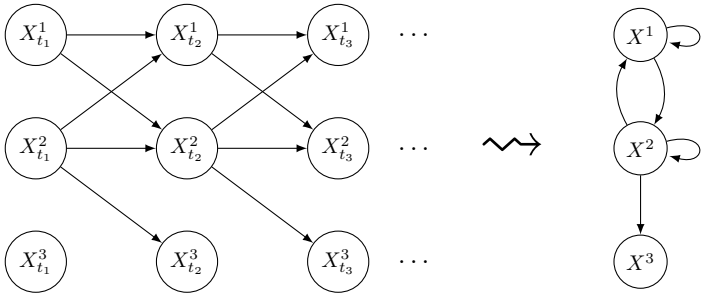


Figure 2: On the left, a discrete dynamical system over 3 nodes, each arrow indicating a non-trivial entry of an assignment function in a SCM. On the right, the system is collapsed to a graph omitting the time.

theory around local independence without any direct reference to a SCM. (Sokol and Hansen 2014) suggests using SDE's as a proxy for SCMs, the independent noises ϵ_j replaced by Levy processes driving the SDE, and define interventions as changes to the (in general non-linear) coefficients of the SDE. While we do not employ this setup, the approach is coherent with the framework presented below[3].

[3]In particular (Sokol and Hansen 2014) show that interventions are determined by the generator, which in a point process is again determined by the intensity, which is used in the framework presented below

3 Graph theory and Markov properties

This section presents the concepts needed for developing graphical models for dynamical systems. In Section 3.1, we first introduce the graph theory that is relevant for dealing with collapsed graphs for dynamical systems. We then introduce a number of graph separation criteria in Section 3.2, which is further used to introduce the Markov properties in Section 3.3 for an abstract independence model. In Section 3.4 we introduce the Causal Analysis algorithm, a dynamical pendant to the PC-algorithm introduced in Section 2.2.

3.1 Elementary graph theory

A directed graph $\mathcal{G} = (V, E)$, consist of a finite set of nodes $V = \{1, \dots, d\}$ and edges $E \subseteq V \times V$. While several applications does not permit the diagonal entries of $V \times V$, this is allowed here, and so self-edges are allowed to appear. We introduce a number of definitions relating to a directed graph $\mathcal{G} = (V, E)$:

Parental set For $v \in V$, the **parental set** $\mathrm{pa}(v)$ denotes the set of nodes v' such that $v' \to v$, i.e.

$$\mathrm{pa}(v) = \{v' \in V \mid (v', v) \in E\}$$

Note that if $v \to v$ then $v \in \mathrm{pa}(v)$.

Walk A **walk** from a to b is a sequence of nodes and edges $p = (v_1, e_1, \dots, e_{n-1}, v_n)$ with $v_1 = a$ and $v_n = b$ and e_j connecting v_j and v_{j+1} in the graph.

If $n = 1$, the walk is called **trivial**.

Since two nodes v_1 and v_2, can be connected by up to 2 edges (one in each direction) at the same time, one needs to specify which exact edges participate in the graph. However, whenever there is no ambiguity, we will just associate the walk with the nodes, i.e. $p = (v_1, \dots, v_n)$.

Path A **path** from a to b is walk between a and b where no node occurs more than once. All trivial walks are paths.

Directed path A **directed path** $p = (v_1, \dots, v_n)$ from a to b is a path from a to b where $v_j \to v_{j+1}$ for every j. All trivial paths are directed.

Ancestral set For $v \in V$ the **ancestral set** $\mathrm{An}(v)$ is the set of nodes $v' \in V$ such that a directed path from v' to v exists. More generally for a set $C \subseteq V$, we define $\mathrm{An}(C) = \cup_{v \in C} \mathrm{An}(v)$. Note that by the convention that all trivial paths are directed, $C \subseteq \mathrm{An}(C)$.

Collider For a path $p = (v_1, e_1, \dots, e_{n-1}, v_n)$, a **collider** is a non-endpoint node v_j such that $e_{j-1} \to v_j \leftarrow e_j$. Let $\mathrm{coll}(p) \subseteq \{v_2, \dots, v_{n-1}\}$ denote the set colliders of p. Further let $\mathrm{noncoll}(p)$ denote the subset of $\{v_2, \dots, v_{n-1}\}$ which appear as non-colliders. Note that $\{v_2, \dots, v_{n-1}\} = \mathrm{noncoll}(p) \cup \mathrm{coll}(p)$, though this union is in general not disjoint.

Front-door thinning For a given subset $A \subseteq V$, let $\mathcal{G}^A$ denote the graph obtained by removing all edges leaving A, i.e. $\mathcal{G}^A = (V, E^A)$ with

$$E^A = E \Big\backslash \left\{ \bigcup_{a \in A, v \in V} (a, v) \right\}$$

Subgraph Consider a subset $F \subseteq E$ of vertices. The **subgraph** $\mathcal{G}_F$ of $\mathcal{G}$ induced by F is the graph:

$$\mathcal{G}_F := (F, E \cap F \times F)$$

i.e. the graph on vertices F where an edge $v \to v'$ occurs in $\mathcal{G}_F$ if $v, v' \in F$, and $v \to v'$ in $\mathcal{G}$.

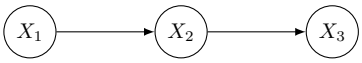
Supergraph If $V \subseteq V'$ and $E \subseteq E'$, we way that $\mathcal{G}' = (V', E')$ is a **supergraph** of $\mathcal{G}$.

Moral graph The moral graph $\mathcal{G}^m$ is obtained by adding an (undirected) edge between all pairs of nodes that have a common child (i.e. marrying unmarried parents, thus the name), and removing directions on all arrows. Both the cases $A \to B$ and $A \leftarrow B$ yield the same result, namely $A - B$.

3.2 μ-separation

When working in a given class of probability distributions, a central task in applying graphical models is to develop graph-separation concepts which in a meaningful way encode the independence structure of the distribution - as we will see, concretely by enabling the global Markov property of the distributions with respect to independence graphs.

Example 3.1. As a motivation, consider 3 random variables X_1, X_2 and X_3 with a joint distribution P. Suppose we are also given the graph:

$\mathcal{G}$: 

Is $\mathcal{G}$ helpful in describing P? Obviously this depends much both on what P concretely is and what we mean by helpful.

Intuitively X_2 separates X_1 from X_3 in $\mathcal{G}$, since all paths from X_1 to X_3 pass through X_2, and so one definition of helpful would be that this separation is reflected in P as an independence of X_1 and X_3 given X_2. This section formalizes different types of graph separation, and the following section introduces exactly such a way of relating separations in $\mathcal{G}$ to independence in P. ∘

Definition 3.2. A **graph separation** $\perp$ is a ternary relation on subsets of nodes. I.e. given a graph $\mathcal{G} = (V, E)$, for any $A, B, C \subseteq V$, it can be determined from the graph separation whether $A \perp B \mid C$ or not. ∘

As a basic concept, one could think of A being separated from B given C if all paths from A to B run through C - this is indeed the case of undirected graph separation defined below - but in general the concepts are more complex.

Classically several such separations have been developed, e.g. d-separation (Pearl 1988), the generalization m-separation (Richardson, Spirtes, et al. 2002) or δ-separation (Didelez 2008). The latter is presented here, and developed further into μ-separation (Mogensen, Malinsky, et al. 2018).

Definition 3.3. In an undirected graph $\mathcal{G}$, we say that A is separated from B given C if every path starting at a node $a \in A$ and ending at a node $b \in B$ intersects C. ∘

This enables us to introduce δ-separation:

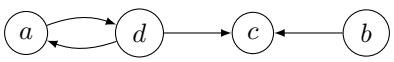
Definition 3.4. Let $\mathcal{G}$ be a DG and let A, B, C be disjoint. We say that A **is δ-separated from B given** C if A is separated from B by C in the undirected graph $(\mathcal{G}^B_{\mathrm{An}(A\cup B\cup C)})^m$.

That is, one considers the subgraph induced by the ancestors of A, B and C, removes edges leaving B and moralizes - and then considers separation in this undirected graph. Notationally we write $A \perp_\delta B \mid C$. ∘

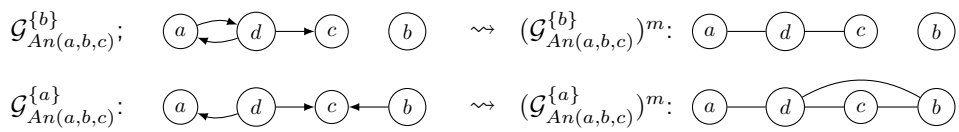
If A, B and C are not disjoint, one can define that A is δ-separated from B given C if $A \backslash (B \cup C) \perp_\delta B \mid C \backslash B$, although this is not in general very satisfactory - for instance if $A = B$, since one always has that $\emptyset \perp_\delta B \mid C \backslash B$, it is always the case that $A \perp_\delta A \mid C$.

Note the asymmetry in Definition 3.4 - contrary to other separation criteria used in graphical models, such as d-separation, the front-door thinning is asymmetric in A and B.

Example 3.5. For a simple example consider the graph:



To check whether $a \perp_\delta b \mid c$ and $b \perp_\delta a \mid c$, we form the moralized, frontdoor-thinned ancestral graphs.



Since there are no paths from a to d in $(\mathcal{G}^{\{b\}}_{An(a,b,c)})^m$, in particular no paths connect a and b without intersecting c. Hence, $a \perp_\delta b \mid c$. On the contrary, in $(\mathcal{G}^{\{a\}}_{An(a,b,c)})^m$, the path $a - d - b$ connects a and b without intersecting c, and so $b \not\perp_\delta a \mid c$. ∘

In many cases, it is easier to check δ-separation from the following equivalent characterization, which uses d-separation.

Definition 3.6. Let $\mathcal{G}$ be a DG with $a \neq b \in V$ and $C \subseteq V \backslash \{a, b\}$. A **$d$-connecting** path p from a to b given C is a path from a to b such that:

1. $\text{coll}(p) \subseteq \text{An}(C)$
2. $\text{noncoll}(p) \cap C = \emptyset$

If no path from a to b is d-connecting given C, we say that a is **d-separated** from b given C and write $a \perp_d b \mid C$. ∘

Proposition 3.7. *For a directed graph $\mathcal{G} = (V, E)$, let $A, B, C \subseteq V$ be disjoint. Then $A \perp_\delta B \mid C$ if and only if there are no d-connecting paths p between an $a \in A$ and $b \in B$ given C such that p contains an edge leaving the set B, i.e. an edge $i \to j$ for $i \in B, j \notin B$.*

Proof. A very detailed proof can be found in (Didelez 2000, Proposition 1.2.5).

Alternatively, it is a classic result (see e.g. (Lauritzen 1996, Proposition 3.25)) that d-separation is equivalent to undirected separation in the moralized ancestral graph, as used in Definition 3.4. The requirement in Definition 3.6 that no edge leaves B is simply analogous to the front-door thinning in Definition 3.4. □

Example 3.8. Consider again Example 3.5. Using the alternative formulation, we see that $a \perp_\delta b \mid c$, because any path from 1 to 4 would include the edge $c \leftarrow b$.

Also we observe that $b \not\perp_\delta a \mid c$: Let p be the path $a \leftarrow d \rightarrow c \leftarrow b$. p has no edges leaving a, $\text{coll}(p) = c \subseteq \text{An}(c)$ and $\text{noncoll}(p) = d$ which is disjoint with $\{c\}$. Thus p is connecting, and $b \not\perp_\delta a \mid c$. ∘

We now introduce μ-separation. First a bit of context: While δ-separation is useful in several cases, it has the disadvantage of assuming that A, B and C are disjoint. As motivated by Figure 2, we are interested in graphs where some, but not all, nodes have self-loops. To deal with a self-loop, one needs to consider separations of the type $a \perp a \mid C$, and so it is essential to have a separation criterion which can handle non-disjoint A and B.

This issue was addressed by (Meek 2014) by introduction of δ^*-separation, which was simply an extension of δ-separation to the non-disjoint cases, but as commented by (Mogensen and Hansen 2018), δ^*-separation aligned badly with local independence (which we introduce below) - in particular one is unlikely to obtain a global Markov property with respect to the local independence graph (which we also introduce below).

Instead (Mogensen, Malinsky, et al. 2018) introduce μ-separation, and for dynamic systems show regularity conditions, such that the local independence model will satisfy the global Markov property.

Definition 3.9. Let $\mathcal{G}$ be a DG with $a, b \in V, C \subseteq V$. A **$\mu$-connecting** walk p from a to b given C is a walk from a to b such that:

1. p is non-trivial and its final edge points to b.
2. $a \notin C$
3. $\text{coll}(p) \subseteq \text{An}(C)$
4. $\text{noncoll}(p) \cap C = \emptyset$

$\circ$

Definition 3.10. For a DG $\mathcal{G}$ we say that b is **μ-separated** from a given C (and write $a \perp_\mu b \mid C$) if no walk from a to b is μ-connecting given C.

Further, for subsets $A, B \subseteq V$, we say that B is μ-separated from A given C if for all $a \in A$ and all $b \in B$, $a \perp_\mu b \mid C$. $\circ$

Clearly μ-separation and δ-separation have very similar definitions. Indeed, if A, B, C are in fact disjoint, in the μ-separation definition, trivially $a \notin C$ and p will be non-trivial. Hence the only difference in this case is between the front-door thinning of δ-separation, and the 'head-at-b' of μ-separation. In fact, (Mogensen and Hansen 2018) show that:

$$A \perp_\delta B \mid C \iff A \perp_\mu B \mid B \cup C \tag{2}$$

for disjoint A, B, C. Consequently μ-separation can be seen as an extension of δ-separation - with the additional benefit, that we can discuss non-disjoint separations like $A \perp_\mu A \mid C$, without having a trivial answer (as is the case for δ-separation).

Example 3.11. Consider again the graph from Definition 3.6, where we verified that $a \perp_\delta b \mid c$ and $b \not\perp_\delta a \mid c$. This is easily reproducible with μ-separation. For instance, since no walk from a to b has a final edge pointing to b, $a \perp_\mu b \mid (b, c)$, as is also predicted by Equation (2).

But for non-disjoint sets, this is not the case. Trivially $a \perp_\delta a \mid c$, while the path $a \to d \to a$ is μ-connecting given c and so $a \not\perp_\mu a \mid c$. $\circ$

To further illustrate this concept, we consider the following two examples, both which may generate conclusions counter-intuitive, to a reader accustomed to dealing with d-separation in DAGs.

Example 3.12. Consider the 2 different graphs in the first column of Table 1. We consider the μ-separations in these. Of particular notice is that $a \perp_\mu a \mid b$ in $a \to b$. A first intuition would be that since collider b on the walk $a \to b \leftarrow a$

Graph	**Given** $\emptyset$	**Given** b
$a \to b$	$a \perp_\mu a \mid \emptyset$	$a \perp_\mu a \mid b$
$a \leftarrow b$	$a \not\perp_\mu a \mid \emptyset$	$a \perp_\mu a \mid b$

Table 1: μ-separation statements of a from a in 2 different graphs and given either $\emptyset$ or b.

is in the conditioning set, this path is μ-connecting. However, since the final arrow does not have a head at a, this path is not connecting. As we shall later, this will become reasonable, when dealing with a process that does not depend on itself. ∘

Example 3.13. Consider the two graphs in Example 3.13 below. We seek to examine whether $b \perp_\mu a \mid \emptyset$.

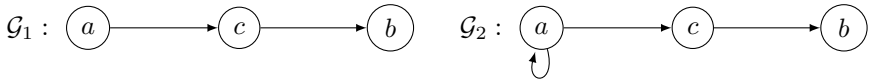


Figure 3

In $\mathcal{G}_1$, this is indeed true, since any walk from b to a will not have a final head at a, and so cannot be μ-connecting. On the contrary, in $\mathcal{G}_2$, the walk $b \leftarrow c \leftarrow a \rightarrow a$ is μ-connecting, and so $b \not\perp_\mu a \mid \emptyset$.

Thus the occurrence of a self-loop at a is crucial for whether the empty set is blocking paths from b to a. ∘

3.3 Markov properties

We introduce the pairwise and global Markov properties, defined for abstract separation-criteria and independence models.

As discussed above, Markov properties are essentially a way of formulating that a graph $\mathcal{G}$ contains the relevant information about the independence structure of some distribution P. Typically we phrase the independence structure of P into an independence model $\mathcal{I}$.

Definition 3.14. For a finite collection of vertices V, an **independence model** $\mathcal{I}$ is a ternary relation on V, consisting of triples $A, B, C \subseteq V$, which we write $\langle A, B \mid C \rangle$. Formally:

$$\mathcal{I} \subseteq \mathcal{P}(V \times V \times V)$$

∘

That is for each combination $A, B, C \subseteq V$, either $\langle A, B \mid C \rangle \in \mathcal{I}$ or $\langle A, B \mid C \rangle \notin \mathcal{I}$.[4] This allows us to define the global Markov property and faithfulness:

Definition 3.15. For a graph $\mathcal{G} = (V, E)$, an independence model $\mathcal{I}$ and a graph separation criterion $\perp$, we say that $\mathcal{I}$ has the **global Markov property**

[4]The obvious motivation for independence models is to define the independence model that corresponds to the conditional independences of a distribution P over a collection of random variables $X_1, \ldots, X_d$ by

$$\langle A, B \mid C \rangle \in \mathcal{I} \iff X_A \perp\!\!\!\perp_P X_B \mid X_C$$

where $A, B, C \subseteq V$ denotes collections of variables.

with respect to $\mathcal{G}$ and $\perp$ if:

$$A \perp B \mid C \implies \langle A, B \mid C \rangle \in \mathcal{I}$$

for all $A, B, C \subseteq V$. ∘

Consequently, if $\mathcal{I}$ satisfies the global Markov property with respect to $\mathcal{G}$, $\mathcal{G}$ contains relevant information, in the sense that several independence statements of $\mathcal{I}$ can be read off from $\mathcal{G}$ as separations.

As demonstrated in Section 2.2, there may be independences of $\mathcal{I}$, which cannot be read from $\mathcal{G}$. But if $\mathcal{I}$ satisfies faithfulness with respect to $\mathcal{G}$, all the independences of $\mathcal{I}$ are included in the separations of $\mathcal{G}$:

Definition 3.16. Given a graph $\mathcal{G} = (V, E)$, an independence model $\mathcal{I}$ and a graph separation criterion $\perp$. We say that $\mathcal{I}$ is **faithful** with respect to $\mathcal{G}$ and $\perp$ if:

$$\langle A, B \mid C \rangle \in \mathcal{I} \implies A \perp B \mid C$$

for all $A, B, C \subseteq V$. ∘

Besides the global Markov property and faithfulness, a number of Markov properties can be defined for a graph, depending on the type of graph. One example is the pairwise Markov property.

Definition 3.17. Let $\mathcal{G} = (V, E)$ be a directed graph (DG), and $\mathcal{I}$ an independence model. $\mathcal{I}$ satisfies the **pairwise Markov property** with respect to $\mathcal{G}$ if:

$$a \notin \mathrm{pa}(b) \implies \langle a, b \mid V \backslash \{a\} \rangle \in \mathcal{I}$$

∘

Definition 3.18. Let $\mathcal{I}$ be an independence model over V. Letting

$$E = \{(a, b) \mid \langle a, b \mid V \backslash \{a\} \rangle \notin \mathcal{I}\},$$

one can define a **local independence graph** by the directed graph $\mathcal{G} = (V, E)$. ∘

Note that an independence model $\mathcal{I}$ will satisfy the pairwise Markov property with respect to the local independence graph $\mathcal{G} = (V, E)$ it induces, because if $a \notin \mathrm{pa}(b)$, then $(a, b) \notin E$ and this is only the case if $\langle a, b \mid V \backslash \{a\} \rangle \in \mathcal{I}$.

So for any independence structure $\mathcal{I}$, one can can get a directed graph $\mathcal{G}$ such that $\mathcal{I}$ satisfies the pairwise Markov property wrt. $\mathcal{G}$. A central question is: For a given independence model $\mathcal{I}$, are there separation criteria $\perp$ such that we are guaranteed to have the global Markov property of $\mathcal{I}$ with respect to $\mathcal{G}$ and $\perp$? If this is indeed the case, we can explore the independence structure of $\mathcal{I}$ simply by reading of the separations of $\mathcal{G}$.

In Section 6 below, we define *local independence*, which is forms an independence model $\mathcal{I}$ for point processes. (Didelez 2008) showed that, under mild regularity assumptions, indeed $\mathcal{I}$ satisfies the global Markov property with respect to the local independence graph $\mathcal{G}$ and $\perp_\delta$. Further (Mogensen and Hansen 2018) argue that when extending $\mathcal{I}$ to include also non-trivial statements regarding overlapping sets, $\mathcal{I}$ still satisfies the global Markov property with respect to $\mathcal{G}$ and $\perp_\mu$.

3.4 CA Algorithm

We now present the CA algorithm as proposed in (Meek 2014). It resembles the PC-algorithm discussed in Section 2.2, but where the PC-algorithm learns undirected edges first and then orientates them, the CA directly learns the directed edges (which essentially is due to the asymmetry of local independence). For now it is assumed that an oracle test is available, i.e. that we have access to the true answers of whether $\langle v', v \mid C\rangle \in \mathcal{I}$ or not.

In a real application, typically one does not know $\mathcal{I}$, and so a major ambition of this paper is to develop empirical tests which can replace this oracle test. See Section 7 for more. Given such an empirical test, this is simply plugged into the below algorithm, to replace the test "**if** $\langle v', v \mid C\rangle \in \mathcal{I}$".

Algorithm 1 Causal Analysis algorithm

Initialize $\mathcal{G} = (V, E_{\text{CA}})$ as a fully connected graph
for $v \in V$ **do**:
  $n = 0$
  while $n < |\text{pa}(v)|$ **do**:
    for $v' \in \text{pa}(v)$ **do**:
      for $C \subseteq \text{pa}(v)\backslash\{v'\}$ with $|C| = n$ **do**:
        if $\langle v', v \mid C\rangle \in \mathcal{I}$ **then** update $\mathcal{G}$ by removing (v', v) from E_{CA}.
    n = n + 1
return $\mathcal{G} = (V, E_{\text{CA}})$

Let $\mathcal{G}_{\text{CA}} = (V, E_{\text{CA}})$ denote the graph returned by algorithm Algorithm 1. If $\mathcal{G}_0 = (V, E_0)$ is the local independence graph of $\mathcal{I}$, we hope that $\mathcal{G}_{\text{CA}}$ equals $\mathcal{G}_0$. The below results from (Meek 2014), gives guarantees in the fully observable case. While (Meek 2014) uses δ^*-separation, we propose the following two purely graphical properties, such that the result generalize to abstract separation.

Definition 3.19. Let $\perp$ be a graph separation. We say that $\perp$ satisfies

P1 If whenever there exist $C \subseteq V\backslash v'$, such that $v' \perp v \mid C$, it follows that $(v', v) \notin E$.

P2 If whenever an edge (v', v) is not in E, it follows that there exist $C \subseteq \text{pa}(v)$, such that $v' \perp v \mid C$.

$\circ$

Lemma 3.20. *d-, δ- and μ-separation in DGs satisfy **P1**. δ- and μ-separation in DGs satisfy **P2**.*

Proof. To show **P1**, suppose $(v', v) \in E$ (in the d- and δ- cases, we only need to consider $v' \neq v$). Then the directed path $v' \to v$ neither has colliders nor non-colliders, and so the path is d-, δ- or μ-connecting for any $C \subseteq V \backslash v'$. By contraposition, **P1** follows.

Suppose $(v', v) \notin E$. We show **P2** by showing that $v' \perp v \mid \mathrm{pa}(v)$. To show δ-case, we use the formulation of Proposition 3.7. If $v' = v$, trivially $v' \perp_\delta v \mid \emptyset$. Else, we must show that no path p from v' to v which has no arrows leaving v is d-connecting given $\mathrm{pa}(v)$. Let $c \to v$ be the last edge (apart from potentially some self-edges). Since c is a non-collider, and $c \in \mathrm{pa}(v)$, it follows that p is not d-connecting. Hence $v' \perp_\delta v \mid \mathrm{pa}(v)$.

In the μ-case, similarly, any potentially μ-connecting walk p from v' to v would have as its last edge $c \to v$. But again c is a non-collider, and $c \in \mathrm{pa}(v)$, and so p is not μ-connecting given $\mathrm{pa}(v)$, and $v' \perp_\mu v \mid \mathrm{pa}(v)$. □

If there could be any ambiguity, we let $A \perp B \mid C \; [\mathcal{G}]$ to denote a graph separation in the graph $\mathcal{G}$. Also $\mathrm{pa}_\mathcal{G}(v)$ denotes the parents of v relative to the graph $\mathcal{G}$.

Lemma 3.21. *If $\perp$ satisfies **P1** and $\mathcal{I}$ satisfies faithfulness wrt. $\mathcal{G}_0 = (V, E_0)$ and $\perp$, then $E_0 \subseteq E_{CA}$.*

Proof. Take an edge $e = (v', v)$ in E_0. If, for a contradiction, $e \notin E_{\mathrm{CA}}$, by the design of the algorithm, there exists a set $C \subseteq V \backslash \{v'\}$ such that $\langle v', v \mid C \rangle \in \mathcal{I}$. By faithfullness this implies that $v' \perp v \mid C \; [\mathcal{G}_0]$. By assumption of **P1**, this implies $e \notin E_0$, which is a contradicition. □

Proposition 3.22. *If $\perp$ satisfies **P1** and **P2** and $\mathcal{I}$ satisfies both faithfulness and the global Markov property with respect to $\mathcal{G}_0$ and $\perp$, then $E_{CA} = E_0$.*

Proof. For a contradiction, suppose that there exists $e = (v', v) \in E_{\mathrm{CA}} \backslash E_0$. Since $e \notin E_0$, by **P2**, there exists a set $C \subseteq \mathrm{pa}_{\mathcal{G}_0}(v) \backslash \{v'\}$[5], such that $v' \perp v \mid C \; [\mathcal{G}_0]$. Hence, by the global Markov property, $\langle v', v \mid C \rangle \in \mathcal{I}$.

By assumption of faithfulness and **P1**, it follows from Lemma 3.21 that $E_0 \subseteq E_{\mathrm{CA}}$. In particular $\mathrm{pa}_{\mathcal{G}_0}(v) \subseteq \mathrm{pa}_{\mathcal{G}_{\mathrm{CA}}}(v)$ and so $C \subseteq \mathrm{pa}_{\mathcal{G}_{\mathrm{CA}}}(v) \backslash \{v'\}$. But then at some point, the algorithm iterator n would have taken the value $|C|$ for the combination (v', v), and because $\langle v', v \mid C \rangle \in \mathcal{I}$, (v', v) would be removed from E_{CA}, contradicting that $e \in E_{\mathrm{CA}}$. It follows that $E_{\mathrm{CA}} \subseteq E$, and thus $E = E_{\mathrm{CA}}$. □

Hence in the fully observable case, with an oracle local independence test and under the assumptions of Proposition 3.22, Algorithm 1 is guaranteed to return the correct local independence graph.

[5] which is really just $\mathrm{pa}_{\mathcal{G}_0}(v)$, because since $e \notin E, v' \notin \mathrm{pa}_{\mathcal{G}_0}(v)$

A very brief note on Markov equivalences

In most of the graphical model literature, Markov equivalence is discussed. Although we shall not discuss this at length, we show here, that for μ-separation in DGs, Markov equivalence yields no identifiability issues:

Two graphs $\mathcal{G}_1 = (V, E_1), \mathcal{G}_2 = (V, E_2)$ are **Markov equivalent**, if they have the same separation statements. As a simple example, $a \to b \to c$ and $a \leftarrow b \leftarrow c$ are Markov equivalent with respect to d-separation (essentially since for any path p, $\text{coll}(p)$ and $\text{noncoll}(p)$ are the same for both graphs).

Let $[\mathcal{G}]$ be the equivalence class consisting of all graphs that are Markov equivalent to $\mathcal{G}$. If an independence model $\mathcal{I}$ satisfies the global Markov property with respect to $\mathcal{G}$, $\mathcal{I}$ satisfies it with respect to all graphs in $[\mathcal{G}]$. In the context of learning algorithms, this leads to identifiability problems, since algorithms like Algorithm 1 can only hope to identify $[\mathcal{G}_0]$ and not $\mathcal{G}_0$.

However Lemma 3.20 and Proposition 3.23 shows that for μ-separation in DGs, no two distinct graphs can have the same separation statements, and so every Markov equivalence class only contains one element.

Proposition 3.23. *Let $\perp$ be a graph separation satisfying **P1** and **P2**. Also let $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ be two directed graphs. Then $\mathcal{G}_1 = \mathcal{G}_2$ if and only if they have all the same graph separations. Formally, $\mathcal{G}_1 = \mathcal{G}_2$ if and only if for all $A, B, C \subseteq V$:*

$$A \perp B \mid C \; [\mathcal{G}_1] \iff A \perp B \mid C \, [\mathcal{G}_2]$$

where the notation $[\mathcal{G}]$ denotes separation in the graph $\mathcal{G}$.

Proof. If $\mathcal{G}_1 = \mathcal{G}_2$, trivially they also have the same separations. For the other direction, suppose for a contradiction that $\mathcal{G}_1 \neq \mathcal{G}_2$ but they have the same separations. Without loss of generality, there is an edge $a \to b$ in $\mathcal{G}_1$, which is not in $\mathcal{G}_2$.

By **P2**, there exist $C \subseteq \text{pa}_{\mathcal{G}_2}(b)$ such that $a \perp b \mid C \; [\mathcal{G}_2]$. Since a is not a parent of b in $\mathcal{G}_2$, $a \notin C$. Since $\mathcal{G}_1$ and $\mathcal{G}_2$ share separations, also $a \perp b \mid C \; [\mathcal{G}_1]$. But then by **P1**, there is no edge $a \to b$ in $\mathcal{G}_1$, which is a contradiction. □

4 Point Processes

In this section, we introduce point processes, the fundamental class of models for which we develop methods in this thesis.

In Section 4.1 we introduce the theoretical foundations for point processes. In Section 4.2 we introduce some martingale theory leading to the compensator, and in Section 4.3 we discuss intensities of point processes. Section 4.4 introduces Hawkes processes, and Section 4.5 regards Marked Point Processes, a particular multivariate form of point processes. Finally in Section 4.6 we introduce the likelihood of a point process, which is to become important for statistical purposes in the following chapters.

4.1 Random measures

Roughly, a point process is a stochastic process describing how a number of points appear in a space $\mathcal{X}$. We will always assume that $\mathcal{X}$ is a complete, separable, metric space, and call it the **outcome space**. Very often $\mathcal{X}$ will be equal to $\mathbb{R}$ or a subset thereof, and we will pay most of our attention to this case. One should however keep in mind that the theory of point processes, including several of the theorems, extends well beyond $\mathbb{R}$.

Definition 4.1. A measure μ on $\mathcal{X}$ (equipped with the Borel σ-field $\mathcal{B}(\mathcal{X})$) is called **boundedly finite** if for every bounded $A \in \mathcal{B}(\mathcal{X})$, $\mu(A) < \infty$. ∘

Let $\mathcal{N}_{\mathcal{X}}^{\#}$ denote the set of all boundedly finite integer-valued measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. As discussed in (Daley and Vere-Jones 2007, chapter 9), one can equip $\mathcal{N}_{\mathcal{X}}^{\#}$ with a meaningful metric, and hence a Borel field $\mathcal{B}(\mathcal{N}_{\mathcal{X}}^{\#})$. Implicitly we shall utilize this over and over again, but we shall only make explicit reference once, namely in the following definition:

Definition 4.2. For a complete separable metric space $\mathcal{X}$ and background space $(\Omega, \mathcal{F}, P)$, a **point process** is a $\mathcal{F}/\mathcal{B}(\mathcal{N}_{\mathcal{X}}^{\#})$-measurable mapping $N : \Omega \to \mathcal{N}_{\mathcal{X}}^{\#}$.

That is, for a fixed $A \in \mathcal{B}(\mathcal{X})$, $\omega \mapsto N_\omega(A)$ is a mapping from Ω to $\mathbb{N}_0 \cup \{\infty\}$. Oftentimes we suppress the dependence on ω from the notation, viewing $N(A)$ as a random variable on $\mathbb{N}_0 \cup \{\infty\}$ (with support on $\mathbb{N}_0$ if A is bounded). ∘

One could in fact consider more generally mappings to the set of boundedly-finite, but not necessarily integer-valued, measures on $\mathcal{X}$. (Daley and Vere-Jones 2007) denotes these **random measures**, and several results generalize from point processes to random measures. We shall however restrict our attention to point processes.

Point processes has a natural interpretation as the count of events occurring randomly in $\mathcal{X}$. For a point process N and a fixed ω, N_ω is an integer-valued measure, and so for a measurable set A, one could regard $N_\omega(A)$ as the count of events occurring in this set. When not keeping ω fixed, $N(A)$ is a random value on $\mathbb{N} \cup \{\infty\}$, which one could regard as being the distribution of the number of events in A.

Definition 4.3. For any (deterministic) boundedly finite measure Λ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we say that N is a **Poisson process** with **intensity measure** Λ if for any bounded disjoint sets $B_1, \ldots, B_m \in \mathcal{B}(\mathcal{X})$ and $n_1, \ldots, n_m \in \mathbb{N}_0$, one has that the joint distribution satisfies:

$$P\Big(N(B_1) = n_1, \ldots, N(B_m) = n_m\Big) = \prod_{j=1}^{m} \frac{\Lambda(B_j)^{n_j} e^{-\Lambda(B_j)}}{n_j!}$$

$\circ$

Essentially, a Poisson process is a point process satisfying two requirements: The count $N(A)$ of any bounded set follows a Pois$(\Lambda(A))$-distribution, and the counts in disjoint intervals are independent.

If $\mathcal{X} = \mathbb{R}$ (or possibly $\mathbb{R}^n$) and Λ is proportional to the Lebesgue measure ℓ, we will denote the process a **homogeneous** Poisson process, denoting the proportionality constant by the *rate*. If Λ is not globally proportional to ℓ, we denote this the **inhomogeneous** case.

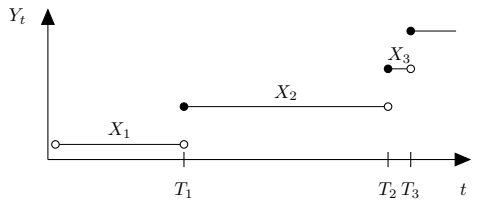


Figure 4: Illustration of the construction in Example 4.4, where the homogeneous Poisson process is created from exponentially distributed waiting times X_n.

Example 4.4. Consider the case with $\mathcal{X} = \mathbb{R}$, and let N be a Poisson process with intensity measure $\Lambda = \ell$. Now $t \mapsto N_t := N(0, t]$ defines a stochastic process, where each N_t marginally is Pois(t) distributed.

A very natural way to think of this process is as follows: Let $X_1, X_2, \ldots$ be i.i.d. exponentially distributed with mean 1, let $T_n = \sum_{j=1}^{n} X_j$ and let $Y_t = \sum_{n \in \mathbb{N}} \mathbb{1}_{T_n \leq t}$. This construction is illustrated in Figure 4. Now for any $m \in \mathbb{N}$:

$$P(Y_t \geq m) = P(T_m \leq t) = 1 - \sum_{n=0}^{m-1} \frac{e^{-t}}{n!} t^n$$

The latter follows because T_m follows an Erlang$(1, m)$-distribution, with this CDF. By rearranging, one recognizes from the CDF of Y_t that Y_t is Pois(t)-distributed for each fixed t, and Y thus has the same marginal distributions as N. In a similar fashion, essentially utilizing the 'forgetfullness' of the exponential distribution, one could show that the finite dimensional distributions of Y match those of N, and so Y and N have the same distribution.

Thus, one can view N_t as the number of events occurring before time t, each event taking an exponentially distributed time to arrive after the former. Heuristically and when simulating, this characterization of a unit rate Poisson process is very helpful. The same result follows if $\Lambda = \lambda \cdot \ell$ for some constant λ and X_n exponentially distributed with mean $\frac{1}{\lambda}$ (and to some extent for an inhomogeneous process as well, as exemplified in the *inverse hazard* simulation scheme (see Hansen 2013)). ∘

Definition 4.5. A point process N is called **simple** if

$$P\bigg(N(\{x\}) < 2 \text{ for all } x \in \mathcal{X}\bigg) = 1$$

∘

Note that the notion of simplicity is *almost everywhere* with respect to the probability space, but *everywhere* with respect to the outcome space.

If point processes constitute an intuition of counting events in $\mathcal{X}$, simpleness can be understood as the requirement that no two events can occur in the same location. Simplicity of a process is very desirable, and we shall very often assume the process to be simple.

4.2 Some martingale theory

We now restrict our attention to the temporal case $\mathcal{X} = \mathbb{R}$, where counting measures are easily interpreted as stochastic processes, as was the case of $N_t = N(0, t]$ in Example 4.4. Though many authors consider temporal processes on $(0, \infty)$, we here do so on $\mathbb{R}$. This eases some of the 'boundary'-problems arising from point processes started at 0, but on the other hand poses some challenge in the existence of a process. For instance when dealing with a Hawkes process, as we are to define below, one needs to exercise care. It is possible for the intensity of the process to grow approximately exponentially towards infinity (without the process exploding). While this process would be perfectly well-defined on $(0, \infty)$, it is not on $\mathbb{R}$.

Often we shall consider a counting measure N on $\mathbb{R}$, but only the counting process $N_t = N(0, t]$ of events after 0. To incorporate knowledge of prior events, we could consider the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0} = (\sigma(\mathcal{K}_t \cup \mathcal{H}_0))_{t\geq 0}$ where $\mathcal{H}_0$ is the σ-field generated by the jumps of N prior to 0 and $\mathcal{K}_t$ is the σ-field generated by the process on the interval $[0, t]$.

Cadlag martingales

A central tool in temporal point processes is martingale theory. Consider a filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$, and let $X = (X_t)_{t\geq 0}$ be an adapted stochastic process. We say that X is a **martingale** if $X_t \in \mathcal{L}^1(\mathcal{F}_t)$ for all $t \geq 0$ and for any $s \leq t$:

$$\mathbb{E}\left[X_t \mid \mathcal{F}_s\right] \overset{(*)}{=} X_s \qquad a.s.$$

or a **submartingale or supermartingale** if $(*)$ is $\geq$ or $\leq$ respectively.

Example 4.6. A counting process $N_t := N(0,t]$ is a submartingale if $t \mapsto N_t$ is adapted to $\mathcal{F}_t$ and $\mathbb{E}N(0,t] < \infty$ for all $t > 0$. This follows because

$$\begin{aligned}\mathbb{E}[N_t \mid \mathcal{F}_s] &= \mathbb{E}\left[N(0,s] \mid \mathcal{F}_s\right] + \mathbb{E}\left[N(s,t] \mid \mathcal{F}_s\right] \\ &= N_s + \mathbb{E}\left[N(s,t] \mid \mathcal{F}_s\right] \geq N_s \qquad a.s.\end{aligned}$$

The last inequality follows from the fact that $N(s,t] = N_t - N_s \geq 0$ a.s. - in fact this is the only point process property we are using, and so the above in general shows that adapted increasing processes are (unsurprisingly!) submartingales. ∘

An nice property in analysis of stochastic processes is for the process to be right continuous with a limit from the left, i.e. almost surely:

$$X_t = \lim_{\epsilon \downarrow 0} X_{t+\epsilon}, \qquad X_{t-\epsilon} \text{ converges for } \epsilon \downarrow 0$$

for all $t \geq 0$. Processes with this property are often denoted **càdlàg**, an acronym for the French *continue à droite, limite à gauche*. The case of left-continuity and right-limit, one similarly denotes **càglàd**.

Example 4.7. A process being càdlàg is often a question of choosing a càdlàg version. Two stochastic processes X and Y are **versions** of each other if $P(X_t = Y_t) = 1$ for all $t \geq 0$ - which is subtly different from them being indistinguishable, in which case $P(X_t = Y_t \text{ for all } t \geq 0) = 1$.

Consider, as in Example 4.4, a unit rate Poisson process N, and let $N_t := N(0,t]$. By boundedly-finiteness, N_t is càdlàg: For any $t > 0$, almost surely $N(0,t]$ is finite, and so for each ω, an ϵ exists such that no events occur in $(t-\epsilon, t)$. In particular the process $N(0, t-\epsilon]$ coming from the left will, for each ω, be constant on this interval, and hence converge. Similarly, N_t almost surely is right-continuous.

Instead, one could have chosen $\tilde{N}_t := N(0,t)$ instead, which would be càglàd, by similar arguments. Since the Poisson process has an intensity measure that is absolutely continuous with respect to the Lebesgue measure (in particular in the unit-rate case it *equals* the Lebesgue measure), for all $t \geq 0$, $P(\tilde{N}_t = N_t) = 1$. The fact that at any time t, the two will be almost surely equal, motivates us to think of them as one process, but in two different versions. Note that since the Poisson process will almost surely have jumps, $P(N_t = \tilde{N}_t \text{ for all } t \geq 0) = 0$. ∘

Oftentimes we are interested in having a càdlàg counting process with an intensity (to be defined below) which is càglàd. A related concept is that of a predictable process:

Definition 4.8. Given a filtration $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$, the **predictable** σ-field is the σ-field on $\Omega \times [0,\infty)$ generated by all left continuous adapted processes (understood as maps from $\Omega \times [0,\infty) \to \mathbb{R}$). We say that a process X is predictable, if the mapping $(\omega, t) \mapsto X_t(\omega)$ is measurable with respect to the predictable σ-field. ∘

Since càglàd processes are in particular left-continuous, any càglàd adapted process is predictable. It can be shown that the predictable sigma-field also is generated by the sets $A \times \{0\}$ for $A \in \mathcal{F}_0$ and $A \times (s,t]$ for $A \in \mathcal{F}_s, s > 0$.

Compensators

A slightly weaker concept than that of martingales are local martingales:

Definition 4.9. An adapted process X is called a **local martingale** if there exists an almost surely increasing sequence $(\tau_n)_{n\in\mathbb{N}}$ of stopping times such that

$$P(\tau_n \to \infty \text{ for } n \to \infty) = 1$$

and for each n, the stopped process $X^{\tau_n} := (X_{\tau_n \wedge t})_{t\geq 0}$ is a martingale. ∘

Note that if X is indeed a true martingale, by choosing $\tau_n = \infty$ for all n, it is also local. Local martingales enable us to introduce the compensator, in the following combined theorem and definition.

Theorem 4.10 (Doob-Meyer)**.** *Let N be a counting process. Then there exists a unique (up do indistinguishability), increasing, predictable process A, denoted the **compensator**, and a local martingale M, such that for all $t \geq 0$:*

$$N_t = M_t + A_t \tag{3}$$

Proof. This follows from (Jacod and Shiryaev 1987, I.1.Theorem 3.18), utilizing that (non-exploding) counting processes are increasing and of locally integrable variation. □

The compensator is an important tool in describing N. In particular, since local martingales have constant marginal means for all $t \geq 0$:

$$\mathbb{E}[M_t] = \mathbb{E}[M_t^t] = \mathbb{E}\left[\mathbb{E}[M_t^t \mid M_0^t]\right] = \mathbb{E}[M_0^t] = \mathbb{E}[M_0]$$

it follows from Equation (3) that $\mathbb{E}[N_t] = \mathbb{E}[A_t]$ for all t. Being monotone, predictable and in many cases continuous, the compensator is often more well-behaved than N, but still preserves the mean structures of N.

4.3 Intensities

Definition 4.11. For a process N with compensator A, we say that the process λ is the **intensity** of N if for all $t > 0$, $\int_0^t \lambda_s \mathrm{d}s$ is a version of A_t. ∘

The following proposition, which appears in (Daley and Vere-Jones 2007, Corollary 14.1.V), yields a guarantee on when a process allows for an intensity.

Proposition 4.12. *Let N be a simple counting process on $\mathbb{R}_+$ with events $\{T_n\}$ and compensator A. The conditional distribution of $T_n - T_{n-1} \mid \mathcal{F}_{T_{n-1}}$ is absolutely continuous with respect to the Lebesgue measure if and only if A is absolutely continuous. In this case, N admits a density, i.e. $A_t = \int_0^t \lambda_s ds$ for some process λ.*

In short, if the waiting times to the next event are absolutely continuous, the compensator allows for an intensity.

Although one can actually derive a closed form expression for λ based on the densities of the conditional distributions $T_n - T_{n-1} \mid \mathcal{F}_{T_{n-1}}$ (in particular the hazard ratio of this distribution), in some cases, one can simply guess a compensator A and check that $N - A$ is martingale.

Example 4.13. Consider a Poisson process N as introduced in Definition 4.3, and suppose that $\Lambda = \lambda\ell$, for some $\lambda > 0$. That is, N is a homogeneous Poisson process with rate λ. Let $N_t = N(0,t]$ be the process that counts the events up until time t. It turns out that the intensity of this process is also λ.

To see this, we guess the compensator $A_t = \int_0^t \lambda \mathrm{d}s = \lambda t$ for $t \geq 0$. Since A is continuous, it is in particular left-continuous, and thus predictable. N_t is $\mathrm{Pois}(\lambda t)$-distributed, so $N_t - A_t \in \mathcal{L}^1$ for all t. Now for any $s \leq t$,

$$\begin{aligned}
\mathbb{E}\left[N_t - A_t \mid \mathcal{F}_s\right] &= \mathbb{E}\left[N(0,t] \mid \mathcal{F}_s\right] - \lambda t \\
&= N(0,s] + \mathbb{E}\left[N(s,t] \mid \mathcal{F}_s\right] - \lambda(t-s) - \lambda s \\
&= N_s - A_s + \mathbb{E}[N(s,t]] - \lambda(t-s) \\
&= N_s - A_s
\end{aligned}$$

We here utilize the properties of a Poisson process. First off, the count in $(s,t]$ is independent of $\mathcal{F}_s$, because counts in disjoint regions are independent. Secondly, $N(s,t]$ follows a $\mathrm{Pois}(\lambda(t-s))$-distribution, and so $\mathbb{E}[N(s,t]] = \lambda(t-s)$.

Of course this result generalizes to the inhomogeneous Poisson process with $\Lambda = \lambda \cdot \ell$.

Also note that the intensity of the Poisson process, λ, is deterministic. In general however, λ will most often be stochastic, as we shall for instance see with the Hawkes process below. ∘

Defining processes by intensities

In the martingale setting above, we have viewed the intensity purely as a quantity derived from the process. Often we shall work reversely: Instead of assuming first a process, we will specify an intensity, and consider processes with this intensity. One has to be a bit careful in doing this: The intensity will be a function of the process it generates, and so the development of the two is heavily entangled.

As will be evident below, we need to start our process somewhere. To still allow for a process on all of $\mathbb{R}$, we let $\mathbb{T} = (\ldots, T_{-1}, T_0)$ be a collection of past events. Suppose that $\lambda_t = \lambda(t, \{N_s\}_{0 \leq s < t}, \mathbb{T})$ is a mechanism, which takes as its input $\mathbb{T}$ and the development of a process N_s on $0 < s < t$. Our intention is then to find a process N such that λ_t is the intensity of N at time t. For some fixed mechanism λ, can we be sure, that a such N exists, such that the intensity of N at all times is λ evaluated in N?

Theorem 4.14 below guarantees the existence of a process satisfying a particular integral equation and Proposition 4.16 shows that such a process indeed has λ as its intensity.

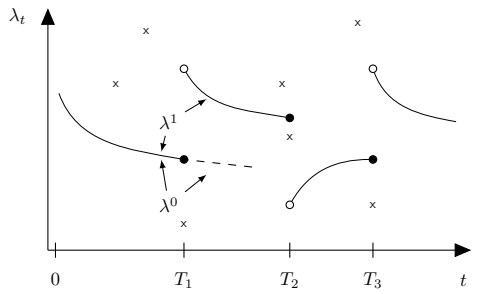


Figure 5: Illustration of a sample path of λ. Also the points of μ, a unit rate Poisson process, are shown as crosses. Our proposed process N is constructed via the points of μ which appear below the λ-curve, here the points at T_1, T_2 and T_3. The proof of Theorem 4.14 below relies heavily on λ being càglàd, such that λ at T_2 cannot 'jump below the point that made it jump' (which would create somewhat of a chicken-or-the-egg paradox). Indeed λ jumps down, but only at times $t = T_2 + \epsilon$ for $\epsilon > 0$. In this example $L_t(\lambda)$ is the region below the curve, and $\mu(L_t(\lambda)) = 3$ in the plotted region. Labels λ^0 and λ^1 are used for reference below.

Consider some positive stochastic process Y. The sample paths of Y define regions under the curve. On an interval $(0, t]$ the region under the curve is:

$$L_t(Y) := \left\{(s, y) \in \mathbb{R}^2 \mid 0 < s \leq t, 0 \leq y \leq Y_s\right\}$$

Now suppose that μ is a unit rate Poisson process on $\mathbb{R}^2$. Then $\mu(L_t(Y))$ is the number of points arriving under the Y-curve in the interval $(0, t]$. Naturally, the larger the values of Y, the more points appear in $L_t(Y)$.

The following result, which appears in (Hansen 2013), essentially has $Y_t = \lambda$ and lets N be the process $t \mapsto \mu(L_t(\lambda))$.

Theorem 4.14. *Suppose a history of past events $\mathbb{T} = (\ldots, T_{-1}, T_0)$ is given.*

Further assume that $\lambda_t = \lambda(t, \{N_s\}_{0 \leq s < t}, \mathbb{T})$ is a mechanism, such that $t \mapsto \lambda(t, \{N_s\}_{0 \leq s < t}, \mathbb{T})$ defines a càglàd process for any càdlàg counting process N. Also let μ be a unit rate homogeneous Poisson process on $\mathbb{R}^2$. Then there exists a stopping time T such that on $[0, T)$ there exists a point process N which for all $t \in [0, T)$ solves the stochastic integral equation:

$$N_t = \int \mathbb{1}_{(0,t]}(s) \mathbb{1}_{[0, \lambda(s, \{N_u\}_{0 \leq u < s}, \mathbb{T})]}(y) \mu(ds \times dy)$$

or in short

$$N_t = \mu(L_t(\lambda))$$

for $t \in [0, T)$ (though this conceals the dependence of λ on N). In the case $T < \infty$, then $N(0, T - \epsilon] \to \infty$ for $\epsilon \downarrow 0$.

Remark. Actually, it is slightly inaccurate to call N a point process (in our definition) in the case $T < \infty$, because the resulting process is not boundedly-finite. However, $N(0, T - \epsilon]$ is finite for any $\epsilon > 0$. ∘

Proof. One creates a sequence of processes $N^0, N^1, \ldots$ and corresponding $\lambda_t^n = \lambda(t, \{N^n\}_{0 \leq u < t), \mathbb{T}})$ as follows: N^0 is set to constant 0, and the following processes are defined by letting

$$T_n = \inf_t \mu(L_t(\lambda^{n-1})) > N_t^{n-1} \tag{4}$$

and $N^n = \sum_{j=1}^n \mathbb{1}_{[T_j, \infty)}$.

By Equation (4), $N_t^0 = \mu(L_t(\lambda^0))$ for $t < T_1$, and so N^0 is a solution on $[0, T_1)$. However also by Equation (4), N^0 is not a solution in the point T_1, and so we create $N^1 = \mathbb{1}_{[T_1, \infty)}$.

Since $N^0 = N^1$ on $[0, T_1)$, also $\lambda_t^0 = \lambda_t^1$ on $[0, T_1]$, where the end-point follows since both are càglàd, hence left-continuous. Thus also $\mu(L_t(\lambda^0)) = \mu(L_t(\lambda^1))$ on $[0, T_1]$, and this allows us to know that $\mu(L_{T_1}(\lambda^1)) = 1$, and so indeed N^1 is a solution on all of $[0, T_1]$.

One could proceed this inductively, again using that $N_t^n = \mu(L_t(\lambda^n))$ for $t \in [0, T_{n+1})$, and showing N^{n+1} is a solution on $[0, T_{n+1}]$.

We terminate this procedure if $T_{n+1} = \infty$[6]. In this case, we have found a solution on all of $[0, \infty)$. If $T_n < \infty$ for all n, let $T = \lim_{n \to \infty} T_n$. $N = \sum_{n=1}^\infty \mathbb{1}_{[T_n, \infty)}$ is a solution on $[0, T)$ whether T is finite or infinite.

The last statement of the theorem follows trivially, because $T_n \to T < \infty$ only if the sequence of events has an accumulation point at T. □

In Figure 5, the intensity λ^0 would be the line section which doesn't jump at T_1, but rather continues its path (since it is a function of N^0, which has no jump at T_1). λ^1 on the contrary, agrees first with λ^0 on $[0, T_1)$, but jumps after T_1. The breaking point in the proof is that by left-continuity of λ^0 and λ^1, they agree on all of the interval $[0, T_1]$, and in particular, have the same points under the curve in $[0, T_1]$.

Example 4.15. For an example where $T < \infty$, consider the intensity

$$\lambda(t, \{N_s\}_{0 \leq s < t}) = (1 + \lim_{s \to t-} N_s)^2$$

on $\mathbb{R}_+$. I.e. the intensity in the interval $(T_n, T_{n+1}]$ is $(1+n)^2$, and so $T_{n+1} - T_n \mid \mathcal{F}_n$ is $\exp\left(\frac{1}{(1+n)^2}\right)$-distributed, and marginally T_n is distributed like a sum $\sum_{j=1}^n X_j$, where X_j is exponentially distributed with mean $\frac{1}{j^2}$.

Thus $\mathbb{E}[T_n] = \sum_{j=1}^n \frac{1}{j^2}$, and $\mathbb{E}[\lim_{n \to \infty} T_n] = \lim_{n \to \infty} \mathbb{E}[T_n] < \infty$. In particular $P(\lim_{n \to \infty} T_n < \infty) = 1$. ∘

[6] Which is possible, since it's the infimum of a set that could potentially be empty.

This example is however slightly sought after, and most often, we will consider 'non-exploding' intensities, where $T = \infty$. As shown in (Jacobsen 2006, Example 4.4.5), this is for instance the case if for all t, $\lambda_t \leq a + bN_{t-}$ for some $a, b \geq 0$.

Theorem 4.14 together with the following proposition, show that for any λ satisfying the requirements in Theorem 4.14, there exists a point process N, such that λ is the intensity of N.

Proposition 4.16. *Let λ satisfy the requirements in Theorem 4.14, and let N be the solution to the integral equation. If N is adapted and $\mathbb{E}N_t < \infty$ for all t, then $N_t - \int_0^t \lambda_u du$ is a martingale (with λ_u shorthand for $\lambda(u, \{N_v\}_{0 \leq v < u}, \mathbb{T})$).*

In other words, λ is indeed the intensity of N.

Proof. Let $s \leq t$. We rewrite:

$$\mathbb{E}\left[N_t - \int_0^t \lambda_u \mathrm{d}u \mid \mathcal{F}_s\right] = N_s - \int_0^s \lambda_u \mathrm{d}u + \mathbb{E}\left[N_t - N_s \mid \mathcal{F}_s\right] - \mathbb{E}\left[\int_s^t \lambda_u \mathrm{d}u \mid \mathcal{F}_s\right]$$

We show that the two last two terms cancel, which will establish the martingale property.

With a slight abuse of notation, let $\mu(\mathrm{d}x \times \mathrm{d}u) = \mu(\mathrm{d}x) \otimes \mu(\mathrm{d}u)$ denote the factorization into marginal measures. Now it follows that:

$$\begin{aligned}\mathbb{E}\left[N_t - N_s \mid \mathcal{F}_s\right] &= \mathbb{E}\left[\int_s^t \int_0^{\lambda_u} 1\mu(\mathrm{d}x)\mu(\mathrm{d}u) \mid \mathcal{F}_s\right] \\ &= \int_s^t \int \mathbb{E}[\mathbb{1}_{[0,\lambda_u]}(x) \mid \mathcal{F}_s]\mathrm{d}x\mathrm{d}u\end{aligned}$$

The last equality uses the fact that the Poisson measure $\mu(\mathrm{d}u)$ is independent of $\mathcal{F}_s$ for $u \in (s, t]$, and so $\mathbb{E}[\mu(\mathrm{d}u) \mid \mathcal{F}_s] = \mathbb{E}[\mu(\mathrm{d}u)] = \mathrm{d}u$. Similarly $\mu(\mathrm{d}x)$ is independent of $\mathcal{F}_s$ for $u > s$, and so also this reduces to $\mathrm{d}x$. Also consider:

$$\begin{aligned}\mathbb{E}\left[\int_s^t \lambda_u \mathrm{d}u \mid \mathcal{F}_s\right] &= \int_s^t \mathbb{E}[\lambda_u \mid \mathcal{F}_s]\mathrm{d}u \\ &= \int_s^t \mathbb{E}\left[\int \mathbb{1}_{[0,\lambda_u]}(x)\mathrm{d}x \mid \mathcal{F}_s\right]\mathrm{d}u \\ &= \int_s^t \int \mathbb{E}\left[\mathbb{1}_{[0,\lambda_u]}(x) \mid \mathcal{F}_s\right]\mathrm{d}x\mathrm{d}u\end{aligned}$$

Since the two expressions coincide, it follows that $\mathbb{E}\left[N_t - \int_0^t \lambda_u \mathrm{d}u \mid \mathcal{F}_s\right] = N_s - \int_0^s \lambda_u \mathrm{d}u$. Consequently $N_t - \int_0^t \lambda_u \mathrm{d}u$ is a martingale. □

In conclusion, Theorem 4.14 and Proposition 4.16 show that, under regularity conditions, for a given mechanism $t \mapsto \lambda(t, \{N_u\}_{0<u<t}, \mathbb{T})$, there exists a process N on $(0, \infty)$ such that λ is the intensity of N, and so to specify a process, we simply need to specify λ.

Our construction of N starts in 0, and so past information is included explicitly as $\mathbb{T}$. If there is no ambiguity, we abuse notation and write $t \mapsto \lambda(t, \{N_u\}_{u<t})$.

4.4 Hawkes processes

We now consider a particular type of processes, namely Hawkes processes (Hawkes 1971). Hawkes processes are so-called self-exciting, in the sense that occurrence of an event will increase the intensity of further events. We introduce Hawkes processes as a cluster process:

Immigrant-Child process

Consider a homogeneous Poisson process N^0 with rate λ_0. We call the events of N^0 **immigrants**, since they are introduced to our system without any predecessors.

Further, for any point t_0, immigrant or not, initiate in t_0, a new (independent) inhomogeneous Poisson process C^{t_0} with intensity measure $\mu^{t_0}(A) = \mu(A - t_0)$, and such that $v = \mu(\mathbb{R}_+) \leq 1$. We denote C^{t_0} the **child process** and observe that the child processes are identically distributed apart from the translation to 'time-of-birth' t_0. Again each child of t_0 may spark its own child processes, and so let D^{t_0} denote the descendant process of t_0, that is the sum of all children-processes which originated from t_0.

We let N denote the sum of both the immigrant process and all child processes, and call N a **Hawkes process**. That is:

$$N(A) = \int_{\mathbb{R}} D^t(A) N^0(\mathrm{d}t) + N^0(A) \tag{5}$$

We show that this is well defined by showing that for bounded A, $N(A) < \infty$ a.s. For an event t_0, let $\mathrm{de}(t_0) = D^{t_0}(\mathbb{R})$ be the total number of descendants of t_0.

Proposition 4.17. *If $v < 1$, for any point t_0:*

$$\mathbb{E}[\mathrm{de}(t_0)] = \frac{v}{1-v}$$

In particular, $P(\mathrm{de}(t_0) < \infty) = 1$

Proof. Let g^n denote the number of n'th generation descendants of t_0. I.e. the number of children of t_0 are g^1 and so forth. Since g^1 is Poisson, $\mathbb{E}[g^1] = v$. Also by the tower property:

$$\mathbb{E}[g^2] = \mathbb{E}\left[\mathbb{E}[g^2 \mid g^1]\right] = \mathbb{E}[v g^1] = v^2$$

And similarly, by induction, $\mathbb{E}[g^n] = v^n$. Thus, since $\mathrm{de}(t_0) = \sum_{n=1}^{\infty} g^n$, the result follows. □

This enables us to show stability of the Hawkes process in the following way:

Proposition 4.18. *For any bounded set A, $P(N(A) < \infty) = 1$.*

Proof. By Proposition 4.17, every immigrant has at most finitely many descendants, and so if $N(A) = \infty$, by the pidgeon-hole principle, infinitely many immigrants have descendants in A. This is only the case if $(D^t(A) > 0 \,\text{i.o.})$ which by the Borel-Cantelli lemma has probability 0 if

$$\int P(D^t(A) > 0) N^0(\mathrm{d}t) < \infty$$

And indeed this is the case because:

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathbb{R}} P(D^t(A) > 0) N^0(\mathrm{d}t)\right] &= \int P(D^t(A) > 0)\,\mathrm{d}t \\
&\overset{i)}{\leq} \int_{\mathbb{R}} \mathbb{E}\left[D^t(A)\right]\mathrm{d}t \\
&= \int_{\mathbb{R}} \mu(A - t)\mathrm{d}t \\
&\overset{ii)}{=} \ell(A)\mu(\mathbb{R}) < \infty
\end{aligned}$$

where $i)$ is the Markov inequality and $ii)$ uses Tonellis theorem and the shift invariance of the Lebesgue measure ℓ. □

Hence we see that the Hawkes process will a.s. only assign finitely many points to any bounded set, and in particular Equation (5) defines a well defined point process.

Hawkes with intensities

Suppose that μ has density with respect to the Lebesgue measure, say $\mu = f \cdot \ell$. Then the child process C^{t_0} of a point t_0 with intensity measure $\mu(A - t_0)$ has intensity $f(t - t_0)$ for any $t \in \mathbb{R}$.

The intensity of the Hawkes process is simply the sum of the child-intensities:

$$\lambda_t = \lambda^0 + \sum_{s<t} f(t - s) = \lambda^0 + \int_{-\infty}^{t-} f(t - s) N(\mathrm{d}s)$$

A frequent choice of μ is for it to be an unnormalized exponential distribution, i.e. with $0 < \beta_0 < \beta_1$ one takes $f(t) = \beta_0 \exp(-\beta_1 t)$ for $t > 0$, or to let f be a mixture of such exponential kernels (with proper restrictions on coefficients such that still $\mu(\mathbb{R}_+) < 1$). The latter possibility, gives rise to the intensity:

$$\lambda(t, \{N_s\}_{0 \leq s < t}) = \lambda_0 + \int_{-\infty}^{t-} \sum_i \beta_0^i e^{-\beta_1^i (t-s)} N(\mathrm{d}s)$$

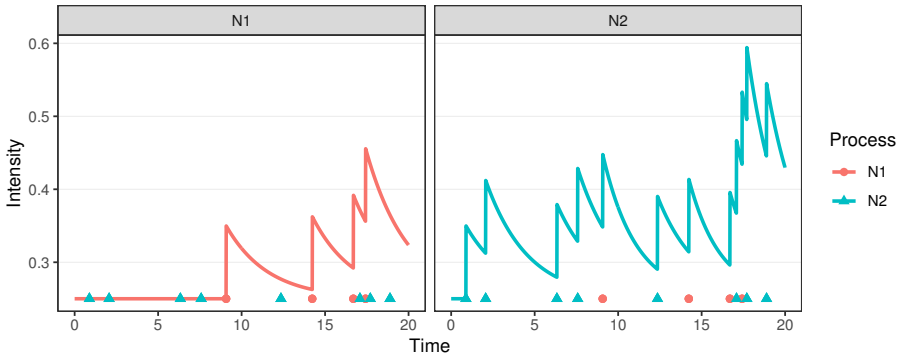


Figure 6: Simulated sample path from the bivariate Hawkes process described in Example 4.21. The left and right plot show intensities for each of the coordinate processes, and both plots show events in N^1 (circles) and N^2 (triangles).

4.5 Marked Point Processes

We now turn to the theory of marked (or multivariate) point processes.

Definition 4.19. Let V be a finite mark space $\{1, \dots, d\}$. A **marked point process** N is a point process on the space $\mathbb{R} \times V$.

The marginal measure $N^g(A) = N(A, V)$ for $A \in \mathcal{B}(\mathbb{R})$ is called the **ground process**, and decomposes due to additivity of measures, such that $N^g(A) = \sum_{v=1}^{d} N(A, \{v\})$. ∘

Some definitions allow for a larger set V with the additional requirement that the marginal measure is boundedly finite (which is e.g. not the case if $V = \mathbb{R}$ and N a homogeneous Poisson process) or $\mathcal{X} \neq \mathbb{R}$, but considering a finite V and $\mathcal{X} = \mathbb{R}$ suffices for our applications.

As such, a marked point process is just a generic point process on this space, but often one interprets a point $x \times v \in \mathbb{R} \times V$ as an event of type v appearing at the position x. The ground process then is the process in $\mathbb{R}$ of occurrences *of any type*, while $N(A, \{v\})$ is count of events of type v occurring within a set $A \in \mathcal{B}(\mathbb{R})$.

We denote the measures $A \mapsto N(A, \{v\})$ the **coordinate processes**, and write them $N^v(A)$. If the coordinate process N^v allows for an intensity, denote this λ^v.

Note that if $t \mapsto N^v(0, t] - \int_0^t \lambda_t^v \mathrm{d}t$ is a $\mathcal{F}_t$-martingale in t for each $v \in V$, then also $t \mapsto N_t^g - \int_0^t \lambda_t^g \mathrm{d}t$ is a martingale with $\lambda^g = \sum_{v \in V} \lambda^v$, and so $\sum_{v \in V} \lambda^v$ is the intensity of the ground process.

We say that a process N has **independent marks**, if given the ground process N^g, the marks are independent, each depending only on its own position. In this case, the distribution of each mark is characterized by the mark kernel, $\gamma(v \mid x)$, which for each $x \in \mathbb{R}$ is a distribution over V.

If a ground process has intensity $t \mapsto f_t^g$, the coordinate process N^v has intensity $t \mapsto \gamma(v \mid t) f_t^g$.

Definition 4.20. We say that N is a **marked Hawkes process** on $\mathbb{R} \times V$, if

- The immigrant process is a marked process with independent marks and a mark kernel $\gamma^0(v)$ which is independent of x, and where the ground process is a homogeneous Poisson process with rate λ_0.
- Given the parent, the child process is a marked process with independent marks, and where the ground process is a inhomogeneous Poisson process, with $v < 1$. Both the intensity measure $\mu_{v_{\text{pa}}}$ and the mark kernel $\gamma(v \mid x, v_{\text{pa}})$ is allowed to vary by the mark of the parent.

and N is the sum of both immigrant and child processes. ∘

Again, if μ_v allows for a ground density f^v for each $v \in V$, the intensity of each coordinate process is:

$$\lambda_t^v = \gamma^0(v)\lambda_0 + \sum_{v' \in V} \int_{-\infty}^{t-} f^{v'}(t-s)\gamma(v \mid t-s, v') N^{v'}(\mathrm{d}s)$$

Example 4.21. As a simple example, consider a marked Hawkes process with $V = \{1, 2\}$. Let the intensity be:

$$\lambda^1(t) = 0.1 + \int_0^{t-} e^{-2(t-s)} N^2(\mathrm{d}s)$$
$$\lambda^2(t) = 0.1 + \int_0^{t-} \frac{1}{2} e^{-(t-s)} N^1(\mathrm{d}s) + \int_0^{t-} e^{-2(t-s)} N^2(\mathrm{d}s)$$

A simulation from this process is seen in Figure 6. One observes that events in N^1 only directly affect the intensity λ^2, whereas events in N^2 affect both λ^1 and λ^2. ∘

4.6 Likelihoods

For a point process, we now introduce the likelihood. As in most of statistics, the likelihood is used for selecting the element of a model class which best fits the data. In particular, we shall parametrize intensities of processes as $\lambda(\beta)$, and maximize the likelihood with respect to β.

Definition 4.22. Let N be a point process on $\mathbb{R}$ which admits the predictable intensity λ. Let $t_1, \ldots, t_n$ be some observed sequence of points in $[0, T]$. Then the **likelihood** is the quantity:

$$\mathcal{L}_T(t_1, \ldots, t_n) = \exp\left(\sum_{j=1}^{d} \log \lambda_{t_j} - \int_0^T \lambda_t \, \mathrm{d}t \right)$$

∘

If it causes no ambiguity, we suppress $t_1, \ldots, t_n$ from the notation, although of course their values are highly important. If we consider a point process with intensity $\lambda(\beta)$ parametrized by $\beta \in \Theta$, we denote the likelihood $\mathcal{L}_T(\beta)$.

We may also be interested in the distribution of the likelihood. If $T_1, \ldots, T_n$ are the events of N in $[0, T]$ (viewed as random variables), the likelihood evaluated in them will be:

$$\begin{aligned}\mathcal{L}_T &= \exp\left(\sum_{j=1}^{n} \log \lambda_{T_j} - \int_0^T \lambda_t \mathrm{d}t\right) \\ &= \exp\left(\int_0^T \log \lambda_t N(\mathrm{d}t) - \int_0^T \lambda_t \mathrm{d}t\right)\end{aligned}$$

(which we also write $\mathcal{L}_T(\beta)$ when parametrizing).

It is curious to consider the difference to the likelihoods typically considered in classical, non-temporal statistics. Here the likelihood often is of a form $\mathcal{L}(\theta) = \prod_i p(x_i \mid \theta)$, similar to the first term of the point process likelihood. The second term $\int_0^T \lambda_t \mathrm{d}t$ however, is not matched by the classical likelihood computations. In effect this is because, although a high intensity in a region explains the occurrence of a point, it also leads to the question: Why didn't any further events occur? This is addressed by the second term in the point process likelihood. While a high intensity benefits in the first term (because it explains an observed event), it is penalized in the second for not explaining *further events that did not happen.*

In the case that we paramatrize λ in a way such that $\lambda(\beta)$ is linear in β (i.e. $\lambda(a\beta_1 + b\beta_2) = a\lambda(\beta_1) + b\lambda(\beta_2)$ for all $a, b \in \mathbb{R}$), the below proposition comes in handy.

Proposition 4.23. *If $\lambda(\beta)$ linear in β, then $\log \mathcal{L}_T(\beta)$ is concave.*

Proof. We consider the two terms of

$$\log \mathcal{L}_T(\beta) = \sum_{j=1}^{n} \log \lambda_{t_j}(\beta) - \int_0^T \lambda_t(\beta)\, \mathrm{d}t$$

Linear functions are (the only functions that are) both convex and concave. Furthermore, composing a concave function with a linear, yields a concave function. Since log is concave, $\log \lambda_{t_j}(\beta)$ is concave for each j, and so is the sum. Since $\lambda_t(\beta)$ is also convex for each t, so is the integral, and thus $-\int_0^T \lambda_t(\beta)\, \mathrm{d}t$ is concave. Adding the two concave summands again yield a concave function. See (Boyd and Vandenberghe 2004) for further details. □

5 Intensity Expansion

In this section we shall develop a representation for intensities of a point process. Throughout we shall consider a multivariate point process $N = (N^1, \dots, N^d)$ with corresponding intensities $\lambda^1, \dots, \lambda^d$, all assumed predictable with respect to the filtration induced by N.

In the classic theory of Volterra series (Volterra 1887; Volterra 1959), in a dynamic time-homogeneous system, one fits the relationship between an input process x_t and output process y_t by the expansion:

$$y_t = h^0 + \sum_{n=1}^{N} \int_{-\infty}^{t} \cdots \int_{-\infty}^{t} h^n(s_1, \dots, s_n) x_{t-s_1} \cdots x_{t-s_n} \mathrm{d}s_1 \cdots \mathrm{d}s_n$$

Here the functions h^n are the so called *Volterra kernels*, which contain the dynamics of the system. Under sufficient regularity conditions (Volterra 1959; Ahmed 1970), y_t can be approximated arbitrarily well when N tends to infinity.

Can this result be translated to the world of point processes, with the coordinate processes as our input and the intensity as the output? There are several differences: The point processes are highly non-continuous and while the above system had a fixed input structure over time, in a point process, the number of influencing events is stochastic in any region.

Yet, in this section, we deliver the proof that a similar representation is possible. The methods we use are completely different to those of the Volterra series, though the result looks surprisingly much the same.

Similarly to (Cohen 2012), our proof is based on martingale convergence, but with the important difference, that we don't restrict ourselves to bounded memories of the kernels and start the process at $-\infty$.

The main result of the section is Theorem 5.1, which we state first, and dedicate Sections 5.1 to 5.4 to proving. Section 5.1 establishes a density of function spaces, in Section 5.2 we prove the core mechanism on the interval $(-\infty, 0]$, and in Section 5.3 we extend to intervals $(-\infty, T]$, all without marks. Section 5.4 shows that the proof can easily be extended to the case of marked point processes.

We note that our interest is primarily in the convergence of such representation, not the concrete functional forms of the kernels. In a real application, one would make various approximations anyways (cf. Section 7.1), and so there's no guarantee that we recover the concrete kernels. But from a theoretical point of view it is valuable to know, that the representation converges.

Like the Volterra theory, our result assumes time homogenity of λ_t, which roughly means that only the relative distance to events affect λ_t (and not the concrete value of t). This is formally stated in Assumption 5.11.

For a simple notation, let $\alpha = (\alpha_1, \dots, \alpha_n)$ denote a generic tuple of length n of the coordinate indices, e.g. $(1, 1, 2)$ is a 3-tuple referring twice to coordinate N^1 and once to N^3. We write the sum over all 3-tuples as $\sum_{|\alpha|=3}$.

Theorem 5.1. *Let λ be the intensity of one of the coordinates of N. Assume that $\lambda_t \in \mathcal{L}^1(\mathcal{F}_t)$ for all $t \in [0,T]$. Further suppose that Assumption 5.11 holds for λ. There exist a sequence of functions h^α_τ, such that for every $t \in [0,T]$, the representation:*

$$h^0_\tau + \sum_{n=1}^{\infty} \sum_{|\alpha|=n} \int_{-\infty}^{t} \cdots \int_{-\infty}^{t} h^\alpha_\tau(t-s_1, \cdots t-s_n) N^{\alpha_1}(ds_1) \cdots N^{\alpha_n}(ds_n)$$

approximates λ_t arbitrarily well in $\mathcal{L}^1$ when $\tau \to -\infty$.

Throughout we will assume that N is a simple, non-exploding point process on $\mathbb{R}$. For simplicity we will often use the notation:

$$\int_A f \, \mathrm{d}N^\alpha(t^n) := \int_A \cdots \int_A f(t_1, \ldots, t_n) N^{\alpha_1}(\mathrm{d}t_1) \cdots N^{\alpha_n}(\mathrm{d}t_n)$$

5.1 Density in $\mathcal{L}^1(\mathcal{F})$

Let $T_1, T_2, \ldots$ be the jumps of N starting at 0 and moving backwards in time, i.e. $\ldots < T_2 < T_1 \leq 0$.

Definition 5.2. For $\tau \leq 0$, let $\mathcal{F}_\tau$ be the σ-algebra generated by events in $[\tau, 0]$. One can write this as:

$$\mathcal{F}_\tau = \sigma(T_1 \vee \tau, T_2 \vee \tau, \ldots)$$

Also let $\mathcal{F} = \sigma(\cup_{\tau \leq 0} \mathcal{F}_\tau)$. As the following proposition shows, we do in this way generate all the information about process N. ∘

Proposition 5.3. *The σ-algebra $\mathcal{F}$ equals $\mathcal{F}^N_0 := \sigma(T_1, T_2, \ldots)$.*

Proof. For any i, $T_i \vee \tau$ is $\sigma(T_i)$-measurable, and in particular $\mathcal{F}^N_0$-measurable. Hence $\mathcal{F}_\tau = \sigma(T_1 \vee \tau, \ldots) \subseteq \mathcal{F}^N_0$, and so $\mathcal{F} = \sigma(\cup_\tau \mathcal{F}_\tau) \subseteq \mathcal{F}^N_0$.

Reversely, we show that for each n, T_n is $\mathcal{F}$-measurable. Since $\mathcal{F}^N_0$ is the smallest σ-algebra making all T_n's measurable, $\mathcal{F}^N_0 \subseteq \mathcal{F}$ will follow.

For any n, $(T_n \vee \tau) \longrightarrow T_n$ for $\tau \to -\infty$ (potentially with $T_n = -\infty$). Since $(T_n \vee \tau)$ is $\mathcal{F}$-measurable for each τ (it is a generator), T_n is $\mathcal{F}$-measurable. □

We shall now consider denseness of the subspaces of measurable and integrable functions with respect to $\mathcal{F}_\tau$. The following definition is from (Schilling 2017, Chapter 17):

Definition 5.4. A set $\mathcal{D} \subseteq \mathcal{L}^p, p \in [1, \infty]$ is **dense** if for every $\lambda \in \mathcal{L}^p$ there exists a sequence $(\lambda_n)_{n \geq 0} \in \mathcal{D}$ such that $\lim_{n \to \infty} \|\lambda - \lambda_n\|_p = 0$. ∘

It turns out that the functions that are $\mathcal{L}^1$-integrable and measurable wrt. $\mathcal{F}_\tau$ for some τ are dense in $\mathcal{L}^1(\mathcal{F})$.

Theorem 5.5. *The union of function spaces $\cup_{\tau \leq 0} \mathcal{L}^1(\mathcal{F}_\tau)$ is dense in $\mathcal{L}^1(\mathcal{F})$.*

Proof. Take any $\lambda \in \mathcal{L}^1(\mathcal{F})$. By the tower property one has that $\lambda_\tau := \mathbb{E}[\lambda \mid \mathcal{F}_\tau] \in \mathcal{L}^1(\mathcal{F}_\tau)$ and further it follows from (Sokol and Rønn-Nielsen 2013, Section 5.4) that $(\lambda_\tau)_{\tau \leq 0}$ is a martingale (in $-\tau$) and that $\mathbb{E}[\lambda \mid \mathcal{F}_\tau]$ converges in $\mathcal{L}^1$ to $\mathbb{E}[\lambda \mid \mathcal{F}] = \lambda$ as $\tau \to -\infty$.

Because each $\lambda_\tau \in \mathcal{L}^1(\mathcal{F}_\tau) \subseteq \cup_\tau \mathcal{L}^1(\mathcal{F}_\tau)$, it follows that $\cup_\tau \mathcal{L}^1(\mathcal{F}_\tau)$ is dense in $\mathcal{L}^1(\mathcal{F})$. □

5.2 The integral representation of λ

In this section, for some $\lambda_0 \in \mathcal{L}^1(\mathcal{F}_\tau)$, we show for each $M \in \mathbb{N}$ that $\lambda_0 \mathbb{1}_{N([\tau,0])=M}$ can be written as a sum of integrals of deterministic functions. These integrands will play the role of the Volterra kernels, but only given the count $N([\tau, 0])$. We then sum over these terms, to obtain a general representation of λ_0.

The following classic result, here as presented in (Sokol and Rønn-Nielsen 2013), motivates expansion of the λ's to functional approximations:

Lemma 5.6. *A real valued random variable λ is $\sigma(Y)$ measurable if and only if there exist a measurable map $\phi : (E; \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that:*

$$\lambda = \phi(Y)$$

In our case this means that if $\lambda_0 \in \mathcal{L}^1(\mathcal{F}_\tau)$, there exists a measurable map f such that:

$$\lambda_0 = f(T_1 \vee \tau, T_2 \vee \tau, \ldots) \tag{6}$$

To obtain an integral representation of λ_0, we can utilize this function. Define $f^n_\tau(t_1, \ldots, t_n) = f(t_1, \ldots, t_n, \tau, \tau, \ldots)$, which we'll denote f^n if τ is clear from the context[7], or $f(t_1, \ldots, t_n)$.

Suppose for a moment that exactly one event occurred in the interval $A := [\tau, 0]$, i.e. $T_1 \in A, T_n \notin A$ for $n \geq 2$. Then one could write:

$$\begin{aligned}\lambda_0 &= f(T_1 \vee \tau, T_2 \vee \tau, \ldots) = f(T_1, \tau, \tau, \ldots) \\ &= \int_\tau^0 f(t, \tau, \tau, \ldots) N(\mathrm{d}t) = \int_\tau^0 f^1 N(\mathrm{d}t)\end{aligned}$$

This however depends heavily on the assumption that $N(A) = 1$. If instead the interval contained say m events, then $\int_\tau^0 f^1(t) N(\mathrm{d}t) = f^1(T_1) + \ldots + f^1(T_m)$ which is not equal to λ_0 (because in this case $\lambda_0 = f^m(T_1, \ldots, T_m)$).

Is there a way to write λ_0 in this integral representation? The following proposition devices a procedure, such that one can obtain $f(T_1)$ exactly if $N(A) = 1$ and else 0, using only integrals of *deterministic* functions (if one could use non-deterministic functions, the problem was easily solved, by including an $\mathbb{1}_{N(A)=1}$ term).

[7] It is only in Theorem 5.10 and after, that we will need to consider f_τ for various τ.

Proposition 5.7. *Assume N is a simple, non-exploding point process. For $\lambda_0 \in \mathcal{L}^1(\mathcal{F}_\tau)$ one can compute $\lambda_0 \mathbb{1}_{N(A)=1}$ (with $A = [\tau, 0]$) by integrals of deterministic integrands as follows:*

$$\lambda_0 \mathbb{1}_{N(A)=1} = \sum_{n=1}^{\infty} \beta_n \int_A f(t_1) \mathbb{1}_{D_n} \, dN(t^n) \tag{7}$$

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

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

Proof. The proof runs constructively in the following way: First we hope that really $N(A) = 1$, and so $\lambda_0 = \int_A f^1 \mathbb{1}_{D_1} \, \mathrm{d}N(t^1)$. If this is not the case, we need to cancel all the terms generated by $\int_A f^1 \mathbb{1}_{D_1} \, \mathrm{d}N(t^1)$, which is sequentially done by integrals of higher and higher order, in each round n choosing a coefficient β_n such that if actually n events occurred in the interval, then all previous terms cancel.

As a primer, consider the term:

$$\int_A f(t_1) \mathbb{1}_{D_1} \, \mathrm{d}N(t^1) - \int_A f(t_1) \mathbb{1}_{D_2} \, \mathrm{d}N(t^2) + \frac{1}{2} \int_A f(t_1) \mathbb{1}_{D_3} \, \mathrm{d}N(t^3)$$

For an $\omega \in (N(A) = 1)$ the first integral equals $f(T_1)$, the latter two are 0 (since there's only one distinct jump so $\mathbb{1}_{D_2}(t_1, t_2) = 0$). Denote this the $n = 1$ case.

If $\omega \in (N(A) = 2)$ ('$n = 2$' case) the first integral yields $f(T_1) + f(T_2)$, but the second yields also $f(T_1) + f(T_2)$, and so by choosing $\beta_2 = -1$, the whole thing becomes 0, which is the desired since also $\mathbb{1}_{N(A)=1} \lambda_0 = 0$ for such ω. (Again $\mathbb{1}_{D_3} = 0$ if there's only 2 jumps).

If instead $\omega \in (N(A) = 3)$, the compensation in the $n = 2$ case overcompensated, since the first term would be $f(T_1) + f(T_2) + f(T_3)$ but the second term would then be $-2\,[f(T_1) + f(T_2) + f(T_3)]$[8]. Now however, the third term will cancel out both the overcompensation of the first and the second term, when choosing a coefficient $\beta_3 = \frac{1}{2}$[9].

And so we continue. In each round n, we assume that n is the true number of events in A, and given this, we calculate the overcompensations of previous rounds. We then choose the coefficient β_n that would cancel out all previous terms, if n was indeed the true number of events. And in that case, there are no distinct pairs of size $n + 1$, and so round n will terminate the scheme. If n is not the true count in A, round $n + 1$ continues the scheme.

Since N is non-exploding, $N(A) < \infty$ a.s. and so this compensation-scheme will almost surely terminate in a finite number of steps. I.e. $\sum_{n=1}^{M} \beta_n \int_A f(t_1) \mathbb{1}_{D_n} \, \mathrm{d}N(t^n)$

[8] Because T_1 is the first element of both the pairs (T_1, T_2) and (T_1, T_3), and similar for T_2 and T_3.

[9] Because T_1 is the first element of both the pairs (T_1, T_2, T_3) and (T_1, T_3, T_2), and similar for T_2 and T_3.

converges almost surely for $M \to \infty$, since the increments will be 0 eventually, justifying the representation in Equation (7).

We can also compute the coefficients β_n. $\beta_1 = 1$ is required, to have the desired if $N(A) = 1$.

The number of $f(T_1)$ terms generated in round n if the true number of events in A is $m \geq n$ is

$$g_n(m) = \frac{(m-1)!}{(m-n)!}$$

Note that we only need to count $f(T_1)$ terms, because by symmetry one generates the same number of $f(T_2), f(T_3), \ldots$ events, and the compensation similarly is always symmetric. Now in round n, believing that n is the true number of event, the number of $f(T_1)$ terms cumulated until now is:

$$c_n = \sum_{j=1}^{n-1} \beta_j g_j(n) = \sum_{j=1}^{n-1} \beta_j \frac{(m-1)!}{(m-j)!}$$

Thus to compensate we choose $\beta_n = -\frac{c_n}{g_n(n)}$, because if n was indeed the true count, the total number of terms after round n would be

$$c_n - \frac{c_n}{g_n(n)} \cdot g_n(n) = 0$$

Hence the coefficients β_n follow the recursion:

$$\beta_n = -\frac{c_n}{g_n(n)} = -\frac{\sum_{j=1}^{n-1} \beta_j \frac{(n-1)!}{(n-j)!}}{\frac{(n-1)!}{(n-n)!}} = -\sum_{j=1}^{n-1} \frac{\beta_j}{(n-j)!} \tag{8}$$

We show that $\beta_n = \frac{(-1)^{n-1}}{(n-1)!}$ is a solution to this equation.

First off, it matches the initial condition $\beta_1 = 1$. Moving everything to the left hand side, Equation (8) reads:

$$\sum_{j=1}^{n} \frac{\beta_j}{(n-j)!} = 0$$

Inserting the proposed solution, one obtains:

$$\begin{aligned}
\sum_{j=1}^{n} \frac{(-1)^{j-1}}{(n-j)!(j-1)!} &= \frac{1}{(n-1)!} \sum_{j=0}^{n-1} (-1)^j \binom{n-1}{j} \\
&= \frac{1}{(n-j)!} \Big(1 + (-1)\Big)^{n-1} \\
&= 0
\end{aligned}$$

And so $\beta_n = \frac{(-1)^{n-1}}{(n-1)!}$ indeed solves the recursions. □

This easily extends to the following corollary:

Corollary 5.8. *Let $\lambda_0 \in \mathcal{L}^1(\mathcal{F}_\tau)$. For $M \in \mathbb{N}$, one has:*

$$\lambda_0 \mathbb{1}_{N(A)=M} = \sum_{n=M}^{\infty} \beta_n^M \int_A f(t_1, \ldots, t_M) \mathbb{1}_{D_n} \mathbb{1}_{O_M} \, dN(t^n)$$

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

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

Proof. The case $M = 1$ is covered in Proposition 5.7. For $M \geq 2$, the compensation scheme essentially is the same, with the additional requirement that the first M jumps should be ordered. To see this, consider the case of $M = 2$ and $\omega \in (N(A) = 2)$. Then $\int_A f^2(t_1, t_2) \mathbb{1}_{D_2} \, \mathrm{d}N(t^2)$ would produce $f(T_1, T_2) + f(T_2, T_1)$ which is undesired, whereas when including $\mathbb{1}_{O_2}$, $f(T_1, T_2)$ is returned as desired.

The 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, namely $g_n^M(m) = \frac{(m-M)!}{(m-n)!}$ and so the scheme runs in a completely similar fashion. □

Extending further on Proposition 5.7 and Corollary 5.8, we may include the base-rate $\lambda_0 \mathbb{1}_{N(A)=0}$. Let h^0 be the value of λ_0 on this set (that is $h^0 = f(\tau, \tau, \ldots)$). Now $\sum_{n=1}^{\infty} \int_A \left[f(t_1) - h^0 \right] \mathbb{1}_{D_n} \, \mathrm{d}N(t^n)$ will return the *additional to base-rate* intensity $f(T_1) - h^0$ if $N(A) = 1$ and 0 else.

Thus, as the following theorem shows, we can in general decompose $\lambda_0 \in \mathcal{L}^1(\mathcal{F}_\tau)$.

Proposition 5.9. *Assume N is a non-exploding point process. For $\lambda_0 \in \mathcal{L}^1(\mathcal{F}_\tau)$ one can compute λ_0 from deterministic integrands h_τ^n and a deterministic constant h_τ^0 as:*

$$\lambda_0 = h_\tau^0 + \sum_{n=1}^{\infty} \int_{(-\infty, 0]} h_\tau^n(t_1, \ldots, t_n) \, dN(t^n) \quad a.s.$$

with

$$h_\tau^n(t_1, \ldots, t_n) = \mathbb{1}_{[\tau, 0]} \mathbb{1}_{D_n} \sum_{M=1}^{n} \beta_n^M \left[f_\tau(t_1, \ldots, t_M) - h_\tau^0 \right] \mathbb{1}_{O_M}$$

Proof. Still we let $A = [\tau, 0]$. Now the result follows directly be writing $\lambda_0 = \mathbb{1}_{N(A)=0} \lambda_0 + \sum_{M=1}^{\infty} \lambda_0 \mathbb{1}_{N(A)=M}$, and write each term $\lambda_0 \mathbb{1}_{N(A)=M}$ as given by Corollary 5.8. By reordering terms, one obtains the desired representation. Note that there are no problems of switching orders of summation, since for each $\omega \in (N(A) < \infty)$, only finitely many of the summands are non-zero. □

We summarize the discussions above, by noting that any $\mathcal{L}^1(\mathcal{F})$ variable can be arbitrarily well approximated using the additive representations from above.

Theorem 5.10. *Let $\lambda_0 \in \mathcal{L}^1(\mathcal{F})$. λ_0 can be arbitrarily well approximated by representations of the form*

$$h_\tau^0 + \sum_{n=1}^{\infty} \int_{(-\infty,0]} h_\tau^n(t_1, \ldots, t_n) N(\mathrm{d}t^n)$$

where the functions $h_\tau^n, n \in \mathbb{N}$ are supported on $[-\tau, 0]$.

Proof. It follows by Theorem 5.5 that there exists a sequence $(\lambda_\tau)_{\tau \in \mathbb{N}}$ such that $\lambda_\tau \in \mathcal{L}^1(\mathcal{F}_\tau)$ and $\lambda_\tau \xrightarrow{\mathcal{L}^1} \lambda_0$. From Proposition 5.9 it follows that each λ_τ has the desired representation, and by construction it also follows that h_τ^n is supported on $[-\tau, 0]$. □

5.3 Representations on $(-\infty, t]$

So far, the representation has been with respect to σ-algebras regarding events on $(-\infty, 0]$. We now extend these to all of $\mathbb{R}$ by shifting events back to 0.

Let N^π denote the shifted point process $N^\pi(A) = N(A + \pi)$. With a substantial abuse of notation, one could also define the mapping $\pi(x) = x - \pi$. It then follows that $\pi(N) = N^\pi$.

Let λ_t^N denote the intensity of our original process at time t. Also let $\lambda_t^{N^\pi}$ be the intensity at time t in the shifted process N^π

One needs to exercise care in interpreting this understanding. If one understands the intensity of a point process as something solely derived after observing the entire process, then of course the intensity would simply shift as one shifted the process in time.

Rather however, if the process is the one *defined* by the intensity, as is the case in Theorem 4.14, the intensity in the shifted process may be different to what one would have in the unshifted process, and so shifting would yield an entirely different intensity, hence process.

One could for instance imagine that λ had a 'hard' dependency on t, such that λ_t would react differently to past events depending on the numerical value of t. In the notation $\lambda(\underline{t}, \{N_s\}_{s<t})$ the essential question is whether the underlined t truly enters the function. It is exactly this types of case the following assumption seeks to rule out.

Assumption 5.11 (Time-homogenity)**.** *Assume that $\lambda_\pi^N = \lambda_0^{N^\pi}$ for all $\pi \in \mathbb{R}$.*

In the language of Theorem 4.14, the intensity mechanism is independent of the time coordinate, i.e.

$$\lambda(\pi, \{N_s\}_{s<\pi}) = \lambda(0, \{N_s^\pi\}_{s<0})$$

From this, the representations above are extended in time under Assumption 5.11:

Theorem 5.12. *Let $\lambda_t \in \mathcal{L}^1(\mathcal{F}_t)$ for all $t \in [0, T]$, and suppose Assumption 5.11 holds.*

Letting h^n_τ be the representations from Theorem 5.10 above, each λ_t can be arbitrarily well approximated by the representations:

$$h^0_\tau + \sum_{n=0}^{\infty} \int_{-\infty}^{t} h^n_\tau(t - s_1, \ldots, t - s_n) N(\mathrm{d}s^n)$$

Proof. The proof runs by shifting λ_t back to 0 using homogenity, and there using the representation results shown above.

To simplify the computation below, observe that for any f, by a change of variable:

$$\int_a^b f(-x) N^\pi(\mathrm{d}x) = \int_a^b f(-x)\pi(N)(\mathrm{d}x) = \int_{a+\pi}^{b+\pi} f(-(x-\pi)) N(\mathrm{d}x)$$

By Assumption 5.11, $\lambda_t = \lambda(t, \{N_s\}_{s<t}) = \lambda(0, \{N^t_s\}_{s<0})$ for each t. $\lambda_0^{N^t}$ is integrable because λ_t^N is, and it is measurable with respect to $\mathcal{F}_0^{N^t}$, the filtration generated by the shifted processes N^t at 0.

Each $\lambda(0, \{N^t_s\}_{s<0}) \in \mathcal{L}^1(\mathcal{F}_0^{N^t})$, and so by the above, each can be arbitrarily well approximated. The essential point is that the functions h^n are common to all $t \in [0, T]$. Remember that the functions h are composed by the functions f, which described how measurable objects depended on the past jumps $T_1, T_2, \ldots$. Suppose $\lambda(0, \{N^{t_1}_s\}_{s<0}), \lambda(0, \{N^{t_2}_s\}_{s<0})$ are the intensities at two different times t_1, t_2 pulled back to 0. If their dependence functions f^{t_1} and f^{t_2} were different, this would clearly violate the time homogeneity.

It now follows by Theorem 5.10 that one can approximate $\lambda_t = \lambda(0, \{N^t_s\}_{s<0})$ arbitrarily well from functions

$$\begin{aligned}
& h^0 + \sum_{n=1}^{\infty} \int_{(-\infty,0]} h^n(-s_1, \ldots, -s_n) N^t(\mathrm{d}s^n) \\
= \; & h^0 + \sum_{n=1}^{\infty} \int_{(-\infty,t]} h^n(-(s_1 - t), \ldots, -(s_n - t)) N(\mathrm{d}s) \\
= \; & h^0 + \sum_{n=1}^{\infty} \int_{(-\infty,t]} h^n(t - s_1, \ldots, t - s_n) N(\mathrm{d}s)
\end{aligned}$$

The negative arguments in h, we can add simply because h is a generic function, and so an extra negative sign could be hidden inside the h.

□

5.4 Extension to Marked Point Processes

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 $A \in \mathcal{B}(X)$ that:

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

When integrating, this factorizes:

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

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

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

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.

Proof of Theorem 5.1. Essentially we just need to show that Theorem 5.12 extends to the marked setting.

In the marked setting, the generated σ-field becomes $\mathcal{F}_\tau = \sigma((T_1\vee\tau, v_1\mathbb{1}_{T_1>\tau}),\ldots)$. A multivariate version of Proposition 5.3 follows because $(T_1\vee\tau, v_1\mathbb{1}_{T_1>\tau}) \to (T_1, v_1\mathbb{1}_{T_1>-\infty})$[10], and so denseness of $\cup_\tau\mathcal{L}^1(\mathcal{F}_\tau)$ also follows in the marked case. Thus the function in Equation (6) could have been written:

$$\lambda_0 = f\left((T_1\vee\tau, v_1\mathbb{1}_{T_1>\tau}),\ldots\right)$$

In the technical Proposition 5.7, 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. If we still compensate based on the event time (that is, use the same construction with D_n and O_M irregardless of marks), the compensation will work in the same way, irregardless of whether or not marks are present.

Therefore also Proposition 5.9 can be generalized such that for $\lambda_0\in\mathcal{L}^1(\mathcal{F}_\tau)$, it holds that:

$$\begin{aligned}\lambda_0 &= h_\tau^0 + \sum_{n=1}^\infty\int_{(-\infty,0]\times V} h_\tau^n(t_1,v_1,\ldots,t_n,v_n)N(\mathrm{d}t^n\times\mathrm{d}v^n)\\ &= h_\tau^0 + \sum_{n=1}^\infty\sum_{|\alpha|=n}\int_{(-\infty,0]} h_\tau^\alpha(t_1,\ldots,t_n)\,\mathrm{d}N^\alpha(t^n)\quad\text{a.s.}\end{aligned}$$

Now by the density of $\cup_\tau\mathcal{L}^1(\mathcal{F}_\tau)$, the result follows for $\lambda_0\in\mathcal{L}^1(\mathcal{F})$, and finally the proof of Theorem 5.12 show that the results on arbitrary intervals $(-\infty,t]$ generalize to the marked case. □

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

6 Local Independence

In this section, local conditional independence for point processes is defined, and a few examples are presented. This will provide a concrete instantiation of the abstract independence models considered in Section 3.

Let $V = \{1, \ldots, d\}$ be a mark space for a point process $N = (N^1, \ldots, N^d)$ with intensities $\lambda^1, \ldots, \lambda^d$.

For a subset C of V, denote by $\mathcal{F}_t^C$ the sub σ-field of $\mathcal{F}_t$ which is generated by processes $\{N^c, c \in C\}$. For each $b \in \{1, \ldots, d\}$, one can define the $\mathcal{F}_t^C$-intensity of b, $\lambda_t^{b,C}$, to be the conditional expectation $\mathbb{E}[\lambda^b \mid \mathcal{F}_t^C]$.

More technically, regularity of this process is guaranteed by taking the optional projection version of $\mathbb{E}[\lambda^b \mid \mathcal{F}_t^C]$ (which will be càdlàg if $\mathcal{F}_t^C$ is) and then taking $\lambda_t^{b,C}$ to be the left limit $\lim_{\epsilon \downarrow 0} \mathbb{E}[\lambda^b \mid \mathcal{F}_{t-\epsilon}^C]$. See (Mogensen, Malinsky, et al. 2018) and references therein.

Following (Mogensen, Malinsky, et al. 2018), we define local independence as follows:

Definition 6.1 (Local independence)**.** Let V be the mark space of the point process, and $A, B, C \subseteq V$. We say that B **is locally independent of** A **given** C if for all $b \in B$, the $\mathcal{F}_t^{A \cup C}$-intensities $\lambda_t^{b,A \cup C}$ have $\mathcal{F}_t^C$-adapted versions. If B is locally independent of A given C, we write

$$A \not\to B \mid C$$

and if this is not the case, $A \to B \mid C$. ∘

Local independence defines an independence model $\mathcal{I}$ where:

$$\langle A, B \mid C \rangle \in \mathcal{I} \iff A \not\to B \mid C$$

Roughly, $A \not\to B \mid C$ if, when observing collections of subprocesses N^A and N^C, our estimate of the intensity of B at any time t doesn't depend on any information from the subprocesses $N^{A \setminus C}$ up to time t.

Remark. Note that a process is allowed to be locally independent of itself, i.e. $a \not\to a \mid C$. For instance this is the case for a unit rate homogeneous Poisson process N^a, in which the intensity is constantly 1, and so $\lambda_t^{a,a} = \mathbb{E}[1 \mid \mathcal{F}_t^a] = 1$ is $\mathcal{F}^{\emptyset}$-measurable, thus $a \not\to a \mid \emptyset$. ∘

Remark. Readers familiar with (usual) conditional independence will recognize many aspect in local conditional independence. There are however some differences. For one, local conditional independence is not (necessarily) symmetric, as is the case with conditional independence. This is evident from the following example. ∘

Example 6.2. Let N be a 3-dimensional counting process on $\mathbb{R}_+$ with coordinate processes N^a, N^b, N^c, with intensities:

$$\lambda_t^a = \mathbb{1}_{N_t^a = N_t^c} \qquad \lambda_t^b = \mathbb{1}_{N_t^b < N_t^a} \qquad \lambda_t^c = \mathbb{1}_{N_t^c < N_t^b}$$

and initial conditions $N_0^a = N_0^b = N_0^c = 0$. Think of this as 3 children throwing a ball between them (in the fixed order $a \to b \to c \to a \to \ldots$), each taking an exponentially distributed amount of time to pick up the ball, before passing it on (emitting a point in process a when child a throws the ball).

Intuitively it is clear, that if a knows the last throw of c, any throw of b is irrelevant for the throw-intensity of a. To b however, the last throw of a is highly important for the intensity of b. Thus we expect that $b \not\to a \mid a, c$ and $a \to b \mid b, c$. This would show that local independence is asymmetric in a and b.

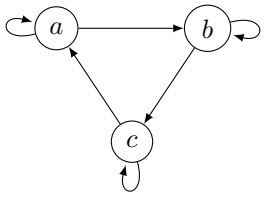
And indeed, this is the case since:

$$\lambda_t^{a,\{a,b,c\}} = \mathbb{E}\left[\lambda_t^a \mid \mathcal{F}_t^{\{a,b,c\}}\right] = \mathbb{1}_{N_t^a = N_t^c} \in \mathcal{F}_t^{a \cup c}$$
$$\lambda_t^{b,\{a,b,c\}} = \mathbb{E}\left[\lambda_t^b \mid \mathcal{F}_t^{\{a,b,c\}}\right] = \mathbb{1}_{N_t^b < N_t^a}$$

The latter intensity cannot have a $\mathcal{F}_t^{b \cup c}$-measurable version, because it depends on the exponentially distributed throwing time of a, which is independent of $\mathcal{F}_t^{b \cup c}$. ∘

Example 6.3. Consider again the process in the ball-throwing example, Example 6.2. We find the local independence graph and check whether the global Markov property holds with respect to this.

It was verified that $b \not\to a \mid a, c$ and $a \to b \mid b, c$. Also one would find that $a \to a \mid b, c$[11], which, by symmetry, shows that all the local independences $\langle v', v \mid V \backslash v' \rangle$ are $b \not\to a \mid a, c$; $c \not\to b \mid a, b$ and $a \not\to c \mid b, c$. Thus the graph below is a local independence graph of N, since these 3 local independence statements correspond to the 3 edges not present in the graph.



We may also check that $b \perp_\mu a \mid a, c$ in the graph. Let $p = (b, v_2, \ldots, v_{n-1}, a)$ be any walk, for a contradiction assumed to be μ-connecting given a and c. We consider the different possibilities of v_{n-1}.

- If $v_{n-1} = a$, the walk will end $a \to a$ (to obtain a final head at a). But then $v_{n-1} = a$ will be a non-collider, but since it intersects with the conditioning set $\{a, c\}$, the walk is non-connecting.
- If $v_{n-1} = b$, the walk does not comply with the 'final-head' requirement.

[11] Essentially, if child a doesn't observe itself, it cannot distinguish between the situations 1) ball is at a and 2) ball is at b. This is however very relevant for the intensity $\lambda_t^{a,\{a,b,c\}}$

- If $v_{n-1} = c$, c will be a non-collider, and again, since it is also in the conditioning set, the walk cannot be connecting.

In summary, no walk from b to a can be μ-connecting given a, c, and thus $b \perp_\mu a \mid a, c$. By symmetry it is clear that also $c \perp_\mu b \mid a, b$ and $a \perp_\mu c \mid b, c$. As a matter of fact, these are the only non-trivial (i.e. where neither of the two separated sets are empty) μ-separations. Since all of these 3 μ-separation statements are matched by a local independence statement, the global Markov property holds with respect to this graph. ∘

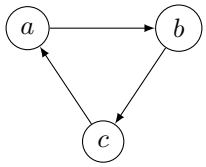
That the local independences for N satisfies the global Markov property with respect to it's local independence graph is not too surprising, since as mentioned in Section 3.3, (Didelez 2008; Mogensen and Hansen 2018) showed that under regularity conditions, local independence models always satisfy the global Markov property with respect to local independence graphs.

Example 6.4. While many point processes will have local independence graphs with self-loops, this doesn't need to be the case. Consider the 3-dimensional Hawkes process $N = (N^a, N^b, N^c)$ on $\mathbb{R}$ with intensities:

$$\lambda_t^a = \lambda_0 + \int_{-\infty}^{t-} \exp(-2s) N^c(\mathrm{d}s)$$
$$\lambda_t^b = \lambda_0 + \int_{-\infty}^{t-} \exp(-2s) N^a(\mathrm{d}s)$$
$$\lambda_t^c = \lambda_0 + \int_{-\infty}^{t-} \exp(-2s) N^b(\mathrm{d}s)$$

To create the local independence graph for this process, we find the local independences of a. Since λ_t^a is $\mathcal{F}_t^c$-measurable, $b \not\to a \mid V \backslash \{b\}$ and $a \not\to a \mid V \backslash a$. On the other hand however, $c \to a \mid V \backslash \{c\}$, because $\mathbb{E}[\lambda_t^a \mid \mathcal{F}_t^V] = \lambda_t^a = \int_{-\infty}^{t-} \exp(-\beta_a s) N^c(\mathrm{d}s)$ which is clearly doesn't have a $\mathcal{F}^{a \cup b}$-measurable version (even though the *intensity* of N_t^c is $\mathcal{F}_t^b$-measurable).

This characterizes all the pairwise local independence statements about a, and by symmetry, also those of b and c. Hence the local independence graph is:



Contrary to Example 6.3, this graph doesn't contain any self-loops. Note however that still a is not independent of its past. One would find that $a \to a \mid \emptyset$, essentially because a high activity in a would stipulate a high activity in b which again would stipulate a high activity in c, which eventually affects a again. ∘

7 A Local Empirical Independence Test

In this section we propose a test for local independence based on the results of Section 5. In Section 7.1 we discuss using tensor basis expansions to fit high-order kernels. Section 7.2 regards estimating the resulting model by constrained, penalized maximum likelihood. In Section 7.3 we derive an approximate normality of maximum likelihood estimates which finally in Section 7.4 we propose to use for the test of local independence.

7.1 Approximating the intensity

We consider the problem of fitting the intensity $\lambda_t^{b,D} = \mathbb{E}[\lambda_t^b \mid \mathcal{F}_t^D]$ for some mark b and a subset D of the process coordinates. We allow for both the cases $b \in D$ and $b \notin D$. The result from Theorem 5.1, which is shown for $\lambda \in \mathcal{L}^1(\mathcal{F}_t)$ (i.e. the history of all coordinates), is easily extended to subsets $D \subseteq V$. That is, $\lambda_t^{b,D}$ can be approximated arbitrarily well by terms:

$$h^0 + \sum_{n=1}^{\infty} \sum_{\substack{|\alpha|=n \\ \alpha \subseteq D}} \int_{-\infty}^{t} h^\alpha(t - s_1, \ldots, t - s_n) N^\alpha(\mathrm{d}s^n) \tag{9}$$

with $\alpha \subseteq D$ meaning that every index $\alpha_i \in D$. When dealing with actual data, one is not able to evaluate the above expression. Instead, we make this expression computable by the following two approximations.

A1: First off, we will truncate the outer sum at some finite step. Truncating the outer sum for instance at $n = 2$, amounts to including the effects of all pairs of events, and we will call this the **second order interactions**, since this corresponds to the effect of a pair of size 2 affecting the intensity of N^b (and similar for other orders). This includes both pairs from within the same mark and with different marks.

A2: Secondly we shall approximate the backward-kernels h^α by a tensor spline expansion as described in Appendix B.2. If $|\alpha| = n$, that is:

$$h^\alpha(x_1, \ldots, x_n) \approx \sum_{j_1=1}^{d} \cdots \sum_{j_n=1}^{d} \beta^\alpha_{j_1,\ldots,j_n} b_{j_1}(x_1) \cdots b_{j_n}(x_n)$$

Here d is the number of basis functions, b_j is the j'th basis function in the basis used and $\beta^\alpha_{j_1,\ldots,j_n}$ is a parameter. For simplicity, we are assuming that for a fixed n, we use the same spline basis in every coordinate, although one in principle could use different ones.

For a simpler notation, by stacking the $\beta^\alpha_{j_1,\ldots,j_n}$'s into a vector β^α and the basis functions into a vector $\Phi_n(x_1, \ldots, x_n)$[12], one has

$$h^\alpha(x_1, \ldots, x_n) = (\beta^\alpha)^T \Phi_n(x_1, \ldots, x_n)$$

[12]Omitting the concrete details of how the entries $(b_{j_1}(x_1) \ldots b_{j_n}(x_n))_{j_1,\ldots,j_n}$ are to be sorted into Φ.

Combining the two approximation, the intensity approximated at degree M then becomes:

$$\lambda_t^{b,D}(\beta) \approx \beta^0 + \sum_{n=1}^{M} \sum_{\substack{|\alpha|=n \\ \alpha \subseteq D}} \int_{-\infty}^{t-} \cdots \int_{-\infty}^{t-} \sum_{j_1=1}^{d} \cdots \sum_{j_n=1}^{d} \beta_{j_1,\ldots,j_n}^{\alpha} \prod_{v=1}^{n} b_{j_v}(t - s_v) N^{\alpha_v}(\mathrm{d}s_v)$$

$$= \beta^0 + \sum_{n=1}^{M} \sum_{\substack{|\alpha|=n \\ \alpha \subseteq D}} \int_{-\infty}^{t-} (\beta^{\alpha})^T \Phi_n(t - s_1, \ldots, t - s_n) N^{\alpha}(\mathrm{d}s^n) \tag{10}$$

Here we are using the same d in all coordinates for simplicity, but other choices could be relevant - for instance, since the parameter space grows rapidly in higher dimensions, one could dampen this by using a more narrow basis for high order interactions.

At first glance, we may still have a practical problem of computing the integral from $-\infty$. Rather than dealing with this explicitly however, this can be resolved by adding to **A2** the assumptions that the basis functions have a limited support, corresponding to limiting the memory of the process.

Note that if it is indeed the case that the kernels h^{α} are supported only on a limited set, the representation Equation (9) becomes exact, i.e. $\lambda^{b,D}$ is indeed equal to such a term, rather than arbitrarily well approximated since, in the language of Section 5, $\lambda^{b,D}$ would be a member of $L^1(\mathcal{F}_\tau)$ for some τ. This is however mainly an aesthetic pleasure, because for actual computability, we still need to make approximations **A1** and **A2**.

We shall only consider the case where we truncate the sum $M = 2$, but the general case is similar. In that case, **A1** and **A2** amounts to approximating $\lambda_t^{b,D}$ by:

$$\lambda_t^{b,D}(\beta) = \beta^0 + \sum_{v \in D} \int_{-\infty}^{t-} (\beta^v)^T \Phi_1(t - s) N^v(\mathrm{d}s)$$
$$+ \sum_{\substack{v_1, v_2 \in D \\ v_2 \geq v_1}} \int_{-\infty}^{t-} (\beta^{v_1 v_2})^T \Phi_2(t - s_1, t - s_2) N^{(v_1, v_2)}(\mathrm{d}s^2)$$

Note that in the sum over $v_2 \in D$ we assume $v_2 \geq v_1$, which is to avoid the over-parametrization of evaluating the symmetric basis twice. This presumes that $V \subseteq \mathbb{N}$, which is obviously possible for any finite-dimensional process.

To understand this, suppose that we didn't include the requirement $v_2 \geq v_1$, and consider $V = \{a, b\}$. Further consider two basis functions, say 1 and 2. Now the coefficient $\beta_{a,b}^{1,2}$ would occur with exactly the same terms as the coefficient $\beta_{b,a}^{2,1}$, which is an over-parametrization.

Since the above is linear in every β^{α}, and each one appears exactly once, we can again pull the β's outside, and stack them into one big vector, such that

$\lambda_t^{b,D} = \beta^T x_t^{b,D}$ with

$$x_t^{b,D} = \Bigg(1, \\ \left[\int_{-\infty}^{t-} \Phi_1(t-s) N^v(\mathrm{d}s) \right]_{v \in D}, \\ \left[\int_{-\infty}^{t-} \Phi_2(t-s_1, t-s_2) N^{v_1,v_2}(\mathrm{d}s^2) \right]_{\substack{v_1,v_2 \in D \\ v_2 \geq v_1}} \Bigg)$$

Computationally it is advantegeous that $x_t^{b,D}$ can be precomputed, and then used for any value of β.

7.2 Estimating λ through maximum likelihood

In the above, we used a generic choice of β without specifying how this could be chosen. Of course however, one has to make an intelligent choice of β.

First off, feasibility: Due to the requirement that $\lambda^{b,D} \geq 0$ one needs to make a feasible choice of β. Secondly optimality: One needs to choose β such that the observed data becomes likely. We do so by constrained, penalized maximum likelihood.

Feasibility

To comply with the feasibility of β, let $[0, T]$ be our region of interest, and let $(t_i)_{i=1,\ldots,n}$ be a grid on $[0, T]$. We add the constraint $\lambda_{t_i}^{b,D} \geq 0$ for each grid point. As discussed above, $\lambda_t^{b,D} = \beta^T x_t^{b,D}$, is linear in β, and so if X is the matrix where each row is the evaluated vector $x_{t_i}^{b,D}$, this constraint becomes:

$$X\beta \geq 0$$

which is a linear inequality constraint in β, and which typically can be solved very quickly.

While this doesn't guarantee that $\lambda^{b,D}$ will be globally positive, in light of the smoothness of $\lambda^{b,D}$, it is unlikely to become substantially smaller than 0.

Optimality

For the maximum likelihood problem, remember from Section 4.6 that the log-likelihood of a process with intensity $\lambda^{b,D}(\beta)$ is given by:

$$\log \mathcal{L}_T^{b,D}(\lambda^{b,D}) = \int_0^T \log \lambda_t^{b,D}(\beta) N^b(\mathrm{d}t) - \int_0^T \lambda_t^{b,D}(\beta) \mathrm{d}t$$

Given an observation, to choose a sensible parameter β for our approximation, we could choose the one that maximizes this likelihood. As discussed in Appendix B.3, we risk that very small gains in the optimization target can be

won only by choosing very large parameters, which produces very non-smooth kernels. To prevent this, we introduce regularization. Concretely we use the roughness penalty of Appendix B.3, which penalizes the curvature of $\beta^T\Phi$ by a quadratic form $\kappa(\beta) = \kappa_0\beta^T\Omega\beta$, where κ_0 is a steering parameter. This yields a penalized maximum likelihood problem.

In terms of β, the likelihood $\mathcal{L}^{b,D}$ of the intensity $\lambda^{b,D}(\beta)$ reads:

$$\log \mathcal{L}_T^{b,D}(\beta) = \int_0^T \log\left(\beta^T x_t^{b,D}\right) N^b(\mathrm{d}t) - \beta^T \int_0^T x_t^{b,D}\mathrm{d}t$$

As discussed in Proposition 4.23, the maximum likelihood problem is concave in the likelihood. Since the roughness penalty $\kappa(\beta)$ is convex in β, the concavity is preserved when subtracting a $\kappa(\theta)$, and so maximizing $\log \mathcal{L}_T^{b,D}(\beta) - \kappa(\beta)$ (or equivalently minimizing $-\log \mathcal{L}_T^{b,D}(\beta) + \kappa(\beta)$) is a convex optimization problem. Formally, we denote by $\hat{\beta}^D$ the solution of the constrained, penalized maximum likelihood problem:

$$\begin{aligned} \max_\beta \quad & \log \mathcal{L}_T^{b,D}(\beta) - \kappa_0\beta^T\Omega\beta \\ \text{s.t.} \quad & X\beta \geq 0 \end{aligned} \tag{11}$$

For concrete methods for solving this for the optimal β, see (Boyd and Vandenberghe 2004).

7.3 Distributions of maximum likelihood estimates

Below we will utilize the distribution of $\hat{\beta}$, the solution of the penalized maximum likelihood equation, and so in this section, we derive distributional results for this (omitting for a while the $\{b, D\}$-superscripts for readability).

If $\lambda(\beta)$ is a parametrization of intensities for $\beta \in \Theta$, let $\ell_T(\beta) = -\log \mathcal{L}_T(\beta)$ denote the negative log-likelihoods obtained for each β. The maximum likelihood estimates from above can be characterized by the (penalized) maximum likelihood equations:

Definition 7.1. For a class of point processes parametrized by $\lambda(\beta)$, the **maximum likelihood equation**[13] is:

$$-\nabla \ell_T(\beta) = 0$$

[13] Observe that if the gradient can be moved into the integral, we have:

$$\begin{aligned} -\nabla \ell_T(\beta) &= \int_0^T \nabla \log \lambda_t(\beta) N(\mathrm{d}t) - \int_0^T \nabla \lambda_t(\beta)\mathrm{d}t \\ &= \int_0^T -\nabla \log \lambda_t(\beta)\left[N - \Lambda(\beta)\right](\mathrm{d}t) \end{aligned}$$

Noticing how $N - \Lambda(\beta)$ resembles a process minus its compensator, this in fact shows that the maximum likelihood equation fits into the bigger framework of martingale estimation equations, which we shall not pursue here. See (Hansen 2013) for details.

Further, if $\kappa(\beta)$ is some penalty on β, the **penalized maximum likelihood equation** is:

$$-\nabla \left(\ell_T(\beta) + \kappa(\beta)\right) = 0$$

$\circ$

Suppose the true data generating procedure has parameter β_0. We define K_T to be the variance matrix of the left hand side of the penalized maximum likelihood equation,

$$\begin{aligned} K_T &= V\Big[-\nabla \left(\ell_T(\beta_0) + \kappa(\beta_0)\right) \Big] \\ &= V\Big[-\nabla \ell_T(\beta_0) \Big] \end{aligned}$$

The penalization term disappears, because this is non-stochastic. Similarly we may define the information matrix as the mean of the differentiated penalized maximum likelihood equation:

$$J_T = \mathbb{E}\Big[-\nabla^2 \left(\ell_T(\beta_0) + \kappa(\beta_0)\right) \Big]$$

It follows from (Hansen 2013, Propositions 4.5 and 4.7) that in the unpenalized case, the variance and information matrices K and J coincide and are given by:

$$J_T^{\kappa=0} = K_T^{\kappa=0} = \mathbb{E}\left[\int_0^T \frac{\nabla \lambda_s(\beta_0) \nabla \lambda_s(\beta_0)^T}{\lambda_s(\beta_0)} \mathrm{d}s \right]$$

Since K_T is anyways independent of κ, $K_T^{\kappa=0} = K_T$. And by the linearity of the mean,

$$J_T = J_T^{\kappa=0} - \nabla^2 \kappa(\beta_0) = K_T - \nabla^2 \kappa(\beta_0)$$

In a practical application, we can compute an empirical version of this, using a solution to the penalized maximum likelihood equation $\hat{\beta}$ (i.e. a maximum likelihood estimate) and by evaluating the concrete realization of λ instead of taking the mean. That is:

$$\hat{K}_T = \int_0^T \frac{\nabla \lambda_s(\hat{\beta}) \nabla \lambda_s(\hat{\beta})^T}{\lambda_s(\hat{\beta})} \mathrm{d}s \tag{12}$$

and $\hat{J}_T = \hat{K}_T - \nabla^2 \kappa(\hat{\beta})$. Using K and J or their empirical versions, one can now approximate the distributions of a solution to an penalized maximum likelihood equation.

(Hansen 2013, Theorem 4.12) shows that a solution $\hat{\beta}$ to the penalized maximum likelihood equation approximately follows a Gaussian distribution $N(\mu_{\hat{\beta}}, \Sigma_{\hat{\beta}})$ with:

$$\begin{aligned} \mu_{\hat{\beta}} &= \beta_0 + J_T^{-1} \nabla \kappa(\beta_0) \\ \Sigma_{\hat{\beta}} &= J_T^{-1} K_T J_T^{-1} \end{aligned}$$

When β_0 is unknown, we may instead compute $\hat{\mu}_{\hat{\beta}}$ and $\hat{\Sigma}_{\hat{\beta}}$ which uses $\hat{K}_T$, $\hat{J}_T$ and $\kappa(\hat{\beta})$. Note that from the above, it is evident that the the penalized maximum likelihood estimate $\hat{\beta}$ is biased away from the true mean by the bias term $J_T^{-1}\nabla\kappa(\beta_0)$.

Likelihood of linear intensities

In the concrete case of $\lambda_t(\hat{\beta}) = \hat{\beta}^T x_t$, we can compute the score and information matrices above. By differentiating, $\nabla\lambda_s(\hat{\beta}) = x_t$, and so:

$$\hat{K}_T = \int_0^T \frac{x_t x_t^T}{\hat{\beta}^T x_t} \mathrm{d}t$$

which can be estimated numerically.

Further, as suggested above, we use $\kappa(\beta) = \kappa_0 \beta^T \Omega \beta$ for some PSD matrix Ω. Since Ω is symmetric, $\nabla\kappa(\beta) = 2\kappa_0\Omega\beta$ and $\nabla^2\kappa(\beta) = 2\kappa_0\Omega$. Hence $J_T = K_T - 2\kappa_0\Omega$, and we can compute:

$$\begin{aligned} \hat{\beta} &\overset{approx}{\sim} \mathcal{N}\left(\beta_0 + 2\kappa_0 J_T^{-1}\Omega\beta_0, J_T^{-1}K_T J_T^{-1}\right) \\ &\overset{\mathcal{D}}{=} \mathcal{N}\left((I + 2\kappa_0 J_T^{-1}\Omega)\beta_0, J_T^{-1}K_T J_T^{-1}\right) \end{aligned}$$

Still we observe the bias $2\kappa_0 J_T^{-1}\Omega\beta_0$ to $\hat{\beta}$ away from β_0 due to the penalization. In practice this would mean, that if we wanted to perform some test $H_0 : \beta_0 = c$, we would not know the true mean of our estimator, leaving a test hard to perform (unless we knew β_0, in which case the test is anyways redundant).

Fortunately however, the bias term $2\kappa_0 J_T^{-1}\Omega$ is multiplied onto β_0, and so in the particular case of testing the hypothesis $\beta_0 = 0$, under the hypothesis, indeed $\hat{\beta}$ has mean 0.

7.4 Testing additive components to 0

We now develop the above into an independence test. First a bit of notation: For a coordinate $v \in D$, let $\hat{\beta}^v$ be the subvector of $\hat{\beta}^D$ corresponding to the kernels which describes the first order effect of N^v. Let similarly $\hat{\beta}^{v_1,v_2}$ be parameters for the second order interactions from (v_1, v_2).

Given a realization of a point process N on V, our overall interest is to test whether $A \not\to b \mid C$ for all $b \in B$ and some $A, B, C \subseteq V$. Remember from Section 6 that this is the case if each $\lambda^{b,A\cup C}$ actually only depends on N^C.

In our model, dependence of events of $A\backslash C$ enter the intensity $\hat{\lambda}^{b,C}$ in Equation (10) additively through the kernels $\Phi_1^T\hat{\beta}^v$ and $\Phi_2^T\hat{\beta}^{v_1,v_2}$ for $v \in A\backslash C$ and pairs (v_1, v_2) where at least one of v_1 and v_2 are in $A\backslash C$. (The generic case of either a first- or second order interaction, we simply denote $\Phi^T\hat{\beta}^\alpha$).

So heuristically, if all those '$A\backslash C$-related kernels' are 0, then $\lambda^{b,A\cup C}$ doesn't depend on the events of $N^{A\backslash C}$. In that case, still heuristically, $\lambda^{b,A\cup C}$ is equal

to something which depends only on events of N^C, i.e. is $\mathcal{F}^C$-measurable. This is the motivation for the following definition:

Definition 7.2. For $b \in V$ and $A, C \subseteq V$ let $\hat{\beta}^{b,A\cup C}$ be the solution to the penalized maximum likelihood problem Equation (11). We say that b **is locally empirically independent of** A **given** C if

$$\Phi_1^T \hat{\beta}^v \equiv 0 \qquad \Phi_2^T \hat{\beta}^{v_1,v_2} \equiv 0$$

for each $v \in A\backslash C$ and for all (v_1, v_2) where at least one is in $A\backslash C$. In this case we write $A \not\to_{\hat{\lambda}} b \mid C$. Similarly for $B \subseteq V$, $A \not\to_{\hat{\lambda}} B \mid C$ if $A \not\to_{\hat{\lambda}} b \mid C$ for every $b \in B$. ∘

Of course we must specify what we mean by this equivalence to 0 - in particular, due to noise, we expect fluctuations around the true mean of $\Phi^T \hat{\beta}^\alpha$. To do this, we perform a statistical test for whether equivalence to 0 is likely.

Example 7.3. We first illustrate this with a very simple example of the concept. Suppose we are given data from the true underlying a graph $a \to c \to b$. To test $a \not\to_{\hat{\lambda}} b \mid b, c$, we fit the $\mathcal{F}^{a,b,c}$-intensity of N^b, $\lambda^{b,a\cup b\cup c}$, by a second order expansion as described above, including the 3 first order effects and the 6 second order effects of a, b and c (3 pairs with distinct partners, 3 pairs (a, a) etc.). Now we would conclude that $a \not\to_{\hat{\lambda}} b \mid b, c$ if the first-order effect h^a and the second order effects $h^{(a,a)}, h^{(a,b)}, h^{(a,c)}$ are all found to be insignificant in the fitted model. ∘

For testing the components $\Phi^T \hat{\beta}^\alpha$ equal to 0, we follow the methodology by (Wood 2012) of evaluating $\Phi^T \hat{\beta}^\alpha$ in a grid $\mathbb{G}$:

Albeit one could test $\hat{\beta}^\alpha = 0$ directly, this isn't our preferred approach. For one, we wouldn't fully encapsulate the functional structure of $\Phi^T \hat{\beta}^\alpha$, since Φ weights the coefficients differently across the domain. Also, especially in the case of little curvature on the basis-functions, one could risk parameters that individually are big but, when weighted with Φ, sum to a function close to 0.

More formally, this is backed up by (Wood 2012), which compares the p-values from directly testing $\beta = 0$ to that of the grid approach, and finds that when the null hypothesis is indeed true, in the direct test, p-values were consistently too small, and so one too often rejects the null hypothesis (see Wood 2012, Fig 2).

Instead (Wood 2012) propose to evaluate Φ in a grid $\mathbb{G}$, and evaluate $\Phi^T \hat{\beta}^\alpha$ in each grid point. Since this is just a linear transformation of $\hat{\beta}^\alpha$, one can again compute a p-value of the joint distribution using the distribution of $\hat{\beta}^\alpha$. Let $\Phi(\mathbb{G})$ be the matrix where each column is Φ evaluated in a grid point. Then $\hat{g}^\alpha := \Phi(\mathbb{G})^T \hat{\beta}^\alpha$ is a vector of $\Phi^T \hat{\beta}^\alpha$ evaluated in $\mathbb{G}$.

Let $\mu_{\hat{\beta}}$ and $\Sigma_{\hat{\beta}}$ be the mean and covariance of $\hat{\beta}^D$ found in Section 7.3. For any α, let μ_α and Σ_α be the mean and covariance of $\hat{\beta}^\alpha$, which is simply obtainable by subsetting. Then the grid-evaluation $\hat{g}^\alpha$ is joint Gaussian with with mean $\mu_{g^\alpha} = \Phi(\mathbb{G})^T \mu_\alpha$ and covariance $\Sigma_{g^\alpha} = \Phi(\mathbb{G})^T \Sigma_\alpha \Phi(\mathbb{G})$.

In particular, we can compute a Wald test statistic for the null hypothesis $\Phi(\mathbb{G})^T \mu_\alpha = 0$ by:

$$\begin{aligned} T^\alpha &= (\hat{g}^\alpha)^T \Sigma_{g^\alpha}^{-1} \hat{g}^\alpha \\ &= (\hat{\beta}^\alpha)^T \Phi(\mathbb{G}) \left[\Phi(\mathbb{G})^T \Sigma_\alpha \Phi(\mathbb{G})\right]^{-1} \Phi(\mathbb{G})^T \hat{\beta}^\alpha \end{aligned}$$

which is $\chi^2_{(1)}$-distributed under the null hypothesis, and so a statistical test based on p-values can be performed at a desired level.

In general Σ_{g_α} can not be expected to be invertible due to the penalization (As also argued by (Wood 2006, Section 4.8.5): When penalizing heavily with smoothness penalty, cubic splines become straight lines, which has fewer degrees of freedom). (Wood 2012) propose to use instead a pseudo-inverse, and devices how to choose the rank of this (and demonstrates that this approach doesn't suffer from the same p-value skew as the direct approach does).

This allows for the following local empirical independence test: To test $A \not\to_{\hat{\lambda}} B \mid C$, for each $b \in B$ and each pair $v_1 \in A \backslash C, v_2 \in A \cup C$, compute the test statistics[14] $T^{v_1,b}$ and $T^{(v_1,v_2),b}$. If any of these are too large, say beyond the 95% quantile, reject the hypothesis of local empirical independence. We summarize the proposed test $A \not\to B \mid C$ in Algorithm 2.

Algorithm 2 Local Empirical Independence test (LEI, $A \not\to_{\hat{\lambda}} B \mid C$)

Input: Quantile q_0 of $\chi^2_{(1)}$-distribution.
Penalty matrix Ω and coefficient κ_0
Observation of $N^{A \cup B \cup C}$

Independent $\leftarrow$ **True**
for $b \in B$ **do**
  Compute the linear coefficients $x_t^{b,A\cup C}$ and $\int_0^T x_t^{b,A\cup C} \mathrm{d}t$
  Solve maximization problem:

$$\hat{\beta}^{b,A\cup C} \leftarrow \arg\max_\beta \int_0^T \log(\beta^T x_t^{b,A\cup C}) N^b(\mathrm{d}t) - \beta^T \int_0^T x_t^{b,A\cup C} \mathrm{d}t - \kappa_0 \beta^T \Omega \beta$$

  for $v_1 \in A \backslash C, v_2 \in A \cup C$ **do**
    Compute $T^{v_1,b}, T^{(v_1,v_2),b}$
    if $T^{v_1,b} > q_0$ or $T^{(v_1,v_2),b} > q_0$ **then**
      Independent $\leftarrow$ **False**
      Break
return Independent

[14] In the above we suppressed from notation that this is the test of effects on some $b \in V$. To avoid ambiguity, we explicitly write $T^{\alpha,b}$ to denote the effect of α on b.

8 Computational cost

We consider here the main computational challenge of local empirical independence algorithm above: Computing the integrated second order effect to be used in the likelihood:

$$\int_0^T x_t^{b,D,\text{second order}} \mathrm{d}t = \left[\int_0^T \int_{-\infty}^{t-} b_{j_1}(t-s_1) b_{j_2}(t-s_2) N^{v_1,v_2}(\mathrm{d}s^2)\mathrm{d}t \right]_{j_1,j_2,v_1,v_2}$$

The computational challenge is much larger than the entire remainder of the algorithm, due to the fact that 3 tasks grow quadratically in size:

C1 First off, the number of combinations of processes (v_1, v_2) grows quadratically with the dimension.

C2 Secondly, for each (v_1, v_2), the number of points (s_1, s_2) to be evaluated in the inner integral grows quadratically with the number of points in each coordinate process.

C3 And thirdly, every pair (s_1, s_2) needs to be evaluated in all combinations of tensor splines (j_1, j_2), which grows quadratically in the number of basis functions.

If $|D|$ is the number of coordinates included, v the average number of points in one coordinate process and M the number of (1d)-splines used in the tensor product, the number of evaluations to be made grows as $\mathcal{O}(|D|^2 v^2 M^2)$.

To give an example, in a process with $|D| = 3$ coordinates, each coordinate having $v = 20$ points and approximating the kernels via a tensor basis of $d = 3$ basis functions tensored with itself means evaluating $3^2 \times 20^2 \times 3^2 = 32.400$ combinations. But doubling each of these figures, yields $6^2 \times 40^2 \times 6^2 = 2.073.600$ combinations, and so doubling each of the 3 drivers has increased the task by a factor 2^6. In both cases, each of these pairs needs to be (numerically) integrated. Naturally, this is a substantial computational challenge, even for small processes. In this section, we present ways of substantially improving this computational task.

The effect **C1** is only truly scaling as $\frac{|D|^2+|D|}{2}$, since, as discussed in Section 7.1 above, one does not need to evaluate the effect of both (v_1, v_2) and (v_2, v_1). While this is still quadratic, it roughly halves the challenge.

A much larger computational saving can be obtained in **C2** by leveraging basis functions with a limited support. If all the basis function are supported only on $[0, y]$, then at a given time t, we need only to consider pairs (s_1, s_2) where $s_1, s_2 < t$ and $\max(t - s_1, t - s_2) = t - \min(s_1, s_2) < y$.

The speed-up from using local basis functions is dependent on the clustering of the process. In a very explosive process, very little is won. On the contrary, if points are very uniform on $[0, T]$, as one would expect of a homogeneous Poisson process, only a fraction $\frac{y^2}{T^2}$ of the points affect the intensity in any region.

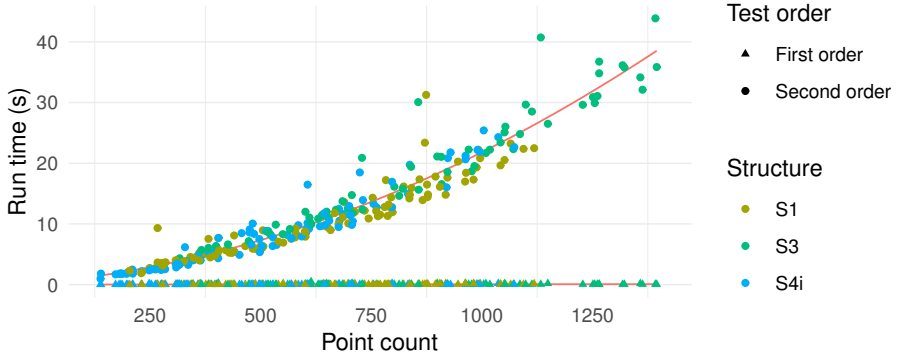


Figure 7: Runtime of 300 invocations of the local empirical independence test Algorithm 2. $a \not\to_{\hat{\lambda}} b \mid b, C$ was tested 100 times in different structure S1, S3 and S4i (see the Section 9 for the concrete structures). For each test, we simulate a Hawkes process (from the same model as in Section 9), and record the number of points in the process and the runtime of computing $\int_0^T x_t^{b,D,\text{first order}} \mathrm{d}t$ and $\int_0^T x_t^{b,D,\text{second order}} \mathrm{d}t$. The processes are simulated on an interval $[0, 1000]$, and the test uses splines supported on $[0, 20]$. As is evident from below, $|S1| = 3$, $|S3| = 4$ and $|S4i| = 2$.
The run-time is the elapsed time measured in seconds, and so is only useful for relative reference, since this depends on the machine running the algorithm. The concrete implementation runs the numerical integration of second order effects in parallel, and so the actual run-time was sped up by a factor of approximately the number of CPUs. The fitted line shows the best quadratic fit of the run-time.

Consider for instance the homogeneous Poisson case, with the expected number of points being $v = \lambda T$. Without local basis functions, $v^2 = \lambda^2 T^2$ pairs have to be evaluated. But using a basis with support $[0, y]$, approximately we have to evaluate $\frac{T}{y}$ regions, each region having λy points and so $(\lambda y)^2$ evaluations per region. In total, the number of evaluations is then $\lambda^2 y T$.

That is, if the number of points v increases due to a longer interval length $[0, T]$, the computational cost is linear in T. And while the cost of increasing the rate is still quadratic, the local case grows slower with a proportionality constant $\frac{y}{T}$.

Finally **C3** can be addressed. By using a sparse basis in making the tensor basis, only a small number of the tensor-splines are non-zero in every point. As found in Appendix B.2, for cubic B-splines in 2 dimensions, in any point in the space, the proportion of non-zero tensor-splines is only $\left(\frac{4}{K+4}\right)^2$ when using a the tensor-basis made from $K + 4$ (1-dimensional) B-splines. If correctly implemented, this eradicates the quadratic growth of **C3** completely.

Thus, although the problem grows rapidly, much can be reduced. For a Poisson process with fixed rate λ on $[0, T]$, the problem only grows as $\mathcal{O}(|D|^2 T)$ in $|D|$ and T.

In Figure 7 the run-times of 300 computations of the second order effect are plotted against the total point count of the process. For comparison, also the cost of computing the integrated first order effects are plotted. One observes that indeed the growth looks quadratic in the number of points. And the computational cost of performing the second order integrals is substantially higher than that of the first (which due to the scaling can barely be distinguished from 0).

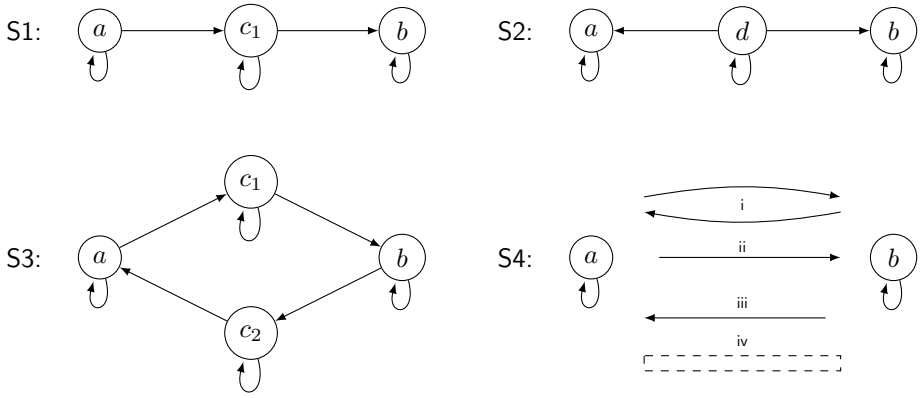


Figure 8: Seven different structures used for testing $a \not\to b|b, C$, where in in S1 $C = \{c_1\}$, in S3 $C = \{c_1, c_2\}$ and in S2 and S4, $C = \emptyset$. Each of the four graphs in S4 represent a different structure, with S4iv representing the case of no arrow between a and b.

9 Experimental results

In this section we evaluate the efficiency of the proposed methods. We conduct 3 different experiments to evaluate the performance of the local empirical independence (LEI) test from Algorithm 2. Section 9.1 evaluates the ability of the LEI to correctly identify whether $a \not\to b|C$ in various graphical structures containing a, b and C, and how this depends on the choice of regularization. Section 9.2 evaluates the performance of the LEI when used as the local independence test in Algorithm 1. Section 9.3 discusses the challenges of partially observed processes, and evaluates the algorithm in this setting, in particular in contrast to a test using only first order approximations.

In each experiment, given some graph $\mathcal{G}$, we sample a Hawkes process from an Ogata scheme with intensity:

$$\lambda_t^v = \alpha_0 + \sum_{v \in \mathrm{pa}_{\mathcal{G}}(v)} \alpha e^{-\beta(t-s)} N^v(\mathrm{d}s) \tag{13}$$

with $\alpha_0 = 0.1, \alpha = 0.4$ and $\beta = 0.8$. We simulate the Hawkes processes in an interval $[0, T]$ with $T = 1000$ and a burn-in period of 20.

9.1 Experiment 1: Choice of κ_0 and accuracy of test

First we test the accuracy of a single evaluation of the local empirical independence test. This is potentially very dependent on the choice of κ_0 in the penalization $\kappa(\beta) = \kappa_0 \beta^T \Omega \beta$, and so we evaluate the accuracy at several choices of κ_0, with the roughness penalty from Appendix B.3. We do so, by fixing a number of graphical structures, as displayed in Figure 8, for each simulating a

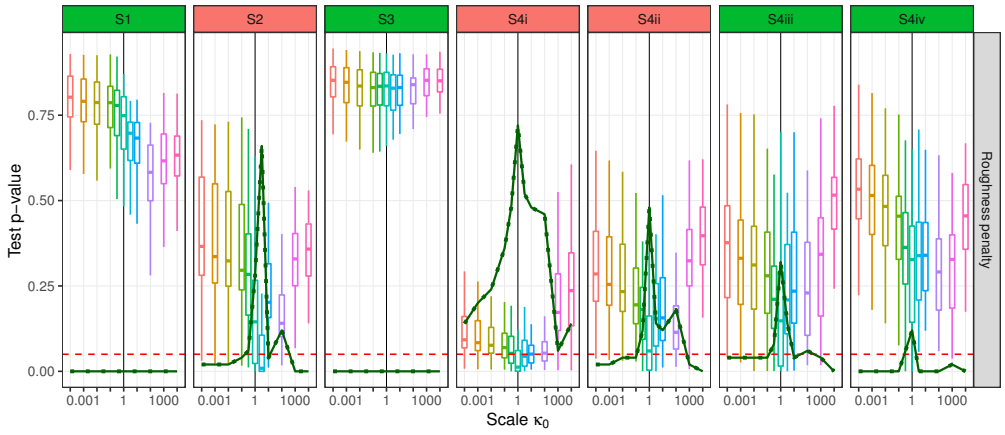


Figure 9: Boxplots of resulting p-values from each of the 7 structures shown in Figure 8. From each structure 100 Hawkes process was simulated, and the local emperical independence test from Algorithm 2 of $a \not\to_{\hat{\lambda}} b | b \cup C$ was run. Each test was run with the roughness penalty at various levels of κ_0 (Shown in color and in the x-axis a $\log_{10}$-scale). Each simulation thus produced a p-value, which is plotted. The red dotted line shows the 5%-level. Large p-values indicate that $a \not\to_{\hat{\lambda}} b \mid b, C$ while low p-values the opposite.

The header colors show the ground truth of whether $a \not\to b \mid b, C$, green indicating that this is the case. The dark-green line show the fraction of the simulated p-values falling below a 5%-level. The vertical line is the choice of κ_0 used in subsequent experiments.

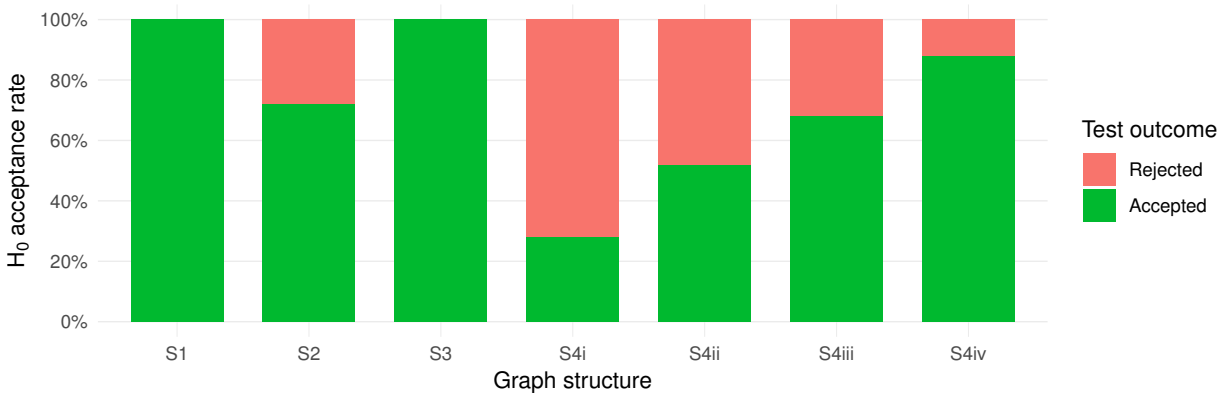


Figure 10: Summary of independence hypothesis acceptance rates in Experiment 1 when using $\kappa_0 = 1$ (Essentially this shows the values of the green line in Figure 9 at the intersection with the vertical $\kappa_0 = 1$ lines). The more each column matches the color of the corresponding header in Figure 9, the better.

Hawkes process with this structure and evaluate the accuracy of the independence test.

From each structure, we simulate 100 point processes. In each sampled process we test $a \not\to_{\hat{\lambda}} b|b, C$ (with C referring to the relevant nodes among c_1, c_2 in the graph), for each choice of penalization.

We note that the correct theoretical answers are that in S2, S4i and S4ii, $a \to b|b, C$, whereas in the others, $a \not\to b|b, C$. Thus a good test provides small p-values in S2, S4i and S4ii, and large in the 4 other structures.

The resulting p-values of each test are plotted by boxplots in Figure 9 at each level of κ_0. Also the proportion of p-values below 5% are indicated by the green lines.

We observe that identification of true local independences in S1 and S3 is very stable, with consistently very large p-values. Also S4ii and S4iv also have low p-values, although at some levels of κ_0, more than 5% are rejected. In the negative cases S2, S4i and S4ii, we observe that for either extremely large or extremely small penalizations, the independence hypothesis is way too often accepted. The best levels at κ_0 are those in the range from 1 to 10.

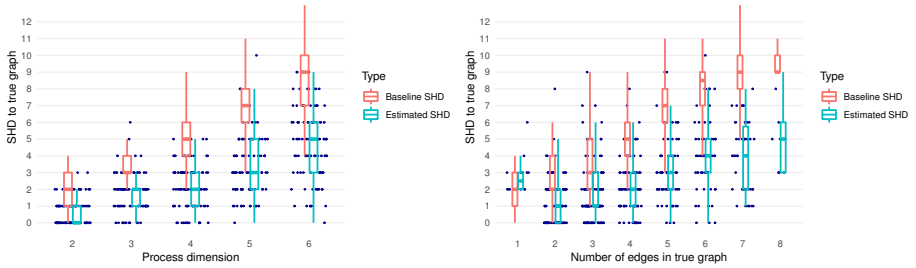
In the following experiment, we evaluate $\kappa_0 = 1$ (marked in vertical lines in Figure 9). Figure 10 shows a basic zoom-in of the results for $\kappa_0 = 1$. While $\kappa_0 = 1$ is best at identifying the true local dependence in S2, S4i and S4ii, it is also likely to make more false negatives in the local independence tests in S4iii and S4iv. Had one for instance used $\kappa_0 = 100$ on the instead, more often one would conclude local empirical independence, and so more aggressively remove edges.

In general we note that there is some evidence that the approximate normal distribution of $\hat{\beta}^{b,D}$ is not very fitting - in particular, although it is of course nice to identify the local independences of S1 and S3 in 100% of the cases, from a theoretical point of view, one would expect to reject these 5% of the time. Similarly, S4iii and S4iv are accepted too seldomly, if we had indeed used the correct distribution of $\hat{\beta}^{b,D}$.

9.2 Experiment 2: Performance in the Empirical Causal Analysis algorithm

In this experiment, we evaluate the local empirical independence test in the context of the ECA presented in Algorithm 1.

For each $d \in \{2, 3, 4, 5, 6\}$ we draw 100 random graphs on $V = \{1, \ldots, d\}$ by letting self-edges $v \to v$ appear with probability $p_0 = 0.7$ and all other edges with probability $p_1 = 0.1$, where occurrence of all edges are drawn independently (the $d = 2$ case is run by $p_0 = 0.8, p_1 = 0.4$, to avoid too sparse graphs). To avoid explosivity, each graph is discarded, if more than 8 edges are present. Given the graph, we simulate a process from the same intensity as above in Equation (13), on $[0, T]$ with $T = 1000$ and a burn-in of 20. Based on the



(a) (b)

Figure 11: Structural Hamming Distances of graphs estimated using the ECA-algorithm with the developed local independence test. For each dimension in $\{2, \ldots, 6\}$, a 100 random graphs were simulated, and a Hawkes process simulated, with parents dictated by the graph. The ECA-algorithm was run twice, once with the developed test, and a baseline ECA-algorithm where every test $A \not\to B|C$ is accepted with probability $\frac{1}{2}$. Boxplots of both baseline and estimated SHD are shown. Further the blue points show the distribution of the estimated SHDs.

results above, we use $\kappa_0 = 1$ for the local independence test.

For each simulated graph, we then run the ECA algorithm, and compare the resulting graph with the true graph used for generation. Running a PC-type algorithm, like the ECA, is known to involve substantial uncertainty. Although, under faithfulness assumptions, an oracle conditional independence test is guaranteed to produce the true graph, a test with uncertainty may not do so. Further, decisions on whether or not to remove edges, affect later decisions, and so once an error is made, this is likely to generate further errors through the algorithm.

So, as a baseline, on each simulated point process, we also run the conditional independence test, which chooses with equal probability between a positive and a negative answer. We compare the resulting graphs to the true graph using the Structural Hamming Distance:

Definition 9.1. For two graphs $\mathcal{G}_1 = (V, E_1), \mathcal{G}_2 = (V, E_2)$, the **Structural Hamming Distance** (SHD) is the minimal number[15] of edge flips, creations or deletions one needs to make in $\mathcal{G}_1$ to obtain $\mathcal{G}_2$.

A large SHD indicates that two graphs are very different, while a small SHD

[15] Given two edge sets, E_1 and E_2, one can compute the SHD by letting

$$B = \{(i,j) \in E_1 \triangle E_2 \mid (j,i) \in E_1 \triangle E_2,\ \{(i,j),(j,i)\} \not\subseteq E_1, E_2\}$$

and $A = (E_1 \triangle E_2) \backslash B$, where $\triangle$ denotes the symmetric difference. Now $SHD(\mathcal{G}_1, \mathcal{G}_2) = |A| + \frac{|B|}{2}$.

that they are similar. ○

Results are plotted in Figure 11. In Figure 11a the SHD is shown for each of the simulated dimensions, and the graphs estimated from the algorithm perform substantially better than the baseline for all dimensions. For instance, in a process of dimension 5, the median SHD is 3, while the baseline test had a median SHD of 7.

In Figure 11b the SHD is shown for the number of edges in the true graph, and also here the algorithm performs substantially better than baseline. In particular, the growth of the SHD with the number of edges in the simulated graph is much slower for the estimated graph.

9.3 Experiment 3: The latent case

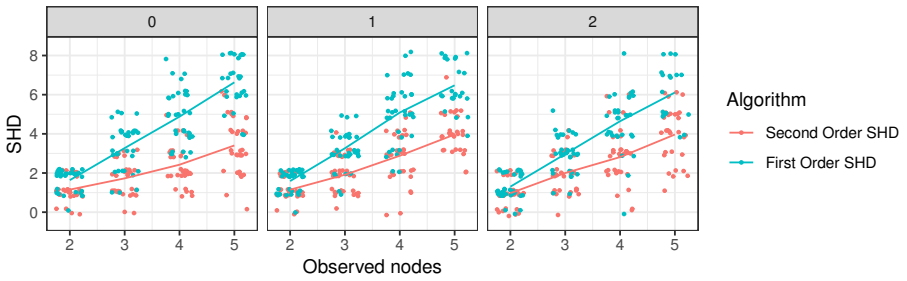


Figure 12: Structural Hamming Distances of graphs estimated using the ECA-algorithm with a first- and second-order local empirical independence test (second being the standard one, used above). Each of the boxes 0, 1 and 2 indicate the number $|V \backslash O|$ of latent variables. The lines represent the average SHD within each group.

We first give a little bit of motivation for this last experiment. In the above, we have tested the ECA algorithm in the case of a fully observed processes, using the local empirical independence test based on a second order expansion.

By now the reader may be wondering: If our interest is primarily in the Hawkes process, which is governed by first-order effects only, why even bothering including higher order interactions (and doing all the tedious work in Section 5)?

The reason is that Hawkes processes are not closed under marginalizations. That is, if $N = (N^1, \dots, N^d)$ is a Hawkes process, there is no guarantee that a $N^O = (N^{o_1}, \dots, N^{o_m})$ is a Hawkes process for a subset $O = \{o_1, \dots, o_m\} \subseteq \{1, \dots, d\}$.

This is problematic for two reasons. First, if one observes only a subset $O \subseteq V$ of the process coordinates, since N^O is not necessarily Hawkes, one is not guaranteed that the system can be described by first order interactions. However by the results of Section 5, no matter the class of the marginalized process,

one can approximate it arbitrarily well, by including higher and higher order interactions. Concretely, in our test, we hope that the 'non-Hawkes' behaviour of marginalized models can be encaptured as second order interactions.

Secondly, even in a fully observed process, when one performs a local independence test $a \not\to b|C$, although the entire process may be a Hawkes process, the process $N^{a,b,C}$ need not to be, and so the local independence test cannot assume that only first-order interactions are present between coordinates a, b and C. As detailed in Section 8, the computational complexity grows quadratically with the number of included processes, and so from a computational perspective, it is very beneficial to have algorithms that remove as many edges as possible with tests $a \not\to b|C$ where C is small. Hence the need for at test which can handle $N^{a,b,C}$ which is not Hawkes.

In this experiment, we seek to test the robustness of the method in the face of latent coordinates. If $\mathcal{G} = (V, E)$ is the graph of the full system, and $O \subseteq V$ are observed coordinates, (Mogensen 2019) show that Algorithm 1 will produce a supergraph $\mathcal{G}_{\mathrm{CA}}$ of the parent graph $\mathcal{P}_O(\mathcal{G})$. We say that $\mathcal{P}_O(\mathcal{G}) = (O, F)$ is a **parent graph** of $\mathcal{G}$ on O if F contains edge an edge $\alpha \to \beta$ if 1) $\alpha \to \beta$ in E or 2) there exist latent nodes $l_1, \dots, l_n \in V \backslash O$ such that $\alpha \to l_1 \dots \to l_n \to \beta$ in E. Consequently, the central property of parent graphs is that $\mathrm{An}_{\mathcal{P}_O(\mathcal{G})}(\beta) = \mathrm{An}_{\mathcal{G}}(\beta) \cap O$ for every $\beta \in O$.

And so, while we do not identify the true graph $\mathcal{G}$, the parent graph tells us that if $a \notin \mathrm{An}_{\mathcal{P}_O(\mathcal{G})}(b)$, then a is not a causal ancestor of b in $\mathcal{G}$.

To test the local independence test under latent coordinates, for each $|O| \in \{2, 3, 4, 5\}$ and each $|V \backslash O| \in \{0, 1, 2\}$ we simulate 30 graphs $\mathcal{G}$ of size $|O| + |V \backslash O|$, by including all self-loops and a 15% probability for every non-diagonal edge to appear. We run Algorithm 1 using the standard second order LEI. Further, we run Algorithm 1 using also the LEI, but where we fit only the first order effects (which was tuned for κ_0 similar to Experiment 1).

For each resulting graph, $\mathcal{G}$, we measure structural hamming distance to the ground truth $\mathcal{G}_0$[16], the results being shown in Figure 12.

We observe that a substantial benefit is earned by including the second order effects. The difference does not seem to increase much with more truly latent coordinates. This could however just be due to the fact that coordinates not in a, b, C appear as latent when testing $a \not\to_{\hat{\lambda}} b \mid C$, also in the fully observed case.

[16]Due to Markov equivalences arising in marginalized graphs, we do not use the parent graph $\mathcal{P}_O(\mathcal{G})$ but rather the parent graph of the maximal element of the markov equivalence class. See (Mogensen 2019) for details.

10 Discussion

In this thesis, we have studied local independence of point processes, and how local independence, through the global Markov property, can be studied in the context of graphs and graph separation.

Further we have proven that point process intensities can be approximated arbitrarily well by considering increasing orders of integrals of kernels, a result similar in nature to that of Volterra series.

We have proposed to use this approximation in a non-parametric test for local independence, which uses basis splines to approximate kernels. We estimate the intensity using penalized maximum likelihood, and we derive distributional results, such that the test relies on the approximate normality of the estimated parameter $\hat{\beta}^{b,D}$. Using this, we test for local empirical independence by testing for insignificance of a collection of kernels.

We have discussed the quadratic growth of the complexity, in particular that while this cannot be alleviated in the number of coordinates, by using limited memories of the kernels, this can in some cases be reduced to linear growth in the number of points.

Finally, we have evaluated the performance of the algorithm in simulated data sets. We examined the consistency of the test and found that in case the underlying processes are truly local independent, there is some evidence that the distributional approximations is not perfectly describing the actual distribution, with acceptance rates both below and above the desired level, depending on the tested structure. Further, in the structures where the underlying processes is not locally independent, we still accept the independence hypothesis in a large number of cases. It is very possible that further work could improve on this lack of strength, for instance by choosing alternative regularizations, which considers the size of the tested structure, or by finding ways of improving the model fit. It may however also be, that in the concrete settings used for experiments, the local independence is so hard to identify that not much can be gained from improving the fit. Nonetheless, we also find that when applying the test in structure learning algorithms, the simulations show very promising results.

The proposed test contributes to the literature on causal structure learning in dynamical systems by providing a non-parametric test which is able to handle latent factors in a computationally feasible manner.

References

Aalen, Odd O (1987). "Dynamic modelling and causality". In: *Scandinavian Actuarial Journal* 1987.3-4, pp. 177–190.

Ahmed, NU (1970). "Closure and completeness of Wiener's orthogonal set {Gn} in the class L2 (Ω, B, μ) and its application to stochastic heriditary dif". In: *Information and Control* 17.2, pp. 161–174.

Boyd, Stephen and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Cohen, Samuel N (2012). "Chaos representations for marked point processes". In: *Communications on Stochastic Analysis* 6.2, pp. 263–279.

Daley, Daryl J and David Vere-Jones (2003). *An introduction to the theory of point processes. Vol. I. Probability and its Applications*. New York). Springer-Verlag, New York,

Daley, Daryl J and David Vere-Jones (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

Didelez, Vanessa (2000). *Graphical models for event history analysis based on local independence*. Logos Berlin.

Didelez, Vanessa (2008). "Graphical models for marked point processes based on local independence". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 245–264.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Fukumizu, Kenji, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf (2008). "Kernel measures of conditional dependence". In: *Advances in neural information processing systems*, pp. 489–496.

Gretton, Arthur, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola (2008). "A kernel statistical test of independence". In: *Advances in neural information processing systems*, pp. 585–592.

Hansen, Niels Richard (2013). *Point Processes*.

Hawkes, Alan G (1971). "Spectra of some self-exciting and mutually exciting point processes". In: *Biometrika* 58.1, pp. 83–90.

Jacobsen, Martin (2006). *Point process theory and applications: marked point and piecewise deterministic processes*. Springer Science & Business Media.

Jacod, Jean and Albert Shiryaev (1987). *Limit theorems for stochastic processes*. Vol. 288. Springer-Verlag Berlin Heidelberg.

Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.

Meek, Christopher (2014). "Toward learning graphical and causal process models". In: *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction-Volume 1274*. CEUR-WS. org, pp. 43–48.

Mogensen, Søren Wengel (2019). *Causal screening for dynamical systems*. `https://soerenwengel.github.io/files/causalScreening.pdf`. Unpublished.

Mogensen, Søren Wengel and Niels Richard Hansen (2018). "Markov equivalence of marginalized local independence graphs". In: *arXiv preprint arXiv:1802.10163*. To appear in Ann. Statist.

Mogensen, Søren Wengel, Daniel Malinsky, and Niels Richard Hansen (2018). "Causal learning for partially observed stochastic dynamical systems". In: *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018Conference on Uncertainty in Artificial Intelligence*. Association For Uncertainty in Artificial Intelligence (AUAI), pp. 350–360.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Ramsey, Joseph D (2014). "A scalable conditional independence test for nonlinear, non-Gaussian data". In: *arXiv preprint arXiv:1401.5031*.

Richardson, Thomas, Peter Spirtes, et al. (2002). "Ancestral graph Markov models". In: *The Annals of Statistics* 30.4, pp. 962–1030.

Schilling, René L (2017). *Measures, Integrals and Martingales*. Cambridge University Press.

Schweder, Tore (1970). "Composable markov processes". In: *Journal of applied probability* 7.2, pp. 400–410.

Sokol, Alexander and Niels Richard Hansen (2014). "Causal interpretation of stochastic differential equations". In: *Electronic Journal of Probability* 19.100, pp. 1–24.

Sokol, Alexander and Anders Rønn-Nielsen (2013). *Advanced Probability*. Department of Mathematical Sciences, University of Copenhagen.

Spirtes, Peter, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson (2000). *Causation, prediction, and search*. MIT press.

Volterra, Vito (1887). *Sopra le funzioni che dipendono da altre funzioni*. Tip. della R. Accademia dei Lincei.

Volterra, Vito (1959). *Theory of functionals and of integral and integro-differential equations*. Dover.

Wood, Simon N (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, Simon N (2012). "On p-values for smooth components of an extended generalized additive model". In: *Biometrika* 100.1, pp. 221–228.

A Enumerations and integrals

In this appendix we show that point processes can be characterized by their jumps. Further, we discuss integration with respect to a point process.

As a motivation, consider Figure 4, where the process obtained by evaluating $t \mapsto N(0,t]$ would have discontinuities at a number of points T_n. This sparks the idea that in the plotted region, $N = \delta_{T_1} + \delta_{T_2} + \delta_{T_3}$. The below guarantee that this construction is always possible.

Considering any fixed element $\mu \in \mathcal{N}_{\mathcal{X}}^{\#}$ (i.e. in the non-random case), μ decomposes as a countable sum of Dirac measures: μ is a finite sum of Dirac measures on any bounded set[17], and by separability, $\mathcal{X}$ is a countable union of bounded sets. Hence $\mu = \sum_{i\in\mathbb{N}} \delta_{t_i}$ for some $(t_i)_{i\in\mathbb{N}} \subseteq \mathcal{X}$.

Actually this is a little inaccurate, because of course μ could be finite, and so this sum should be finite. Since it however is more convenient to always have an infinite sum, one could make some extension of $\mathcal{X}$ to $\bar{\mathcal{X}}$, and exclude the surplus summands by letting $t_i \in \bar{\mathcal{X}}$, ensuring that $\sum_{i\in\mathbb{N}} \delta_{t_i}(A)$ is a correct representation of μ for all $A \in \mathcal{B}(\mathcal{X})$. The obvious motivation for doing so is the case of $\mathcal{X} = \mathbb{R}$, with $\bar{\mathcal{X}} = \mathbb{R} \cup \{-\infty, \infty\}$ - 'events t_i that doesn't occur will put at $\pm\infty$'.

Now if we don't consider a fixed element $\mu \in \mathcal{N}_{\mathcal{X}}^{\#}$, but rather a probability distribution over $\mathcal{N}_{\mathcal{X}}^{\#}$ (i.e. a point process), we can obtain a similar result:

Proposition A.1. *For a given point process N on $\mathcal{X}$ there exists a sequence $T_1, T_2, \ldots$ of random variables on $\bar{\mathcal{X}}$ such that for any set $B \in \mathcal{B}(\mathcal{X})$ and $\omega \in \Omega$, one has:*

$$N_\omega(A) = \sum_{i\in\mathbb{N}} \delta_{T_i(\omega)}(A)$$

Proof. This follows from the arguments preceding (Daley and Vere-Jones 2007, Lemma 9.1.XIII). □

We call the sequence $T_1, T_2, \ldots$ an **enumeration** of N, or less formally the points, jumps or events of N.

In the case of $\mathcal{X} = \mathbb{R}$, we will often enumerate this sequence by $\mathbb{Z}$ instead of $\mathbb{N}$ and ensure that the events are ordered, i.e. $i < j \implies T_i \leq T_j$. This can for instance be done by letting:

$$T_i = \begin{cases} \inf\{t > 0 | N(0,t] \geq i\} & i \geq 1 \\ \sup\{t \leq 0 | N(t,0] > -i\} & i \leq 0 \end{cases}$$

with the conventions that $\inf(\emptyset) = \infty, \sup(\emptyset) = -\infty$. The assumption that a process is simple is equivalent to the assumption that almost surely the (finite) T_i's are distinct. With a representation as a sum of Dirac measures, the definition of integrals with respect to point processes are straightforward:

[17] This is not trivial, but follows by the assumption that $\mathcal{X}$ is complete and separable. See (Daley and Vere-Jones 2003, Appendix 1.6) for details.

Definition A.2. Let $f^+ : \mathcal{X} \to \mathbb{R}$ be a positive measurable function. Define by $\int_{\mathcal{X}} f^+(t)N(\mathrm{d}t)$ the (not-necessarily finite-valued) random variable obtained by for each $\omega \in \Omega$ integrating f^+ with respect to the measure N_ω.

If $f : \mathcal{X} \to \mathbb{R}$ is any measurable function, and $P\left(\int_{\mathcal{X}} |f|(t)N(\mathrm{d}t) < \infty\right) = 1$, we define the (a.s. finite-valued) integral $\int_{\mathcal{X}} f(t)N(\mathrm{d}t)$ similarly. ∘

Since integration with respect to a Dirac measure simply means plugging in, i.e. $\int_{\mathcal{X}} f(t)\delta_x(\mathrm{d}t) = f(x)$, it follows that $\int_A f(t)N(\mathrm{d}t) = \sum_{i:T_i \in A} f(T_i)$ for integrable f.

B Tensor splines

In this appendix we introduce B-splines, and discuss how one splines in one variable can be extended to splines of multiple variables. We also discuss the roughness penalty, a way of biasing estimates towards smoother functions.

To approximate the kernels we use a basis expansion of tensor splines. A major advantage of using a spline expansion is the fact that most oftenly, one uses an additive decomposition $\beta_1 B_1(x) + \ldots + \beta_n B_n(x)$, which is linear in the parameter β. We stack the basis functions into a vector function $\Phi(x) = (B_1(x), \ldots, B_n(x))$ and thus a linear combination of basis functions is written as $\beta^T \Phi$. For ease of notation, below we do this also in the case of tensor bases, even though the vector ordering is not as obvious in this case.

B.1 B-splines in one dimension

One such basis spline are the B-splines. They're constructed by functions that are piecewise polynomial, with the pieces defined on regions that overlap at 'knots', and with continuity assumptions at the boundary on the function and derivatives.

Concretely, one defines a B-spline on an interval $[\xi_0, \xi_{K+1}]$ from it's knots $(\xi_i)_{i=0,\ldots,K+1}$, and it's order M. A order M-spline consists of polynomials of order $M-1$ and has continuous derivatives up to order $M-2$, the most frequent case being $M = 4$, which is known as *cubic* splines. Following (Friedman et al. 2001) we let:

$$\begin{aligned} \tau_1 = \ldots = \tau_M &= \xi_0 \\ \tau_{M+1} = \xi_1, \ldots, \tau_{M+K} &= \xi_K \\ \tau_{M+K+1} = \ldots = \tau_{M+K+K} &= \xi_{K+1} \end{aligned}$$

and define recursively the functions:

$$\begin{aligned} B_{i,1}(x) &= \mathbb{1}_{[\tau_i, \tau_{i+1})}(x) \\ B_{i,m}(x) &= \frac{x - \tau_i}{\tau_{i+m-1} - \tau_i} B_{i,m-1}(x) + \frac{\tau_{i+m} - x}{\tau_{i+m} - \tau_{i+1}} B_{i+1,m-1}(x) \end{aligned}$$

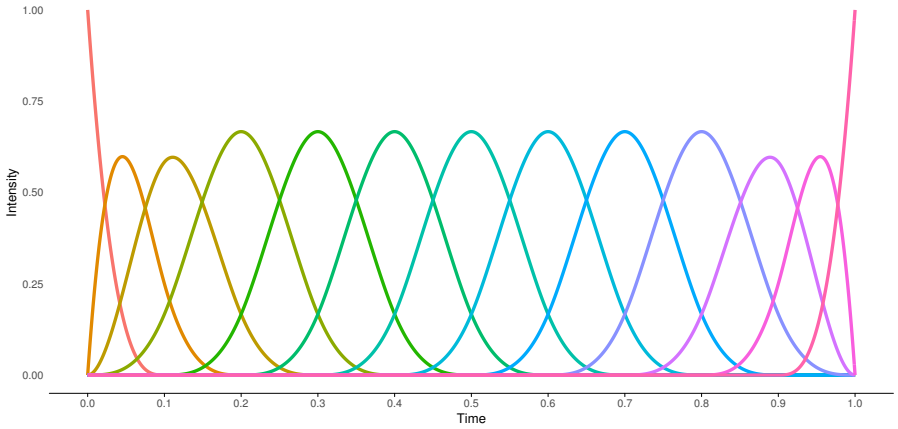


Figure 13: Plot of B-splines with order $M = 4$ and internal knots $0.1, \ldots, 0.9$. In total there are $K + M = 13$ functions.

for $i = 1, \ldots, K + 2M - m$ and m up to order M. Also, while zero-division is possible if $\tau_{i+m-1} = \tau_i$ this only occurs if also the denominator $B_{i,m-1} \equiv 0$, so we adopt the convention that in this case, the result is zero.

Now the functions $B_{i,M}, 1 \leq i \leq K + M$ constitute what is referred to as the B-splines of order M. The case of $M = 4$ and $K = 9$ with $\xi_i = \frac{i}{10}$ for $i = 0, \ldots, 10$ is shown in Figure 13.

Once a basis $\Phi = (B_{1,M}, \ldots, B_{K+M,M})$ is chosen, one can now consider the function space spanned by linear combinations $\beta^T \Phi$ for $\beta \in \mathbb{R}^{K+M}$, all of which also have continuous derivatives up to order $M - 2$.

An important feature of the B-splines is the fact that each spline is supported only on a set of M intervals between $M + 1$ knots (splines near the boundary on even fewer). This is clear for the first order splines, since they're simply indicator functions on one interval, and since higher-order splines are sums of adjacent lower-order splines, this is true for all orders. That implies that for any x, at most M splines will be non-zero when evaluated in x. For instance in Figure 13, for any $x \in [0, 1]$, at most 4 of the 13 spline-functions is non-zero in x.

The fact that B-splines are locally supported is a very important computational feature, since this will produce sparse representations, with a 'degree-of-sparsity' of $\frac{K}{K+M}$. As we shall see below, upon considering tensor-splines, this sparseness is only fortified in higher dimensions.

B.2 Tensor splines

How can we generalize basis expansions to higher dimensions? One way to do so is via **tensor splines**. For instance in the two dimensional case, let $\Phi_1 =$

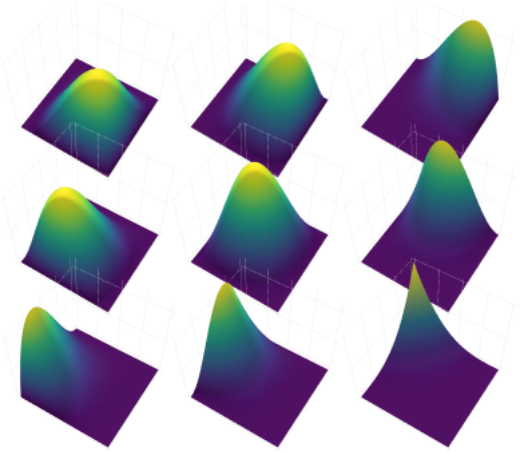


Figure 14: Plot of the tensor spline basis made from 3 B-spline basis functions tensored with themselves. Note how the plots are mirrored through the diagonal.

$(B_1^1, \ldots, B_{m_1}^1)$ and $\Phi_2 = (B_1^2, \ldots, B_{m_2}^2)$ be two basis expansions on the intervals $[a_1, b_1]$ respectively $[a_2, b_2]$. Now define a basis expansion $\Phi = (B_{i,j})_{i \leq m_1, j \leq m_2}$ on $[a_1, b_1] \times [a_2, b_2]$ by the functions:

$$B_{i,j}(x, y) = B_i^1(x) B_j^2(y)$$

I.e., in vector form, Φ consists of the entries of the outer product $\Phi_1 \Phi_2^T$. In dimensions higher than 2, one proceeds similarly. See (Wood 2006, Section 5.6) for further details.

Example B.1. We consider in particular the d-dimensional case on an interval $[a, b]^d$, in each coordinate using the same cubic B-splines with K internal knots (and so $K + 4$ basis functions).

The size of the basis grows exponentially as $(K + 4)^d$ in the dimension. However, since for any $x \in [a, b]^d$, at most 4 basis functions are active in each coordinate, only 4^d basis functions are non-zero, which also grows exponentially but at a much slower rate. Thus the sparseness of the basis is $1 - \left(\frac{4}{K+4}\right)^d$.

If for instance $d = 2$ and $K = 10$, the dimension of the parameter space becomes 196, but in any point in the region, only 16 functions are active. ∘

B.3 Regularization of splines

The function space spanned by a spline basis is very rich, especially when including many basis functions. When optimizing some functional of $\beta^T \Phi$ over β, one risks that the optimization produces parameters β such that $\beta^T \Phi$ behave very 'wiggly', to be able to fit the optimization target a little bit better.

To dampen this, one can add some regularization $\kappa(\beta)$, to avoid the resulting estimate behaving too wildly.

For our purposes we choose the roughness penalty, cf. (Friedman et al. 2001), which penalizes curvature of $\beta^T\Phi$. In a d-dimensional tensor spline, we let $\Phi(x)$ be shorthand for the vector form $\Phi(x_1, \ldots, x_d)$. For each pair of coordinates (i, j), we square the curvature $\frac{\partial^2}{\partial x_i x_j}\beta^T\Phi(x)$, and sum this up. This we integrate over our area of interest $[a, b]^d$. Finally, we scale the penalty with some constant κ_0, which steers our prior belief on the smoothness of the estimate. That is:

$$\begin{aligned}
\kappa(\beta) &= \kappa_0 \int_{[a,b]^d} \sum_{i,j}^{d} \left(\frac{\partial^2}{\partial x_i x_j} \beta^T \Phi(x) \right)^2 \mathrm{d}x \\
&= \kappa_0 \int_{[a,b]^d} \sum_{i,j}^{d} \left(\beta^T \frac{\partial^2}{\partial x_i x_j} \Phi(x) \right) \left(\beta^T \frac{\partial^2}{\partial x_i x_j} \Phi(x) \right)^T \mathrm{d}x \\
&= \kappa_0 \beta^T \underbrace{\int_{[a,b]^d} \sum_{i,j}^{d} \left(\frac{\partial^2}{\partial x_i x_j} \Phi(x) \right) \left(\frac{\partial^2}{\partial x_i x_j} \Phi(x) \right)^T \mathrm{d}x}_{:=\Omega} \beta \\
&= \kappa_0 \beta^T \Omega \beta
\end{aligned}$$

where when pulling β outside of the integral, we consider the entry-wise integral of the matrix. Thus we see that the penalty is just a quadratic form in β.

Note the similarity to having chosen simply a L^2 penalty, which would correspond to the choice of $\Omega = I$.

C Implementing integration of 2nd order effects

This appendix adds technical details to the computational analysis of Section 8.

When computing the second order integrals, it can be beneficial for implementation purposes to switch orders of integration, such that the time integral is evaluated first.

That is:

$$\int_0^T \int_{-\infty}^{t-} b_{j_1}(t - s_1) b_{j_2}(t - s_2) N^{v_1, v_2}(\mathrm{d}s^2)\, \mathrm{d}t = \int_{-\infty}^{T} \int_{(0 \vee s_1 \vee s_2)+}^{T} b_{j_1}(t - s_1) b_{j_2}(t - s_2) \mathrm{d}t\, N^{v_1, v_2}(\mathrm{d}s^2)$$

Where we maintain the notation of collapsing the double integral wrt. N into one. This form is simple for implementation purposes: For each pair of events, one should compute the inner integral, and sum all those computed values together.

If one chooses a kernel with bounded support, say on an interval $[0, y]$, the outer integrals $\int_{-\infty}^{T} \cdots N^v(\mathrm{d}t)$ can be reduced to $\int_{-y}^{T} \cdots N^v(\mathrm{d}t)$, because for $s < -y$, one has $t - s > y$ when $t \in (0 \vee s_1 \vee s_2)+$.

Further, the upper limit of the integral can be reduced to $T \wedge ([s_1 \wedge s_2] + y)$, because for $t > y + [s_1 \wedge s_2]$ either $t - s_1 > y$ or $t - s_2 > y$, and so either b_{j_1} or b_{j_2} is off it's support. In conclusion one needs to compute:

$$\int_{-y}^{T} \int_{(0 \vee s_1 \vee s_2)+}^{T \wedge ([s_1 \wedge s_2]+y)} b_{j_1}(t - s_1) b_{j_2}(t - s_2) \mathrm{d}t \, N^{v_1, v_2}(\mathrm{d}s^2)$$

This allows for a straight-forward implementation: One loops over pairs (s_1, s_2), and if $|s_1 - s_2| < y$, one computes the inner integral, and adds it to the total output sum.